7-1. Generalized Linear Mixed Models

Subject-Specific Models

- Assumptions
  Given the subject-specific effects b_i (a $q-$vector), the responses Y_{ij} are independent and follow a distribution from the exponential family. Let

$$E(Y_{ij}|b_i) = \mu_{ij},$$

  then

$$g(\mu_{ij}) = \eta_{ij} = X_{ij}^T\beta + Z_{ij}^T b_i$$

  where η_i is the linear predictor and g is the link function. X_{ij} and Z_{ij} are $p-$ and $q-$ vectors of covariates with Z often being a subset of X.

- Approaches of Inference

  1. Conditional likelihood

- appropriate when only interested in regression coefficients that do not vary across subjects (i.e., not the intercept in a model with subject-specific intercept).
- Subject-specific effects $b_1, b_2, \cdots, b_m$ are treated as nuisance parameters.
- estimate β using the conditional likelihood given the "sufficient" (more or less) statistics for b_i.

2. Marginal likelihood
   - appropriate when subject-specific coefficients are of interest or conditioning discards too much information.
   - treat b_i as unobserved random variables and integrate them out to get the marginal likelihood of β.
   - The random effects b_i are independent and identically distributed with mean 0 and variance $D(\alpha)$. Its distribution G is completely specified with parameter α. That is, G does not depend on any covariates.

Sufficiency and Conditional Inference

- Suppose a random vector Y has a density with parameter θ, and $s = s(y)$ is a statistic. Then s is said to be sufficient for θ if

$$f(y;\theta) \propto f(s;\theta) f(y;s).$$

  The inference for θ can be based on the marginal density of s and no information is lost. The conditional density $f(y;s)$ is useful for model checking but not in inference for θ.

- If $f(y;\theta) \propto f(s|t;\theta) f(t)$, then $t = t(y)$ is said to be ancillary for θ. In this case, the conditional density $f(s|t;\theta)$ is used for inference about θ.

- When there is a nuisance parameter λ, an extension of the factorization is

$$f(s;\theta,\lambda) = f(s_1|s_2;\theta) f(s_2;\lambda).$$

  – s_2 is sufficient for λ and is ancillary for θ.

– s_1 is conditionally or partially sufficient for θ.

We can use the conditional density $f(s_1|s_2;\theta)$ for inference about θ and the marginal density $f(s_2;\lambda)$ for inference about λ. Often we only have

$$f(y;\theta,\lambda) = f(y|s_2;\theta)f(s_2;\theta,\lambda).$$

We can still use $f(y|s_2;\theta)$ for inference (for more pragmatic reasons) about θ but there is potential information in $f(s_2;\theta,\lambda)$ about θ.

Conditional Likelihood

- Preliminary
  - We will consider the binary (Bernoulli) and count (Poisson) data.
  - Assume $a(\phi) = 1$ (no overdispersion) to simplify the discussion.
  - Restrict to the canonical link, thus

$$\theta_{ij} = \eta_{ij} = g(\mu_{ij}) = x_{ij}^T\beta + z_{ij}^T b_i.$$

- Sufficient Statistics
  Treating b_i as fixed (as parameters), the likelihood function for β for individual i is

$$\prod_{j=1}^{n_i} f(y_{ij}|\beta, b_i) \propto \prod_{j=1}^{n_i} \exp\left\{\theta_{ij} y_{ij} - \psi(\theta_{ij})\right\}$$

$$= \exp\left\{\beta^T \sum_{j=1}^{n_i} x_{ij} y_{ij} + b_i^T \sum_{j=1}^{n_i} z_{ij} y_{ij} - \sum_{j=1}^{n_i} \psi(\theta_{ij})\right\} \quad (1)$$

  Hence the sufficient statistic for b_i is

$$T_i = \sum_{j=1}^{n_i} z_{ij} Y_{ij}.$$

  Let

$$S_i = \sum_{j=1}^{n_i} x_{ij} Y_{ij} \quad \text{(sufficient for } \beta)$$

- Conditional likelihood

The conditional distribution of $y_i|T_i = t$ is

$$\begin{aligned} p(y_i|T_i = t_i) &= \frac{p(Y_i = y_i, T_i = t_i)}{p(T_i = t_i)} \\ &= \frac{\exp\left(\beta^T s_i + b_i^T t\right)}{\sum_{y_i^* \in R_{t,i}} \exp\left(\beta^T s_i + b_i^T t\right)} \\ &= \frac{\exp\left(\beta^T s_i\right)}{\sum_{y_i^* \in R_{t,i}} \exp\left(\beta^T s_i^*\right)} \end{aligned}$$

where $R_{t,i} = \{(y_{i1}, \cdots, y_{in_i}) : T_i = t\}$, that is, the set of outcomes for which the statistic T_i takes the value t (Note: T_i depends on i only through covariates Z_i).

For all the data, the conditional likelihood is proportional to

$$\prod_{i=1}^{m} \frac{\exp\left(\beta^T s_i\right)}{\sum_{y_i^* \in R_{t,i}} \exp\left(\beta^T s_i^*\right)}.$$

– The conditional likelihood uses part of the data that does not contain information about

$(b_1, \cdots, b_m)$ to estimate β.

