Chapter 7

Likelihood Models for Repeated Binary Data

In Section 6.1 we showed the close correspondence between exponential family likelihood theory and quasi-likelihood for generalized linear models in the univariate setting. In the multivariate case, it is not possible to completely generalize the GLM theory to likelihoods in a way which is entirely satisfactory for non-normal data. Here we will deal with some approaches to constructing likelihood models for repeated binary data. To fix ideas, we will first consider a single sample, no covariates, with $n_i = n$. We also drop the subscript i for much of the discussion for simplicity.

With measured multivariate responses, likelihood based analyses are invariably based on the Multivariate Normal (MVN) distribution, although how the mean and variance are parameterized may differ for different models. The MVN has many attractive features:

1. In the general case, the distribution is indexed by n parameters for the mean μ and $n(n+1)/2$ parameters for the variance-covariance matrix Σ.

2. Any subset of the n-vector Y, say Y_s, also has a multivariate normal with μ_s and Σ_s being the corresponding subsets of (μ, Σ). This property is known as reproducibility.

3. The MVN ensures consistency of ML estimates of μ and Σ, *even if* MVN does not hold. In particular, if $E(Y) = X\beta$ and $V(Y) = \Sigma(\alpha)$, then the ML estimates of (β, α) will be consistent even if Y is *not* MVN.

4. The parameters μ and Σ are distinct.

In contrast, many of these features do not hold when $Y^T = (Y_1, \ldots, Y_n)$ and each Y_j is binary. One attractive feature of binary data is that it is always true that the joint pdf of Y is multinomial with 2^n points in the sample space; that is, it is not an assumption like the MVN, but can be derived from first principles.

1. In the most general case, the joint pdf has $2^n - 1$ parameters, which grows much more rapidly than $n + n(n+1)/2$ as n gets large. This implies we need models which permit parsimonious parameter specification. The parsimonious parameter models are derived by making assumptions on the multivariate distribution.

2. Given a subset Y_s, the joint distribution of the Y_s is also multinomial, but the parameters of the pdf of Y_s are now sums of the parameters of the pdf of Y. For example, if $n = 3$ and

   $$\pi_{i_1 i_2 i_3} = P(Y_1 = i_1 \text{ and } Y_2 = i_2 \text{ and } Y_3 = i_3), \qquad i_1, i_2, i_3 = 0, 1,$$

   denote the parameters of the cell probabilities for (Y_1, Y_2, Y_3), then the marginal distribution of (Y_1, Y_2) is again multinomial, but with parameters

   $$\begin{aligned} \pi_{i_1 i_2} &= \sum_{i_3=0}^{1} \pi_{i_1 i_2 i_3} \\ &= P(Y_1 = i_1 \text{ and } Y_2 = i_2). \end{aligned}$$

3. The parameters for the mean and variance are functionally related:

   $$\text{var } (Y_j) = E(Y_j)(1 - E(Y_j)), \qquad j = 1, \ldots, n,$$

   and

   $$\text{cov } (Y_j, Y_k) = P(Y_j = Y_k = 1) - E(Y_j)E(Y_k), \qquad j, k = 1, \ldots, n.$$

4. As we have discussed in Chapter 6, we also typically use a nonlinear link function to relate covariates to $E(Y_j) = \mu_j$.

All of these features mean that likelihood models for binary responses are more difficult to formulate than was the case with linear models in the MVN setting.

When dealing with the GLM model for multivariate data with covariates, we seek to specify a distribution for each $Y_{i\,(n\times 1)}$, $i = 1, \ldots, N$, while retaining the basic regression model of interest, namely that

$$\mu_i = E(Y_i) = g(X_i\beta) \tag{7.1}$$

for

$$\ell\,(\mu_i) = X_i\beta. \tag{7.2}$$

By definition, the distribution of Y_i is multinomial; we can view the mean restriction as imposing a model on the n one-way margins of an n dimensional array with cell probabilities $\pi_{i\,(2^n\times 1)}$. The cell probabilities depend upon i through the covariates specifying the margins. It may also be desirable to let the association parameters depend upon covariates. Thus the main issue for likelihood modeling is how to model higher order associations in the table in a flexible and interpretable way that is consistent with the model for the mean μ_i. Several general approaches have been suggested; we focus here on two approaches, both of which are related to the general log-linear modeling system for contingency tables. We will first briefly review this approach to modeling associations with multivariate binary responses, and then discuss two modifications for longitudinal data.

