Chapter 8

Random Effects Models for Repeated Binary Data

The models and methods for repeated binary data which were considered in Chapter 7 are most appropriate when the data are balanced, that is, there are n common occasions of measurement and imbalance ($n_i \neq n$ for some i) arises because of missing observations. When the unequal n_i arise because of inherently unbalanced data or because of clustered designs, the most natural approach is to consider extending the LMM using random effects to the GLM setting.

By analogy to the linear case, we assume each subject has a vector of subject-specific effects, b_i, and we add $Z_i b_i$ to the linear predictor $X_i \beta$. Letting Y_i denote the $n_i \times 1$ vector of binary outcomes, we have

$$E(Y_i \mid b_i, X_i) = \mu_i^* = g(X_i \beta^* + Z_i b_i) \tag{8.1}$$

where

$$\ell\left(\mu_i^*\right) = X_i \beta^* + Z_i b_i, \tag{8.2}$$

ℓ is the link function, and g the inverse link function. As before, we assume that $E(b_i) = 0$ and $\mathrm{var}(b_i) = D$. Generally, we also assume that given b_i, the Y_{ij}'s are independent.

We use the μ_i^*, β^* notation to emphasize that μ_i^* and β^* are conditional and not marginal parameters. Recall that for μ_i, β defined in Chapters 6 and 7, we assume that

$$E(Y_i \mid X_i) = \mu_i = g(X_i \beta).$$

But here we have

$$E(Y_i \mid X_i, b_i) = \mu_i^* = g(X_i \beta^* + Z_i b_i)$$

so that

$$E(Y_i \mid X_i) = E(\mu_i^*) \neq g(X_i\beta^*)$$

for nonlinear link functions. Thus the fixed effects in this mixed effects model are not directly comparable to those in the GLM, the 'mixed' model parameterization or the GEE.

In some cases, these conditional parameters are comparable to marginal parameters. For example, with the log-link where

$$E\{E(Y_{ij} \mid X_{ij}, b_i)\} = E\left(e^{X_{ij}^T\beta^* + b_i}\right)$$

and b_i is scalar, we have

$$\begin{aligned} E(Y_{ij} \mid X_{ij}) &= e^{X_{ij}^T\beta^*}\ E(e^{b_i}) \\ &= e^{X_{ij}^T\beta^* + \ln E(e^{b_i})} \end{aligned}$$

so that $\ln E(e^{b_i})$ merely acts like a constant offset for each observation. Thus apart from the intercept, $\beta^* = \beta$.

For the logistic link function

$$E(Y_{ij} \mid X_{ij}) = E(\mu_i^*) = E\left\{\exp(X_{ij}\beta + Z_{ij}b_i)/\left(1 + \exp(X_{ij}\beta + Z_{ij}b_i)\right)\right\},$$

where expection is over the distribution of b_i. Various authors (e.g., Newhaus *et al.*, 1991, Diggle *et al.*, 1994) have shown that for the logistic link function the components of β^* are typically attenuated relative to the components of β. Because the expression for μ_i^* is an integral, in the binary setting, μ_i^* usually cannot be computed except by approximation. This makes estimation complex. We will consider two general approaches to estimating the parameters in random effects models: moment estimation and likelihood approaches. We now drop the superscript * on β and μ_i for notational simplicity, it being understood that they refer to the parameters in (8.1)–(8.2). We shall also use the following notation:

$$\begin{aligned} g'\left(X_{ij}^T\beta + Z_{ij}^Tb_i\right) &= g'(\ell_{ij}) = \partial\, g/\,\partial\ell_{ij}, \\ \partial g\left(X_{ij}^T\beta + Z_{ij}^Tb_i\right)\Big/\,\partial b_i &= g'\left(X_{ij}^T\beta + Z_{ij}^Tb_i\right)Z_{ij}^T, \end{aligned}$$

where X_{ij}^T and Z_{ij}^T are column vectors denoting the jth rows of X_i and Z_i, respectively. Further,

$$\partial g\,(X_i\beta + Z_ib_i)/\,\partial\, b_i = \operatorname{Diag}\left\{g'\left(X_{ij}\beta + Z_{ij}b_i\right)\right\}Z_i.$$

Note that for the logistic transform, $g'(\ell_{ij}) = p_{ij}(1 - p_{ij})$ where $p_{ij} = g(\ell_{ij})$. Hence

$$\partial g\left(X_i\beta + Z_ib_i\right)/\partial b_i = R_iZ_i$$

where

$$R_i = \text{ diag } \{p_{ij}(1-p_{ij})\} = \text{ diag } \left\{g'(X_{ij}\beta + Z_{ij}b_i)\right\}.$$