- For simple cases such as the random intercept model, the conditional likelihood is relatively easy to maximize.
- It is not necessary to specify the distribution of b_i.
- When the distribution of b_i depends on covariates, an important assumption for random effects model is violated. In the case of a random intercept model, using the conditional likelihood will still give a consistant estimate of β.

- Random Intercept Model

  In the random intercept model, the linear predictor is

$$\eta_{ij} = \gamma_i + x_{ij}^T \beta \quad (+ \text{ offset if necessary})$$

  where $\gamma_i = \beta_0 + b_i$ and x_{ij} does not include an intercept term.

The sufficient statistic for γ_i is

$$T_i = \sum_{j=1}^{n_i} Y_{ij} = Y_{i+}.$$

Conditional likelihood for β is proportional to

$$\prod_{i=1}^{m} \frac{\exp(\sum_{j=1}^{n_i} y_{ij} x_{ij}^T \beta)}{\sum_{\sum_j y'_{ij} = y_{i+}} \exp(\sum_{j=1}^{n_i} y'_{ij} x_{ij}^T \beta)} \tag{2}$$

Logistic Regression for Binary Responses

- 2×2 Crossover Trial
  For $n_i = 2$, $i = 1, \cdots, m$, compare two treatments: A (active drug), B (placebo)
  Responses: 1 for a normal electrocardiogram reading; 0 for an abnormal reading

	$y_1 = 1$ $y_2 = 1$	$y_1 = 0$ $y_2 = 1$	$y_1 = 1$ $y_2 = 0$	$y_1 = 0$ $y_2 = 0$
AB	22	0	2	6
BA	18	4	6	9

If $y_{i+} = 2$ or 0 (2 successes or no success in two trials), R_{i2} has a single element

$$(y_{i1}, y_{i2}) = (1, 1) \text{ or } (0, 0)$$

and the contribution to (2) is 1. Therefore, we do not need to consider responses (1,1) or (0,0) in calculating the conditional likelihood.

Let

$$x_1 = \begin{cases} 1, & \text{if A (active);} \\ 0, & \text{if B (placebo).} \end{cases} \quad x_2 = \begin{cases} 1, & \text{period 2;} \\ 0, & \text{period 1.} \end{cases} \quad x_3 = x_1 x_2.$$

The conditional likelihood for β is

$$\begin{aligned} L(\beta) &= \prod_{i:y_{i1}=1} \frac{\exp\left(x_{i1}^T\beta\right)}{\exp\left(x_{i1}^T\beta\right)+\exp\left(x_{i2}^T\beta\right)} \\ &\quad \times \prod_{i:y_{i2}=1} \frac{\exp\left(x_{i2}^T\beta\right)}{\exp\left(x_{i1}^T\beta\right)+\exp\left(x_{i2}^T\beta\right)}. \end{aligned} \quad (3)$$

Write

		Group AB Period 1 (A)	
		1	0
Period 2 (B)	1	a_1	c_1
	0	b_1	d_1

		Group BA	
		Period 1 (B)	
		1	0
Period 2 (A)	1	a_2	c_2
	0	b_2	d_2

a_1, d_1, a_2 and d_2 do not contribute to the conditional likelihood.

- Values of $\sum_{j=1}^{2} x_{ij}^T y_{ij}$.
  - We can ignore calculating these for $(y_{i1}, y_{i2}) = (1,1)$ or $(0,0)$.
  - For $i \in$ group AB,

$$X_i = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

So,

$$x_{i1}^T \beta = \beta_1,$$
$$x_{i2}^T \beta = \beta_2.$$

and there are c_1 and b_1 of those terms.

– For $i \in$ group BA,

$$X_i = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 1 & 1 \end{pmatrix}$$

So,

$$x_{i1}^T \beta = 0,$$
$$x_{i2}^T \beta = \beta_1 + \beta_2 + \beta_3$$

and there are c_2 and b_2 of those terms.

– Conditional likelihood is proportional to

$$\left(\frac{e^{\beta_2}}{e^{\beta_1}+e^{\beta_2}}\right)^{c_1}\left(\frac{e^{\beta_1}}{e^{\beta_1}+e^{\beta_2}}\right)^{b_1} \quad \text{(AB group)}$$
$$\times\left(\frac{e^{\beta_1+\beta_2+\beta_3}}{1+e^{\beta_1+\beta_2+\beta_3}}\right)^{b_2}\left(\frac{1}{1+e^{\beta_1+\beta_2+\beta_3}}\right)^{c_2} \quad \text{(BA group)}$$

– Special case: When $\beta_2 = \beta_3 = 0$. i.e., only consider the treatment effect, the conditional

likelihood reduces to

$$\begin{aligned} & \left(\frac{1}{1+e^{\beta 1}}\right)^{c_1}\left(\frac{e^{\beta 1}}{1+e^{\beta 1}}\right)^{b_1}\left(\frac{e^{\beta 1}}{1+e^{\beta 1}}\right)^{b_2}\left(\frac{1}{1+e^{\beta 1}}\right)^{c_2} \\ = \quad & \left(\frac{e^{\beta 1}}{1+e^{\beta 1}}\right)^{b_1+b_2}\left(\frac{1}{1+e^{\beta 1}}\right)^{c_1+c_2}. \end{aligned}$$