7.1 A brief overview of log-linear models

These models were introduced by Birch (1963), and they have enjoyed considerable success for 1) studying associations among a set of n binary (or categorical) variables or for 2) logistic regression when one of the n variables is an outcome, and the remainder are categorical predictors (Bishop, Fineberg and Holland, 1975).

To explain basic ideas, we first consider $n = 3$, and assume no covariates. Again, we generally drop the subscript i. The basic principle is that the multinomial cell probabilities are not especially useful for studying associations or the effects of covariates, so we make a transformation from π to some other parameter set which is useful. The transformation needs to be one-to-one and invertible, and we need the new parameters to have meaningful interpretations.

The log-linear transform can be written in the form

$$\underset{(2^n-1)\times 1}{\lambda} = \underset{(2^n-1)\times 2^n}{C_1^T}\ \ln\ \underset{2^n\times 1}{\pi}$$

for a certain C_1. The elements of λ can be partitioned as:

Label	**Effects**	**Number**
main effects	$\lambda_1, \lambda_2, \lambda_3, \ldots, \lambda_n$	n
$2-$ way effects	$\lambda_{12}, \lambda_{13}, \ldots, \lambda_{n(n-1)}$	$\binom{n}{2}$
$3-$ way effects	$\lambda_{123}, \lambda_{124}, \ldots, \lambda_{n(n-1)(n-2)}$	$\binom{n}{3}$
$\vdots$	$\vdots$	$\vdots$

where, for $n = 3$,

$$\lambda_1 = \ln \frac{\pi_{1**}}{\pi_{0**}},$$

and similarly, for λ_2, λ_3,

$$\lambda_{12} = \ln \frac{\pi_{11*}\pi_{00*}}{\pi_{10*}\pi_{01*}},$$

and similarly for λ_{13}, λ_{23}, and

$$\lambda_{123} = \ln \left\{ \frac{\pi_{111}\pi_{001}}{\pi_{101}\pi_{011}} \div \frac{\pi_{110}\pi_{000}}{\pi_{100}\pi_{010}} \right\},$$

and where $*$ indicates taking the geometric mean over the omitted subscripts. Thus

$$\lambda_1 = \ln \frac{\pi_{1**}}{\pi_{0**}} = \ln \pi_{1**} - \ln \pi_{0**}$$

where

$$\pi_{1**} = (\pi_{111}\pi_{110}\pi_{101}\pi_{100})^{1/4}$$

so that

$$\ln\ \pi_{1**} = \frac{1}{4}\, \ell_{1++} = \bar{\ell}_{1..}$$

where

$$\ell_{ijk} = \ln\ \pi_{ijk}.$$

It follows that

$$\lambda_1 = \bar{\ell}_{1..} - \bar{\ell}_{2...}$$

Similar expressions can be derived for λ_{12} and λ_{123}. From this it is apparent that $\lambda = C_1^T \ln \pi$ for an appropriate C_1 since each element λ is a linear combination of the elements of $\ln \pi$. Note that λ is $(2^n - 1)$ and π is 2^n, but $\sum \pi = 1$, hence there are only $2^n - 1$ unique cell

probabilities. As we will discuss, the "higher order" λ's $(\lambda_{12}, \ldots, \lambda_{123})$ have interpretation as associations.

The log-linear transformation has many attractive features:

1. Apart from $\lambda_1, \ldots, \lambda_n$, which are often called the main effects, the higher order parameters can be thought of as log odds-ratios and differences of log odds-ratios, etc. To see this, consider the highest order term, or λ_{123} when $n = 3$. Rewriting λ_{123} as

$$\lambda_{123} = \ln \left\{ \frac{P(Y_1 = 1, Y_2 = 1 \mid Y_3 = 1)\, P(Y_1 = 0, Y_2 = 0 \mid Y_3 = 1)}{P(Y_1 = 1, Y_2 = 0 \mid Y_3 = 1)\, P(Y_1 = 0, Y_2 = 1 \mid Y_3 = 1)} \Big/ \right.$$
$$\left. \frac{P(Y_1 = 1, Y_2 = 1 \mid Y_3 = 0)\, P(Y_1 = 0, Y_2 = 0 \mid Y_3 = 0)}{P(Y_1 = 1, Y_2 = 0 \mid Y_3 = 0)\, P(Y_1 = 0, Y_2 = 1 \mid Y_3 = 1)} \right\},$$

   where we use the fact that

$$P(Y_1 = i_1, Y_2 = i_2, \mid Y_3 = i_3) = \pi_{i_1 i_2 i_3} / \pi_{++i_3}.$$

   Thus λ_{123} is the log odds-ratio measuring association between Y_1 and Y_2 given $Y_3 = 1$, minus the same log odds-ratio conditional on $Y_3 = 0$; i.e., $e^{\lambda_{123}}$ is a ratio of odds-ratios. Note that by symmetry, λ_{123} can also be thought of measuring odds-ratios for variables Y_1 and Y_3 given Y_2, or Y_2 and Y_3 given Y_1.