8.1 GEE approach to estimating β

The basic idea here is to find the marginal means and variances of Y_i in terms of the conditional β vector and D, and use a GEE-type approach to estimation. This is attractive, because as we show in Section 8.2, full likelihood approaches are numerically difficult. We here further assume that the Y_{ij}'s given b_i are independent, so that

$$\text{var } (Y_i \mid X_{ij}, b_i) = \text{diag } \{p_{ij}(1-p_{ij})\}$$

where

$$p_{ij} = P(Y_{ij} = 1 \mid X_{ij}, b_i) = g(X_{ij}\beta + Z_{ij}b_i).$$

Provided we specify the distribution of b_i, we may in principle compute

$$E(Y_i \mid X_i) = E[g(X_i\beta + Z_ib_i)]$$

and

$$\text{var } (Y_i \mid X_i) = E[\text{diag } \{p_{ij}(1-p_{ij})\}] + \text{var } p_i.$$

Notice that because of the nonlinear link function, specifying simply the moments of b_i is not generally sufficient.

For certain link functions, and distributional assumptions on b_i, e.g., probit or log and $b_i \sim N(0, D)$, it is possible to find closed form expressions for the marginal means and variances. In other cases various approximations have been used. For a probit link and $b_i \sim N(0, D)$, Zeger *et al.* (1988) show that

$$E(Y_{ij} \mid X_i) = \Phi\left(a_p(D)X_{ij}^T\beta\right)$$

where $a_p(D) = \mid DZ_{ij}Z_{ij}^T + I \mid^{-q/2}$. For the logit link there is no corresponding closed form solution, but Johnson and Kotz (1970) show that

$$E(Y_{ij} \mid X_i) \approx \text{anti logit } \{a_\ell(D)X_{ij}^T\beta\}$$

where

$$a_\ell(D) = \mid c^2\ DZ_{ij}Z_{ij}^T + I \mid$$

and

$$c^2 = 16\sqrt{3}/(15\ \pi).$$

Zeger *et al.* (1988) use this approximation plus Taylor series approximations for var $(Y_{ij} \mid X_i)$ to get GEE type estimating equations for β and D.

This same general approach has been used by Gilmour *et al.* (1985), who used the probit link and a normal assumption for b_i to get closed form solutions for the binomial case. They develop estimating equations for β and D similar to those given by Zeger *et al.* (1988).

A related approach (Goldstein, 1991) is to write a linear model as

$$Y_{ij} \doteq p_{ij} + e_{ij}$$

where for the binary case, the e_{ij}'s are independent with

$$\text{var}\,(e_{ij} \mid b_i) = p_{ij}(1 - p_{ij}).$$

Using a first order approximation for $g\,(X_i\beta + Z_i b_i)$ around $b_i = 0$, we can write

$$Y_i \doteq g\,(X_i\beta) + (\partial g/\partial b_i)\Big|_{b_i=0} b_i + e_i.$$

Letting R_{i0} denote $\text{var}(Y_i|b_i)$ with b_i evaluated at zero, when the logit link function is used we can approximate the variance of Y_i with

$$\text{Var}\,(Y_i|X_i) = R_{i0} + R_{i0}Z_iD\;Z_i^T R_{i0}.$$

A more general expression for the case where $g'(\ell_i) \neq g(\ell_i)\,(1 - g(\ell_i))$ is given by

$$\text{var}\,(Y_i|X_i) = R_{i0} + \Delta_{i0}Z_iDZ_i^T\Delta_{i0}$$

where Δ_{i0} is diagonal with $(g'(\ell_i))$ on the diagonal, and b_i is evaluated at zero.

Since now it is assumed that

$$E(Y_i) \doteq g(X_{ij}^T\beta),$$

it is straightforward to implement GEE to estimate β given D. Breslow and Clayton (1993) show that using Fisher scoring to solve GEE can be expressed, as in the LMCD case, as iteratively weighted least squares regression of $\tilde{Y}_i$ on X_i with weight matrix W_i being proportional to the inverse of $\text{var}(Y_i|X_i)$:

$$W_i = (\Delta_{i0}^{-1}R_{i0}\Delta_{i0}^{-1} + Z_iDZ_i)^{-1},$$

and $\tilde{Y}_i$ is the "working" variable $\tilde{Y}_i = X_i^T\beta + \Delta_{i0}^{-1}(Y_i - g(X_i^T\beta))$. Since the model for Y_i has been linearized, estimates for b_i can be taken as

$$\widehat{b}_i = DZ_i^TW_i\;(\tilde{Y}_i - X_i^T\widehat{\beta}).$$

Estimates of D may be obtained using methods discussed in Section 8.4.