Therefore,

$$\hat{\beta}_1 = \log\left(\frac{b_1+b_2}{c_1+c_2}\right) = \log\left(\frac{6+4}{0+2}\right) = 1.61,$$

$$\hat{se}(\hat{\beta}_1) = \sqrt{\frac{1}{b_1+b_2}+\frac{1}{c_1+c_2}} = 0.77.$$

– When $\beta_3 = 0$, the conditional likelihood is

$$\begin{aligned} & \left(\frac{e^{\beta 2}}{e^{\beta 1}+e^{\beta 2}}\right)^{c_1}\left(\frac{e^{\beta 1}}{e^{\beta 1}+e^{\beta 2}}\right)^{b_1}\left(\frac{e^{\beta 1+\beta 2}}{1+e^{\beta 1+\beta 2}}\right)^{b_2}\left(\frac{1}{1+e^{\beta 1+\beta 2}}\right)^{c_2} \\ = \quad & (1-p_1)^{c_1}p_1^{b_1}p_2^{b_2}(1-p_2)^{c_2} \end{aligned}$$

where $\text{logit}(p_1) = \beta_1 - \beta_2$, $\text{logit}(p_2) = \beta_1 + \beta_2$.

Then

$$\hat{p}_1 = \frac{b_1}{b_1 + c_1}, \quad \hat{p}_2 = \frac{b_2}{b_2 + c_2},$$

$$\hat{\beta}_1 = \frac{1}{2}\left(\text{logit}\hat{p}_1 + \text{logit}\hat{p}_2\right) = \frac{1}{2}\log\left(\frac{b_1 b_2}{c_1 c_2}\right),$$

$$\hat{se}(\hat{\beta}_1) = \frac{1}{2}\sqrt{b_1^{-1} + c_1^{-1} + b_2^{-1} + c_2^{-1}}.$$

Since $c_1 = 0$, use ad hoc convention of adding 0.5 to give

$$\hat{\beta}_1 = \frac{1}{2}\log\left(\frac{6 \times 4}{0.5 \times 2}\right) = 1.59,$$

$$\hat{se}(\hat{\beta}_1) = 0.85.$$

– Note

1. In both models, $\hat{\beta}_1 \approx 1.6$ is marginally significant at a (2-sided) 5% level.
2. $\exp(1.6) \approx 5$ indicates the odds of a normal result for a treated patient is about 5 times of the odds for a non-treated patient.

3. Compare with marginal model where

$$\hat{\beta}_1 \approx 0.57, \quad e^{\hat{\beta}_1} \approx 1.77$$

compare to Theorem (Neuhaus): $|0.57| \leq |1.59|$.

Example: Conditional logistic Regression

Poisson Regression for Count Responses

$y_i = (y_{i1}, \cdots, y_{in_i})$ are independent Poisson r.v.'s with

$$\log E(y_{ij}|\gamma_i, \beta) = \gamma_i + x_{ij}^T \beta + \log(t_{ij}),$$

where $\gamma_i = \beta_0 + b_i$ and x_{ij} does not include the intercept term. The likelihood contributed by the $i-$th individual is proportional to

$$\exp\left\{\gamma_i \sum_j y_{ij} + \beta^T \sum_j x_{ij} y_{ij} + \sum_j y_{ij} \log(t_{ij}) - \sum_j t_{ij} \exp\left(\gamma_i + x_{ij}^T \beta\right)\right\} \frac{1}{\Pi_j \, y_{ij}!}.$$

Then $y_{i+} = \sum_j y_{ij}$ is a sufficient statistic for γ_i.

The distribution of y_{i+} is Poisson with mean

$$\sum_j e^{\gamma_i + x_{ij}^T \beta + \log t_{ij}} = \lambda_i.$$

Let $s_i = \sum_j x_{ij} y_{ij}$. Then the conditional distribution of y_i given y_{i+} is

$$\begin{aligned} P(y_i|y_{i+}) &= \frac{\exp\left\{\gamma_i y_{i+} + \beta^T s_i + \sum_j y_{ij} \log(t_{ij}) - \sum_j t_{ij} \exp\left(\gamma_i + x_{ij}^T \beta\right)\right\}}{\left(\prod_j y_{ij}!\right) \lambda_i^{y_{i+}} \exp(-\lambda_i)/y_{i+}!} \\ &= \binom{y_{i+}}{y_{i1}, \cdots, y_{in_i}} \frac{\exp\left\{\beta^T s_i + \sum_j y_{ij} \log(t_{ij})\right\}}{\left(\sum_j \exp\left\{x_{ij}^T \beta + \log(t_{ij})\right\}\right)^{y_{i+}}} \end{aligned}$$

Example: Seizure Data
The model is

$$\log E(y_{ij}|\gamma_i, \beta) = \gamma_i + \beta_1 x_{ij1} + \beta_2 x_{ij2} + \beta_3 x_{ij3} + \log t_{ij},$$

where $x_{ij1} = 1$ or 0 for progabide or placebo, respectively; $x_{ij2} = 1$ if $j = 1, 2, 3, 4$ or 0 if $j = 0$;

$x_{ij2} = x_{ij1}x_{ij2}$ (interaction term). Note that β_3 is the parameter of interest. We also note that e^{γ_i} is the expected baseline count for individual i.