   If $\lambda_{123} = 0$, it implies that these 2-way conditional odds-ratios are independent of the value of the variable being conditioned on, i.e.,

$$\mathrm{OR}(Y_1, Y_2 \mid Y_3 = 1) = \mathrm{OR}(Y_1, Y_2 \mid Y_3 = 0),$$

   and similarly for $\mathrm{OR}(Y_1, Y_3 \mid Y_2)$, etc. In this case, the two-way parameters λ_{12}, etc., are directly interpretable as log odds-ratios, and it follows from setting $\lambda_{123} = 0$, and using the definition of λ_{12}, etc., that

$$\begin{aligned} \lambda_{12} &= \mathrm{OR}(Y_1, Y_2 \mid Y_3), \\ \lambda_{13} &= \mathrm{OR}(Y_1, Y_3 \mid Y_2), \\ \lambda_{23} &= \mathrm{OR}(Y_2, Y_3 \mid Y_1). \end{aligned}$$

2. To specify parsimonious models, we can take the higher associations (λ's) to be zero and get meaningful reduced models. The pairwise model is a popular choice:

$$\begin{aligned} \lambda &= (\lambda_1, \lambda_2, \lambda_3, \lambda_{12}, \lambda_{13}, \lambda_{23}), \\ \lambda_{123} &= 0, \end{aligned}$$

partly because this reduces us to n parameters for the mean $(\lambda_1, \lambda_2, \lambda_3)$ and $n(n-1)/2$ for the associations. With longitudinal data, we might also consider first order Markov models:

$$\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_n, \lambda_{12}, \lambda_{23}, \ldots, \lambda_{n(n-1)}),$$

where

$$\lambda_{13} = \lambda_{14} = \cdots \lambda_{123} = \lambda_{234} = 0.$$

3. It is easy to characterize and compute the MLE's of λ by using Iterative Proportional Fitting (IPF).

4. The λ's are variation independent, i.e., $\lambda \in \Re^k$, where $k = \dim \lambda$, so there are no restrictions on λ's, i.e., log-odds ratios do not depend on the margins $(\lambda_1, \ldots, \lambda_n)$.

5. Properties 4 and 5 are related to the fact that λ is the vector of canonical parameters in the exponential family representation of the multinomial. As shown by Cox (1972), for $n = 3$; we may write

$$\pi_{i_1 i_2 i_3} \propto \exp\left\{ \sum_j \lambda_j i_j + \sum_{j<k} \lambda_{ij} i_j i_k + \lambda_{123} i_1 i_2 i_3 \right\}.$$

The log-linear representation of π forms the basis for approaches (discussed in Section 7.3) suggested by Zhao and Prentice (1990) and Fitzmaurice and Laird (1993). The major difficulty with the log-linear model is that the "main effects," $\lambda_1, \ldots, \lambda_n$, are not very interesting or meaningful; this might be expected from a model designed to study only associations. If we are interested in $\mu_{ij} = E(Y_{ij}|X_{ij})$ as a function of covariates, this transformation is not attractive because the μ_{ij}'s are not a simple function of the λ's. We now discuss the Multivariate Logistic Transform which can be viewed as very similar to the log-linear, but the model uses marginal rather than conditional odds-ratios.