8.2 Likelihood Approaches

In principle, one can generalize the ML normal theory approach, estimating β and D by marginal ML, that is, integrating out the b_i and maximizing the resulting likelihood:

$$L(\beta, D) = \Pi_{i=1}^{N} \int_{R^q} f(Y_i \mid \beta, b_i) f(b_i \mid D) db_i, \tag{8.3}$$

and using empirical Bayes for b_i:

$$\widehat{b_i} = E(b_i \mid Y_i, \beta, D) \mid_{\widehat{\beta}, \widehat{D}},$$

where β, D are evaluated at the MLE's. For the binary case, assuming independence given b_i, we have

$$f(Y_i \mid \beta, b_i) = \Pi_{j=1}^{n_i} p_{ij}^{Y_{ij}} (1 - p_{ij})^{1 - Y_{ij}}$$

for

$$p_{ij} = E(Y_{ij} \mid X_{ij}, b_i) = g\left(X_{ij}^T \beta + Z_{ij}^T b_i\right).$$

What makes this approach difficult is that the integral in (8.3) does not have a closed form solution in the general case, nor do the derivatives of $L(\beta, D)$. If b_i is a scalar normal random variable, so that $Z_i^T = (1, \ldots, 1)$, then Gaussian quadrature can be used to give a very good approximation to $L(\beta, D)$ using

$$L(\beta, D) = \Pi_{i=1}^{N} \sum_{l=1}^{K} h_l(Y_i, X_i\beta, D, S_l) W_l$$

where S_l and W_l are the known mass points and weights for the $N(0, 1)$ integral and depend only on the number of grid points. Here

$$h_l(Y_i, X_i\beta, D, S_l) = \Pi_{j=1}^{n_i} p_{ijl}^{Y_{ij}} (1 - p_{ijl})^{1 - Y_{ij}}$$

and

$$p_{ijl} = g\left(X_{ij}^T \beta + \sqrt{D}\, S_l\right).$$

Anderson and Aitken (1985) point out that using this approach, $L(\beta, D)$ can be maximized using ordinary logistic regression, with Kn_+ responses and linear predictors $X_{ij}^T \beta + \sqrt{D} S_l, \ell = 1, \ldots, K$. This approach is easily implemented, but can be unmanageable if n_+ is large. A similar approach was used for the compound Poisson by Hinde (1982). The approach taken by most authors in this setting is to seek other approximations to $L(\beta, D)$ and its derivatives.

8.3 An Approximate Likelihood Approach: *PQL*

Penalized Quasi-Likelihood (PQL) (Green, 1987) is a method for approximate quasi-likelihood estimation with random effects. Similar approaches were proposed by Laird (1978), Stiratelli *et al.* (1984), Schall (1991) and McGilchrist and Aishett (1991). These approaches are reviewed in the paper by Breslow and Clayton (1993).

The PQL approach is more general than marginal ML, since $f(Y_i \mid \beta, b_i)$ in (8.3) is replaced by its quasi-likelihood equivalent

$$ql_i = \exp\left\{-\sum_{j=1}^{n_i} d_i(Y_{ij}, g(X_{ij}^T\beta + Z_{ij}b_i))/2\phi\right\}$$

where $d_i(\cdot)$ is an appropriate deviance function. We restrict attention to the binary case where $\phi = 1$ and the deviance is $\ln\left(p_{ij}^{Y_{ij}}(1-p_{ij})^{1-Y_{ij}}\right)$, up to an additive constant. The PQL approach is to use Laplace's method for integral approximations. After various approximations and much simplification, the log penalized quasi-likelihood can effectively be written as

$$\begin{aligned} ql(\beta, D) &= \frac{1}{2}\sum_{i=1}^{N} \ln \mid I + Z_i^T(\Delta_i^{-1}V_i\Delta_i^{-1})^{-1}Z_iD \mid \\ &\quad -\sum_{i=1}^{N}\sum_{j=1}^{n_i} \ln\left(p_{ij}^{Y_{ij}}(1-p_{ij})^{1-Y_{ij}}\right) - \frac{1}{2}\sum_{i=1}^{N} b_i^T\, D^{-1}b_i, \end{aligned}$$

where

$$V_i = (\Delta_i^{-1}R_i\Delta_i^{-1} + Z_iDZ_i^T).$$

Assuming that the dependence of V_i on β can be ignored and D is known, approximate likelihood equations for β and estimating equations for the b_i can be obtained by jointly maximizing the last two terms, i.e., maximizing

$$\sum_{i=1}^{N}\left\{\sum_{j=1}^{n_i} \ln\left(p_{ij}^{Y_{ij}}(1-p_{ij})^{1-Y_{ij}}\right) - \frac{1}{2}\Sigma b_i^T D^{-1} b_i\right\}, \tag{8.4}$$

as a function of β and $B^T = (b_1^T, \ldots, b_N^T)$, while D is held constant. For the binary case with logit link function, this leads to