- For the placebo group,

$$x_{ij}^T = \begin{cases} (0,0,0), & \text{if } j = 0; \\ (0,1,0), & \text{if } j = 1,2,3,4. \end{cases}$$

- For the treatment group,

$$x_{ij}^T = \begin{cases} (1,0,0), & \text{if } j = 0; \\ (1,1,1), & \text{if } j = 1,2,3,4. \end{cases}$$

Thus

$$\begin{aligned} s_i &= \sum_{j=0}^{4} x_{ij}y_{ij} \\ &= x_{i0}y_{i0} + x_{i1}\sum_{j=1}^{4} y_{ij} \\ &= x_{i0}y_{i0} + x_{i1}(y_{i+} - y_{i0}) \end{aligned}$$

If $i \in$ placebo group,

$$s_i^T \beta = \beta_2 (y_{i+} - y_{i0}),$$

otherwise it equals

$$y_{i0}\beta_1 + (y_{i+} - y_{i0})(\beta_1 + \beta_2 + \beta_3) = (y_{i+} - y_{i0})(\beta_2 + \beta_3) + y_{i+}\beta_1.$$

The conditional likelihood is proportional to

$$\prod_i \frac{\exp(s_i^T \beta) \exp(\sum_j y_{ij} \log(t_{ij}))}{\left(\sum_{j=0}^4 \exp\{x_{ij}\beta + \log t_{ij}\}\right)^{y_{i+}}}.$$

Here we have $t_{i0} = 8$ and $t_{i0} = 8$ and $t_{i1} = t_{i2} = t_{i3} = t_{i4} = 2$.

- For $i \in$ placebo group, the likelihood is proportional to

$$\frac{e^{y_{i+}\beta_1} e^{(\beta_2+\beta_3)(y_{i+}-y_{i0})}}{e^{y_{i+}\beta_1}\left(1 + e^{\beta_2+\beta_3}\right)^{y_{i+}}} = \pi_2^{y_{i0}} (1 - \pi_2)^{y_{i+}-y_{i0}}$$

where $\pi_2 = \frac{1}{1+e^{\beta_2+\beta_3}}$.

- In summary, the conditional likelihood is proportional to

$$\prod_{i=1}^{28} \pi_1^{y_{i0}}(1-\pi_1)^{y_{i+}-y_{i0}} \prod_{i=29}^{59} \pi_2^{y_{i0}}(1-\pi_2)^{y_{i+}-y_{i0}}.$$

- π_1 and π_2 are the probabilities that an individual's seizure occurs before rather than after the randomization, for placebo and progabide groups, respectively.

- If progabide is helpful in reducing seizures we would observe $\pi_1 < \pi_2$, or equivalently, $1 + e^{\beta_2} > 1 + e^{\beta_2+\beta_3} \quad \Leftrightarrow \quad \beta_3 < 0$.

- Results (patient 207 deleted)

	GEE	Cond. Like.
β_2	0.11 (0.12)	0.11 (0.047)
β_3	-0.30 (0.17)	-0.30 (0.07)

Notes:

1. Conditional likelihood inference leads to conclusion that progabide's effect is highly significant.
2. But (see DHLZ) the fitted model is grossly inadequate on a Pearson χ^2 statistic.
3. Because of (2) the estimated s.e.'s in conditional likelihood approach may be too small in this example.
4. The homogeneity assumption that everyone's response to the treatment is the same is inadequate, that is, a random slope is needed.

Maximum Likelihood for GLMM

We need specify a distribution for b_i to use maximum likelihood estimation. For generalized linear mixed models, we assume

$$b_i \sim^{iid} N(0, D(\alpha))$$

where α is the variance parameter for the random effects.

The joint likelihood for (β, α) is the marginal distribution of Y

$$L(\beta, \alpha) = \prod_{i=1}^{m} \int \prod_{j=1}^{n_i} f(y_{ij}|\beta, b_i)\phi(b_i|\alpha)db_i$$

where ϕ is the multivariate normal density function. Note that there are q levels of integral (q is the dimension of b_i).

For f in exponential family, we have

$$L(\beta, \alpha; y_i) = (2\pi)^{-q/2}|D|^{-1/2} \exp\left\{1^T c(y_i)\right\} J(\beta, \alpha)$$

where

$$J(\beta,\alpha)=\int_{R^q}\exp\left\{y_i^T(X_\beta+Z_ib_i)-1^Tb_i(X_i\beta+Z_ib_i)-\frac{1}{2}b_i^TD^{-1}b_i\right\}db_i.$$

Here, b_i refers to the $b(\theta)$ part of exponential family density. In general, the marginal likelihood has no closed form and the integration is quite difficult.