7.2 The Multivariate Logistic Transform (MLT)

The MLT is quite similar in nature to the log-linear transform. It is defined for the general case by

$$\underset{2^n \times 1}{\delta} = \underset{2^n \times 2^n}{C_2^T} \ln \underset{2^n \times 2^n}{L} \underset{2^n \times 1}{\pi}$$

for appropriate matrices C_2 and L. For $n = 3$, define

$$\delta_0 = \ln\left(\sum \pi\right) = 0,$$
$$\delta_1 = \text{logit}\ \mu_1 = \ln(\pi_{1++}/\pi_{0++}),$$

and similarly for δ_2 and δ_3;

$$\delta_{12} = \ln \frac{\pi_{11+}\pi_{00+}}{\pi_{10+}\pi_{01+}},$$

and similarly for δ_{13} and δ_{23}, and

$$\delta_{123} = \lambda_{123}.$$

Here δ_0 is a normalizing constant, ensuring $\sum \pi = 1$. So the MLT is exactly like the log-linear transform, except a "+" replaces a "*" in the omitted subscript. Thus we can partition δ as

Label	Parameter	Number
main effects	$\delta_1, \delta_2, \delta_3$	3
two-way associations	$\delta_{12}, \delta_{13}, \delta_{23}$	3
three-way association	δ_{123}	1.

Hence the main effects are precisely the parameters of interest, and the associations are marginal, rather than the conditional odds ratios.

The MLT was originally introduced by Grizzle, Starmer and Koch (1969), who proposed a weighted least squares analysis. McCullagh and Nelder (1989) coined the MLT phrase for $n = 3$, and proposed a maximum likelihood analysis. The general case is considered in Glonek and McCullagh (1995).

REMARKS.

1. Multivariate logistic models are attractive because the main effects are those of most interest in many longitudinal data settings. As with log-linear models, we can set higher order terms to zero and get meaningful models.

2. Unlike log-linear models, δ is not variation independent, and there may be margin incompatibility.

3. As with log-linear models, there is no closed form solution for estimating δ under ML, nor, except in special cases, can one invert the transform to express π in terms of δ. This must be done iteratively, and makes computations complex.

4. The parameter sets are not orthogonal. To be explicit, suppose now that each individual has covariate vectors X_i for Y_i and F_i for the association parameters, and we write

$$\delta_i = \begin{pmatrix} \delta_{i1} \\ \vdots \\ \delta_{in} \\ \delta_i^{HO} \end{pmatrix} = \begin{pmatrix} X_i\beta \\ F_i\alpha \end{pmatrix} \begin{matrix} p \times 1 \\ m \times 1 \end{matrix}$$

where δ_i^{HO} denotes the higher order effects (i.e., δ's with at least two subscripts) and α is a parameter indexing the associations. The information matrix, $\mathcal{I}\left(\begin{smallmatrix}\beta \\ \alpha\end{smallmatrix}\right)$, is *not* block diagonal, as it is in the MVN setting.

5. Because δ involves marginal moments of all order, it is relatively easy to handle missing responses (assuming MAR).

6. The MLT is reproducible.

7.3 A Mixed Parameter Transform

One way to retain the attractive features of both the log-linear and the MLT models is to use a mixed parameterization: Make a transformation from π to (δ^L, λ^{HO}) where $\delta^L = (\delta_1, \ldots, \delta_n)$ and λ^{HO} is λ without the main effects. Thus

$$\pi \rightarrow \begin{pmatrix} \delta^L \\ \lambda^{HO} \end{pmatrix}$$

and given covariate matrices X_i and F_i for the ith subject we can write

$$\pi_i \rightarrow \begin{pmatrix} \delta_i^L \\ \lambda_i^{HO} \end{pmatrix} \rightarrow \begin{pmatrix} X_i\beta \\ F_i\alpha \end{pmatrix}.$$

Since the λ^{HO} are the canonical parameters, several of the attractive features of log-linear models are retained.

A general expression for the likelihood equations under the mixed model for (β, α) are given by:

$$\sum_{i=1}^{N} \begin{pmatrix} \frac{\partial \mu_i}{\partial \beta} & 0 \\ \frac{\partial \tau_i}{\partial \beta} & \frac{\partial \tau_i}{\partial \alpha} \end{pmatrix}^{-1} \begin{pmatrix} \text{var}(Y_i) & \text{cov}(Y_i, T_i) \\ & \text{var}(T_i) \end{pmatrix}^{T} \begin{pmatrix} Y_i - \mu_i \\ T_i - \tau_i \end{pmatrix} = 0,$$

where T_i is the vector of sufficient statistics for λ^{HO}: $T_i = (Y_{i1}Y_{i2}, Y_{i1}Y_{i3}, \ldots, Y_{i1}Y_{i2}\ldots Y_{in})$ and $\tau_i = E(T_i)$. If we fit a parsimonious model, some of the higher order λ's will be set to zero, and the dimension will be reduced. For the mixed model parameterization, Fitzmaurice and Laird (1993) have shown that these likelihood equations can be rewritten as