$$\sum_{i=1}^{N} X_i^T\ (Y_i - \widehat{p}_i) = 0$$

and

$$\widehat{b}_i = DZ_i^T(Y_i - \widehat{p}_i)$$

where

$$\widehat{p}_i = g(X_i^T\widehat{\beta} + Z_i\widehat{b}_i).$$

Precisely this same set of estimating equations for (β, b) was derived by Stiratelli *et al.* (1984) (see also Knuiman and Laird, 1988), using an empirical Bayes approach. Drawing on the normal analogy, the unified empirical Bayes approach to estimating β, B and D is to assign β a flat prior, estimate (β, B) by their joint posterior means, and var $(\widehat{\beta}, \widehat{B})$ by their posterior variance matrix, holding D fixed. Then D is estimated by maximizing the marginal likelihood, obtained by integrating β out of (8.3) as well. Notice that with multivariate normality, the joint posterior moments are equal to marginal posterior moments for (β, D), the posterior means equal the posterior modes and the posterior variances will be the inverse second derivative matrices.

Because of the intractability of the posteriors, posterior modes rather than means are used. But, with a flat prior for β, the posterior mode for (β, B) given D is obtained by maximizing (8.4), i.e. the empirical Bayes and PQL estimates for β, B coincide. Notice that the likelihood equations for β look exactly like ordinary logistic regression, except that the linear predictor is $X_i^T\beta + Z_i\widehat{b}_i$, not $X_i^T\beta$. When $D \doteq 0$, $\widehat{b}_i \doteq 0$ and $\widehat{\beta}$ is approximately the ordinary logistic regression estimator, since $D \doteq 0 \Rightarrow b_i \doteq 0$. Alternately, when D is very large so that $D^{-1}\widehat{b}_i \doteq 0$, the estimate for $\widehat{b}_i$ requires that

$$Z_i^T(Y_i - \widehat{p}_i) \doteq 0$$

which would be the same as treating the b_i's as fixed constants and maximizing

$$\sum_{i=1}^{N}\sum_{j=1}^{n_i} \ln\left\{p_{ij}^{Y_{ij}}(1-p_{ij})^{1-Y_{ij})}\right\}$$

as a function of (β, B). Breslow and Clayton (1993) suggest estimating var$(\widehat{\beta})$ by $\widehat{V}$, where $\widehat{V} = (\Sigma X_i^T W_i X_i)^{-1}$.

Now consider estimation of D. In the normal case, the variance components may be found using the profile likelihood approach. Here estimating D is more complicated because of the dependence of W_i on $(\widehat{\beta}(\theta), \widehat{B}(\theta))$ where θ denotes the parameters in D. If we ignore that dependence, and replace $f(Y_{ij} \mid p_{ij})$ by a standard kernel for $(\tilde{Y}_{ij} -$

$X_{ij}^T\widehat{\beta})$, one can again obtain a similar set of scoring equations for D by maximizing

$$\frac{1}{2}\sum_{i=1}^{N} \ln \mid W_i \mid -\frac{1}{2}\sum_{i=1}^{N}(\tilde{Y}_i - X_i\widehat{\beta})^T W_i(\tilde{Y}_i - X_i\widehat{\beta})$$

for maximum likelihood, or

$$\frac{1}{2}\sum_{i=1}^{N} \ln \mid W_i \mid -\frac{1}{2}\ln \mid \sum_{i=1}^{N} X_i^T \widehat{V}^{-1} X_i \mid -\frac{1}{2}\sum_{i=1}^{N}(\tilde{Y}_i - X_i\widehat{\beta})^T W_i(\tilde{Y}_i - X_i\widehat{\beta})$$

for REML.

The alternative approximation of Stiratelli *et al.* (1984) is to maximize the marginal likelihood to estimate D, integrating out both β and B. They avoid direct approximation of this integrated likelihood by approximating the derivatives instead. If the b_i's were observed, $\Sigma b_i b_i^T$ would be the sufficient statistic for D. Hence, as in the normal theory case, EM estimating equations yield

$$\widehat{D} = \left[\sum_{i=1}^{N} E\left((b_i \mid Y_i, \widehat{D})(E(b_i^T \mid Y_i, \widehat{D})\right)\right.$$
$$\left. +E\left(\mathrm{var}(b_i \mid Y_i, \beta, \widehat{D})\right) + \mathrm{var}\ E(b_i \mid Y_i, \beta, \widehat{D})\right] \Big/ N.$$

These expectations can be readily evaluated by assuming that the joint posterior of (β, B) is approximately normal, with mean given by $(\widehat{\beta}(D)$, $\widehat{B}(D))$ and variance given by the inverse second derivative matrix. These approximations should work well provided q is small relative to each n_i and N is large, but may be poor otherwise.

The PQL approach for estimating θ and β gives biased results, and in some cases the bias can be substantial. Breslow and Lin (1995) and Lin and Breslow (1996) study the bias and give simple correction formulas for using the PQL approach when $q = 1$.