Estimation Methods for GLMM

- Generalized Estimating Equations (GEE)

- Numerical evaluation of the likelihood

  1. Gauss-Hermite quadrature.
  2. Adaptive quadrature (SAS PROC NLMIXED).
  3. Importance sampling and Monte Carlo integration.

- Numerical maximization of the likelihood (without evaluating it).

  1. Expectation-maximization (EM) algorithm.
  2. Monte Carlo EM.

3. Monte Carlo Newton-Raphson.

- Approximate likelihood

1. Penalized quasi-likelihood (PQL), Bias-corrected PQL.
2. Linearization.

- Bayesian Markov chain Monte Carlo (Zeger and Karim, 1991; Clayton, 1996).

1. Gibbs sampling.
2. Metropolis-Hastings algorithm.

GEE for GLMM

An obvious, though by no means easier, alternative to maximizing the marginal likelihood for GLMM, which often has no closed form, is to derive the estimating equation for the conditional coefficients β (Zeger et al., 1998):

$$U(\beta) = \sum_{i=1}^{m} \left(\frac{\partial \mu_i^M}{\partial \beta} \right)^T V_i^{-1}(\alpha)(Y_i - \mu_i^M)$$

where

$$\mu_{ij}^M = E(Y_{ij}) = E\left[E(Y_{ij}|b_i)\right] = \int \mu_{ij} dF(b_i),$$

$\mu_{ij} = g^{-1}(X_{ij}^T\beta + Z_{ij}^T b_i)$ is the conditional mean

$$V_i = cov\left[E(Y_i|b_i)\right] + E\left[cov(Y_i|b_i)\right].$$

For binary data, the (j,k) element of the marginal variance matrix is

$$\begin{aligned}[V_i]_{jk} &= \int (\mu_{ij} - \mu_{ij}^M)(\mu_{ik} - \mu_{ik}^M) dF(b_i) \\ &\quad + I_{j=k} \int \mu_{ij}(1-\mu_{ij}) dF(b_i).\end{aligned}$$

Note that the distribution F of the random effects may involve additional unknown parameters. Here they are treated as fixed and known. If F is Gaussian with mean 0 and variance matrix D, these expression simplifies (identity link) or can be approximated.

Example: GEE for Logistic-Gaussian Models.

For binary data with probit link,

$$\Phi^{-1}(\mu_{ij}^M) = a_p(D) X_{ij}^T \beta$$

where $a_p(D) = |DZ_{ij}Z_{ij}^T + I|^{-q/2}$ and q is the dimension of b_i.

For logit link, no closed form exists for μ_{ij}^M, but it can exist by approximation

$$\text{logit}\left(\mu_{ij}^M\right) \approx a_l(D) X_{ij}^T \beta$$

where $a_l(D) = |cDZ_{ij}Z_{ij}^T + I|^{-q/2}$ and $c = 16\sqrt{3}/(15\pi)$.

Except for identity link, no simple formula exists for V_i. However, because of the natural of GEE, only an approximation is needed. Zeger et al. (1988) used Taylor expansion about $b_i = 0$ to get an approximation. Zeger et al. (1988) also argues that estimation of the unknown parameters in F (or D), which are not asymptotically orthogonal to β, is not crucial for inference about β for Gaussian random effects.

Numerical Evaluation of the Likelihood

·Gauss-Hermite Quadrature

The Gauss-Hermite quadrature uses a fixed set of K ordinates and weights (z_k, w_k) to approximate an integral:

$$\int_{-\infty}^{\infty} g(x)e^{-x^2}dx \approx \sum_i w_i g(z_i).$$

Consider a single scalar random effect $b_i \sim N(0, \tau^2)$. Then the GLMM likelihood for subject i is given by

$$\begin{aligned} L(\beta, \tau|y_i) &= \int \left\{ \prod_{j=1}^{n_i} f(y_{ij}|\beta, b_i) \right\} f(b_i; \tau) db_i \\ &= \int \left[\exp\left\{ \sum_j \log f(y_{ij}|\beta, b_i) \right\} \right] \frac{1}{\tau}\phi\left(\frac{b_i}{\tau}\right) db_i \\ &\approx \sum_{k=1}^{K} w_k \left[\exp\left\{ \sum_j \log f(y_{ij}|\beta, b_i = z_k/\tau) \right\} \right]. \end{aligned}$$

Note

- Gauss-Hermite quadrature is less accurate when the τ increases.

- There are situations where a 20-point G-H quadrature may fail.