$$\sum_{i=1}^{N} \begin{pmatrix} \frac{\partial \mu_i}{\partial \beta} & 0 \\ 0 & F_i \end{pmatrix}^{T} \begin{pmatrix} \Sigma_i^{-1} & 0 \\ -\text{cov}\,(Y_i, T_i)\,\Sigma_i^{-1} & I \end{pmatrix} \begin{pmatrix} Y_i - \mu_i \\ T_i - \tau_i \end{pmatrix} = 0,$$

where $\Sigma_i^{-1} = \text{var}(Y_i)$. Thus the likelihood equations for β are identical to GEE, where $W_i = \Sigma_i^{-1}$:

$$\sum_{i=1}^{N} \left(\frac{\partial \mu_i}{\partial \beta}\right)^{T} W_i\,(Y_i - \mu_i) = 0.$$

The difference is that $\text{var}\,(Y_i)$ is derived from the model for the distribution of Y_i.

In addition, Fitzmaurice and Laird (1993) show that the Fisher information matrix for (β, α) is block diagonal, thus the β and α parameters are orthogonal:

$$\begin{aligned} \text{Avar}\,(\widehat{\beta}, \widehat{\alpha}) &= \mathcal{I}^{-1}(\beta, \alpha) \\ &= \begin{bmatrix} \sum \left(\frac{\partial \mu_i}{\partial \beta}\right)^{T} \Sigma_i^{-1} \left(\frac{\partial \mu_i}{\partial \beta}\right) & 0 \\ 0 & \sum F_i^T(\Gamma_i - \xi_i\Sigma_i^{-1}\xi_i^T)F_i \end{bmatrix}^{-1} \end{aligned}$$

where $\Gamma_i = \text{var}\,(T_i)$ and $\xi_i = \text{cov}\,(Y_i, T_i)$. Thus avar $(\widehat{\beta})$ is the same as the GEE variance when $W_i = \Sigma_i^{-1}$.

By analogy with GEE, we will obtain a consistent estimate of β even if the model

$$\delta_i^{HO} = F_i \alpha$$

is wrong so that var (Y_i) is misspecified. In this case, however the Fisher information will not give the correct variance.

REMARKS.

- Because odds ratios are invariant to changes in the margins, the k-vector (δ^L, λ^{HO}) is variation independent in $\Re^k$.
- A computational algorithm for obtaining MLE's is given in Fitzmaurice and Laird (1993).
- A drawback to this parameterization is that because the association parameters are *conditional* rather than marginal odds ratios, this approach cannot be generalized to the arbitrary unbalanced case, and the interpretation of the association parameters is not always meaningful in the longitudinal setting, i.e., the log odds ratio expressing association between Y_1 and Y_2 conditional on the future value Y_3. A related feature is that the distribution is not reproducible, i.e., any subset of Y, say Y_s, does not have a mixed parameter model representation.

Fitzmaurice, Laird and Lipsitz (1995) discuss estimation when imbalance arises due to missing responses. Although the computations are generally straightforward, many of the nice features of the complete data case are lost. In particular, the likelihood equations become

$$\sum_{i=1}^{N} \begin{pmatrix} \frac{\partial \mu_i}{\partial \beta} & 0 \\ 0 & F_i \end{pmatrix}^T \begin{pmatrix} \Sigma_i^{-1} & 0 \\ -\xi_i \Sigma_i^{-1} & I \end{pmatrix} \begin{pmatrix} E(Y_i|Y_i^{OBS}) - \mu_i \\ E(T_i|Y_i^{OBS}) - \tau_i \end{pmatrix} = 0,$$

For β this yields

$$\sum_{i=1}^{N} \left(\frac{\partial \mu_i}{\partial \beta}\right)^T \Sigma_i^{-1} \left(E(Y_i|Y_i^{OBS}) - \mu_i\right) = 0$$

so that these are no longer the same as the GEE. In particular, $\widehat{\beta}$ will generally not be consistent for β with either MCAR or MAR data, unless the distribution of Y_i is correctly specified. This differs from ML with multivariate normal responses where estimates for β are always consistent

under MCAR even when the multivariate normal is misspecified. In addition, the asymptotic covariance of (α, β) is no longer zero, and the parameters are not orthogonal. Some simulations in simple cases suggest that the bias due to model misspecification in MCAR settings is small.