·Adaptive Gauss Quadrature

The adaptive Gaussian quadrature centers the Gaussian approximation at the posterior mode of the random effects. For subject i,

$$\begin{aligned}
L(\beta,\tau|y_i) &= \int \left[\exp\left\{\sum_j \log f(y_{ij}|\beta,b_i)\right\}\right] \frac{1}{\tau}\phi\left(\frac{b_i}{\tau}\right) db_i \\
&= \int \left[\exp\left\{\sum_j \log f(y_{ij}|\beta,b_i)\right\}\right] \frac{1}{\tau}\frac{\phi\left(\frac{b_i}{\tau}\right)}{\phi\left(\frac{b_i-a}{b}\right)}\phi\left(\frac{b_i-a}{b}\right) db_i \\
&= \int \left[\exp\left\{\sum_j \log f(y_{ij}|\beta,b_i=a+bz)\right\}\right] \frac{1}{\tau}\frac{\phi\left(\frac{a+bz}{\tau}\right)}{b\phi\left(z\right)}\phi\left(z\right) dz \\
&\approx \sum_{k=1}^{K} w_k \exp\left\{\sum_j \log f(y_{ij}|\beta,b_i=a+bz_k)\right\} \frac{1}{\tau}\frac{\phi\left(\frac{a+bz_k}{\tau}\right)}{b\phi\left(z_k\right)}.
\end{aligned}$$

In the adaptive quadrature method, different values a_i and b_i are used for each subject where a_i is the posterior mode and b_i is the approximate posterior curvature. Adaptive quadrature requires less points to achieve the same accuracy as fixed-point quadrature.

But the quadrature method do not perform very well for higher-dimension integrations.

·Monte Carlo Methods for Integration

Suppose we want to calculate this integral

$$E(h(X)) = \int h(x)f(x)dx$$

where X can be a vector and f is a density function.

- Naive Monte Carlo
  If we can draw a random iid sample $x^{(1)}, x^{(2)}, \cdots, x^{(M)}$ from f, then

  $$\int h(x)f(x)dx \approx \frac{1}{M}\sum_{j=1}^{M} h(x^{(j)}).$$

  It also works when dependent samples are drawn from a Markov chain with stationary distribution $f(x)$ (Markov chain Monte Carlo).

- Importance Sampling
  It is often difficult or impossible to draw independent

samples from an arbitrary distribution f (which may not even have a closed form). We can find another density function g whose support includes that of f. Since

$$\int h(x)f(x)dx = \int \frac{h(x)f(x)}{g(x)}g(x)dx,$$

we can draw a random iid sample $x^{(1)}, x^{(2)}, \cdots, x^{(M)}$ from g, then

$$\int h(x)f(x)dx \approx \frac{1}{M}\sum_{j=1}^{M} w_j h(x^{(j)})$$

where $w_j = \frac{f(x^{(j)})}{g(x^{(j)})}$.
while the algorithm works for any g, the efficiency depends on the choice of g (the variance of w_j).

·Simulated Maximum Likelihood for GLMM

Geyer and Thompson (1992) and Gelfand and Carlin (1993) suggested simulation to estimate the GLMM

likelihood which can be numerically maximized

$$\begin{aligned} L(\beta, D|y) &= \int f(y|\beta, b) f(b|D) db \\ &= \int \frac{f(y|\beta, b) f(b|D)}{g(b)} g(b) db \\ &\approx \frac{1}{M} \sum_{k=1}^{M} \frac{f(y|\beta, b^{(k)}) f(b^{(k)}|D)}{g(b^{(k)})} \end{aligned}$$

where $b^{(k)}$ is drawn from the importance sampling-distribution $g(b)$.

- To maximize the likelihood, we need to evaluate it at different values of (β, D). The advantage of importance sampling method is that only one sample from g is needed.

- To improve Monte Carlo efficiency, multiple g's that depend on the value of (β, D) can be used.

- When g itself is difficult to draw independent samples from, Markov chain Monte Carlo (MCMC)

can be used to draw dependent samples from a Markov chain with g as stationary distribution.

Expectation-Maximization Algorithm for GLMM

Treat the random effects as missing data. Then the complete data is (y, b). The complete data log-likelihood is given by

$$l^c(\beta, D) = \sum_i \{\log f(y_i|\beta, b_i) + \log f(b_i|D)\} .$$

E-step: evaluate the expected complete data log-likelihood

$$E(l^c|y)$$

where the expectation is taken with regard to the conditional distribution $f(b|y)$ using current estimates of $\beta^{(m)}$ and $D^{(m)}$.

M-step: maximize the expected complete data log-likelihood. Since the parameters (β, D) are separated into two terms, the maximization can be done separately

$$\beta^{(m+1)} = \text{arg}_\beta\text{max} \sum_i E \{\log f(y_i|\beta, b_i)|y\} ,$$

$$D^{(m+1)} = \text{arg}_D\text{max} \sum_i E \{\log f(b_i|D)|y\} .$$

For GLMM, the EM algorithm involves solving estimating equations iteratively:

$$0 = \sum_{i=1}^{m} \sum_{j=1}^{n_i} x_{ij} \left\{ y_{ij} - E\left(g^{-1}(X_i\beta + Z_i b_i) | y_i \right) \right\}, \quad (4)$$

$$0 = \frac{1}{2} D^{-1} \left\{ \sum_{i=1}^{m} E(b_i b_i^T | y_i) \right\} D^{-1} - \frac{m}{2} D^{-1}. \quad (5)$$

Monte Carlo EM

The conditional expectation required in the EM algorithm typically does not have a closed form since the conditional distribution is

$$f(b|y) = \frac{f(y,b)}{\int f(y,b) db}$$

and the integration in the denominator is what we want to avoid.

·Monte Carlo EM

We can draw dependent samples from $f(b|y)$ using

Metropolis algorithm without calculating $f(y)$ and use Monte Carlo method to calculate the conditional expectation (McCulloch, 1997). Using $f(b|D)$ as proposal distribution, the Metropolis-Hastings ratio becomes

$$min\left\{1, \frac{\prod_i f(y_i|b^*, \beta)}{\prod_i f(y_i|b, \beta)}\right\}.$$

Monte Carlo Newton-Raphson

Note that in EM, we can use Newton-Raphson or Fisher scoring method to solve the score equation iteratively:

$$\beta^{(k+1)} = \beta^{(k)} + \left\{\left(\frac{\partial \mu^M}{\partial \beta}\right)^T V^{-1}\left(\frac{\partial \mu^M}{\partial \beta}\right)\right\}^{-1}\left(\frac{\partial \mu^M}{\partial \beta}\right)^T V^{-1}(y - \mu^N). \quad (6)$$

For GLMM, we have

$$\mu^M = E_{b|y}(\mu).$$

So

$$\begin{aligned}\frac{\partial \mu^M}{\partial \beta} &= \frac{\partial E(\mu)}{\partial \beta} = E\left(\frac{\partial \mu}{\partial \beta}\right) = E\left(\frac{\partial \mu}{\partial \eta}\frac{\partial \eta}{\partial \beta}\right) \\ &= E\left(\frac{\partial \mu}{\partial \eta}\frac{\partial (X\beta + Zb)}{\partial \beta}\right) \\ &= E\left(\frac{\partial \mu}{\partial \eta}\right) X.\end{aligned}$$

So we

$$\begin{aligned}\beta^{(j+1)} &= \beta^{(j)} + E_{b|y}\left[X^T W\left(\beta^{(j)}, b\right) X\right]^{-1} X^T E_{b|y}\left[W\left(\beta^{(j)}, b\right) \frac{\partial \eta}{\partial \mu}|_{\beta^{(j)}}\right] \\ &\quad \times \left(y - E_{b|y}\left(\mu(\beta^{(j)}, b)\right)\right), \qquad (7)\end{aligned}$$

where

$$W(\beta^{(j)}, b) = \left(\frac{\partial \mu}{\partial \eta}\right)^T V^{-1} \left(\frac{\partial \mu}{\partial \eta}\right)|_{\beta^{(j)}}.$$

The conditional expectation can again be calculated using Markov chain Monte Carlo.

Note

- When the likelihood is not unimodal, EM may converge to the wrong place and Newton-Raphson may not converge.

- For accurate estimates, many iterations are required. It might be desirable to combine difference approaches.

Penalized Quasi-Likelihood (PQL)

Use conditional models rather than conditional means in the score function and thus avoid integrations.

- This is equivalent to approximating the conditional distribution of b_i given y_i by a Gaussian distribution with the same mode and curvature (incorporated into the M-step).

- In other words, we plug in the posterior mode or BLUP $\hat{b}$ for b. From another perspective, consider the random effects b as fixed parameters, we can maximize the joint likelihood with respect to (β, b)

$$\log f(y|\beta, b) - \frac{1}{2} b^T D^{-1} b.$$

This penalized likelihood idea is similar to REML, and this approach is called penalized quasi-likelihood (Breslow and Clayton, 1993). Direct differentiation of the penalized likelihood leads to the score equations for (β, b) (assuming canonical exponential family).

$$\begin{pmatrix} X^T(y-\mu) \\ Z^T(y-\mu) - D^{-1}b \end{pmatrix} = 0,$$

where $g(\mu) = X\beta + Zb$. Alternating solving these two sets of score equations gives the previous algorithm which can be implemented easily in standard software. In PQL, only the mean and variance need to be specified in the conditional mean model of $y|b$.

PQL and Laplace Approximation

More formally, PQL can be derived via Laplace approximation to the GLMM likelihood

$$L(\beta, D) \propto |D|^{-1/2} \int_{R^q} \exp(h(b))db$$

where

$$h(b) = y^T(X\beta + Zb) - 1^T b(X\beta + Zb) - \frac{1}{2}b^T D^{-1} b.$$

The Laplace approximation of $L(\beta, D)$ starts with the Taylor series approximation of $h(b)$

$$h(b) \approx h(b_0) + \dot{h}(b_0)(b - b_0) + \frac{1}{2}(b - b_0)^T \ddot{h}(b_0)(b - b_0).$$

We can (for example) choose b_0 to solve

$$\dot{h}(b_0) = 0. \text{ i.e., mode.}$$

When then have

$$h(b) \approx h(b_0) + \frac{1}{2}(b - b_0)^T \ddot{h}(b_0)(b - b_0).$$

Note that

$$\int_{R^q} \frac{1}{(2\pi)^{q/2}} |\Sigma|^{-1/2} \exp\left\{ -\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu) \right\} dx = 1.$$

We have (up to a constant),

$$L(\beta, D) \approx |D|^{-1/2}|-\ddot{h}(b_0)|^{-1/2}\exp(h(b_0)).$$

Here

$$\dot{h}(b) = \left\{y - \dot{b}(X\beta + Zb)\right\}^T Z - b^T D^{-1},$$
$$\ddot{h}(b) = -Z^T diag(\ddot{b}(X\beta + Zb))Z - D^{-1}.$$

The resulting approximation (up to a constant) to the log-likelihood is

$$\begin{aligned} l(\beta, D) \quad &\approx \quad y^T(X\beta + Zb_0) - 1^T b(X\beta + Zb_0) - \frac{1}{2}b_0^T D^{-1} b_0 \\ &\quad -\frac{1}{2}\log|I + DZ^T diag(\ddot{b}(X\beta + Zb_0))Z|. \end{aligned} \tag{8}$$

PQL uses an additional approximation by assuming $\ddot{b}(X\beta + Zb_0)$ to be relatively constant with respect to be β. Hence it can be dropped and we are left with

$$l(\beta, D) \approx h(b_0) = \log f(y|\beta, b_0).$$

- Expand h at the true mean of b leads to another approximation of the likelihood.

- The approximation adopted by PQL induces bias in the estimates of β. Better approximating can be achieved by considering high-order expansions. Breslow and Lin (1995) and Lin and Breslow (1996) derived the bias-corrected estimates.

Linearization

Another approximation approach is based on the idea of linearization. The data Y is decomposed into the mean μ which is a non-linear function of the linear predictor and an appropriate error term

$$\begin{aligned} E(Y|b) &= \mu = g^{-1}(X\beta + Zb) = g^{-1}(\eta), \\ var(Y|b) &= C^{1/2}RC^{1/2}, \\ b &\sim N(0, G). \end{aligned}$$

Taylor series expansion of μ about $\tilde{beta}$ and $\tilde{b}$ is

$$g^{-1}(\eta) \approx g^{-1}(\tilde{\eta}) + \tilde{\Delta}X(\beta - \tilde{\beta}) + \tilde{\Delta}Z(b - \tilde{b}) \qquad (9)$$

where $\tilde{\Delta} = \frac{\partial}{\partial \eta} g^{-1}(\eta)|_{\tilde{\beta},\tilde{b}}$. Rearranging (9) is

$$\tilde{\Delta}^{-1}(\mu - g^{-1}(\tilde{\eta})) + (X\tilde{\beta} + Z\tilde{b}) \approx X\beta + Zb.$$

Now define pseudo-data Y^* as

$$Y^* = \tilde{\Delta}^{-1}(Y - g^{-1}(\tilde{\eta})) + (X\tilde{\beta} + Z\tilde{b}).$$

Then we have

$$\begin{aligned} E(Y^*|b) &= X\beta + Zb, \\ var(Y^*|b) &= \tilde{\Delta}^{-1} C^{1/2} R C^{1/2} \tilde{\Delta}^{-1}. \end{aligned}$$

Thus, we can consider the linear mixed model:

$$Y^* = X\beta + Zb + \epsilon$$

where ϵ has a Gaussian distribution. The model on pseudo-data can be fitted using ML or REML.
Double-iterative algorithm:

- Start with $\tilde{\beta}$ and $\tilde{b}$, compute the pseudo data, and fit the linear mixed model, which may itself require an iterative algorithm.

- Update $\tilde{\beta}$ and $\tilde{b}$.

Choice of $\tilde{\beta}$ and $\tilde{b}$:

- Penalized quasi-likelihood, subject-specific expansion: BLUP of the random effects

$$\tilde{\beta} = \hat{\beta}, \quad \tilde{b} = \hat{b}.$$

- Marginal quasi-likelihood, population-average expansion:

$$\tilde{\beta} = \hat{\beta}, \quad \tilde{b} = \hat{b}.$$

The full likelihood of Y is not required (hence, 'quasi-likelihood'). A model without random effects can be fitted, which specifies a marginal model like in GEE (but the estimates are different). Accuracy depends on the distribution assumption of the pseudo-data. There can be substantial bias for binary data with relatively small number of repeated observations (not "sufficiently continuous").

Fitting GLMM: Software

- SAS
  - PROC NLMIXED uses Gauss-Hermite or Adaptive Quadrature methods.
  - PROC GLIMMIX uses linearization idea to fit penalized/marginal quasi-likelihood models with Gaussian random effect.

- R
  - glmmPQL (MASS): penalized quasi-likelihood, allows the use of an additional correlation structure.
  - GLMM (lme4 a newer reimplementation of nlme): similar to (same as?) glmmPQL.
  - glmm (repeated): Gauss-Hermite quadrature, models with random intercept.
  - gnlmix (repeated): non-linear regression with mixed random effects for the location parameters. Non-Gaussian mixing distributions are allowed.
  - glmm (GLMMGibbs): Gibbs sampling.

- BUGS (OpenBUGS/WinBUGS)