

Likelihood-Based Methods for Repeated Binary Data

Multivariate Normal Distribution

- In the most general case, an n -MVN distribution is completely specified by n mean parameters and $n(n-1)/2$ variance-covariance parameters.
- A subset of the n -vector \mathbf{Y} , say \mathbf{Y}_s , also has MVN distribution with mean $\boldsymbol{\mu}_s$ and variance Σ_s the corresponding subsets of $\boldsymbol{\mu}$ and Σ . (*Reproducibility*)
- The MVN theory ensures the consistency the MLEs of $\boldsymbol{\mu}$ and Σ even when MVN does not hold (only needs the correct specifications of the mean and variance).
- The parameters $\boldsymbol{\mu}$ and Σ are distinct (and can be estimated orthogonally).

Joint Multinomial Distribution

- An n -vector of binary variables \mathbf{Y} has an *exact* joint multinomial distribution with 2^n points in its sample space.
- In the most general case, the multinomial distribution has $2^n - 1$ number of parameters.
- A subset of the n -vector \mathbf{Y} , say \mathbf{Y}_s , also has a multinomial distribution. The parameters of $\Pr(\mathbf{Y}_s)$ are sums of the parameters of $\Pr(\mathbf{Y})$.
- The variances are functions of the means.
- To relate covariates to the means $\boldsymbol{\mu}$, a nonlinear link function is typically used (logit, probit).

Issues with Modeling Repeated Binary Data

- Parsimony: constrains higher-order associations to be zero.
- Flexibility: allows dependence on covariates.
- Interpretability: e.g., odds ratio is more natural than correlation.

The Log-Linear Model

- Log-linear models (Bishop et al, 1975) have been popular in studying multiple correlated categorical (binary) variables.
- The general form for the log-linear model:

$$\begin{aligned} & \log \Pr(\mathbf{Y} = \mathbf{y}) \\ &= c(\boldsymbol{\theta}) + \sum_{j=1}^n \theta_j y_j + \sum_{j_1 < j_2} \theta_{j_1, j_2} y_{j_1} y_{j_2} + \cdots + \theta_{1, \dots, n} y_1 \cdots y_n, \quad (1) \end{aligned}$$

where $c(\boldsymbol{\theta})$ is a normalizing constant.

- $\boldsymbol{\theta}$ is a $2^n - 1$ -vector of canonical parameters:

$$\boldsymbol{\theta} = (\theta_1, \dots, \theta_n, \theta_{12}, \dots, \theta_{n-1, n}, \dots, \theta_{1, \dots, n})^T.$$

- $\boldsymbol{\theta}$ can be viewed as a log-linear transformation of the multinomial cell probabilities $\boldsymbol{\pi}$ (an 2^n vector),

$$\boldsymbol{\theta} = \mathbf{C}_1^T \log \boldsymbol{\pi},$$

where \mathbf{C}_1 is an $2^n \times (2^n - 1)$ matrix.

- The elements of $\boldsymbol{\theta}$ can be partitioned as:

main effects	$\theta_1, \dots, \theta_n$	n
2-way effects	$\theta_{12}, \theta_{13}, \dots, \theta_{n-1, n}$	$\binom{n}{2}$
3-way effects	$\theta_{123}, \theta_{124}, \dots, \theta_{n-2, n-1, n}$	$\binom{n}{3}$
\vdots	\vdots	\vdots
n -way effect	$\theta_{1, \dots, n}$	1

Interpretation

- For $n = 3$,

$$\theta_1 = \log \frac{\pi_{100}}{\pi_{000}},$$

For the higher order parameters, we have

$$\theta_{12} = \log \frac{\pi_{110}\pi_{000}}{\pi_{100}\pi_{010}},$$

and

$$\theta_{123} = \log \left\{ \frac{\pi_{111}\pi_{001}}{\pi_{101}\pi_{011}} \div \frac{\pi_{110}\pi_{000}}{\pi_{100}\pi_{110}} \right\}.$$

- So it is apparent that each θ is a linear combination of $\log \pi$.
- The higher order parameters can be interpreted as log odds ratios and differences of log odds ratios and so on.
- Consider $n = 3$, θ_{123} can be rewritten as

$$\theta_{123} = \log \left\{ \frac{\Pr(Y_1 = 1, Y_2 = 1 | Y_3 = 1) \Pr(Y_1 = 0, Y_2 = 0 | Y_3 = 1)}{\Pr(Y_1 = 1, Y_2 = 0 | Y_3 = 1) \Pr(Y_1 = 0, Y_2 = 1 | Y_3 = 1)} \right\} \\ - \log \left\{ \frac{\Pr(Y_1 = 1, Y_2 = 1 | Y_3 = 0) \Pr(Y_1 = 0, Y_2 = 0 | Y_3 = 0)}{\Pr(Y_1 = 1, Y_2 = 0 | Y_3 = 0) \Pr(Y_1 = 0, Y_2 = 1 | Y_3 = 0)} \right\}.$$

So

$$\theta_{123} = \text{OR}(Y_1, Y_2 | Y_3 = 1) - \text{OR}(Y_1, Y_2 | Y_3 = 0).$$

When $\theta_{123} = 0$, θ_{12} etc., can be directly interpreted as log of the *conditional* odds ratios, that is,

$$\theta_{12} = \log \text{OR}(Y_1, Y_2 | Y_3).$$

Pros and Cons of Log-Linear Models

- By setting higher order parameters to 0, we get reduced parsimonious models that are interpretable.
- It is easy to characterize and compute the MLEs for $\boldsymbol{\theta}$.
- The range of $\boldsymbol{\theta}$ is not constrained, i.e., the log odds ratios do not depend on the marginal means (variation independent).
- The log-linear model is not convenient to model the marginal means as a function of the covariates because the marginal means are not simple functions of $\boldsymbol{\theta}$.
- The interpretation of the canonical parameters depends on the number of responses. Hence this formulation is not suitable for unbalanced data.

Bahadur Model

- The Bahadur model uses marginal means, correlations and higher-order moments to parameterize the multinomial distribution.
- Let $\mu_j = E(Y_j)$,

$$R_j = \frac{Y_j - \mu_j}{[\mu_j(1 - \mu_j)]^{1/2}},$$

$$\rho_{jk} = \text{Cor}(Y_j, Y_k) = E(R_j R_k),$$

$$\rho_{jkl} = E(R_j R_k R_l),$$

$$\rho_{1,\dots,n} = E(R_1 R_2 \cdots R_n).$$

$$\Pr(\mathbf{Y} = \mathbf{y}) = \prod_{j=1}^n \mu_j^{y_j} (1 - \mu_j)^{(1-y_j)} \times$$

$$\left(1 + \sum_{j < k} \rho_{jk} r_j r_k + \sum_{j < k < l} \rho_{jkl} r_j r_k r_l + \cdots + \rho_{1,\dots,n} r_1 r_2 \cdots r_n \right)$$

- Uses marginal means (parameters of interest) and correlations (familiar from continuous variables).
- The correlations are constrained by the marginal means (not variation independent) in a complicated manner.

Multivariate Logistic Model

- In the general form, the multivariate logistic transformation is defined by

$$\boldsymbol{\gamma} = \mathbf{C}_2^T \log \mathbf{L}\boldsymbol{\pi},$$

where $\boldsymbol{\gamma}$ and $\boldsymbol{\pi}$ are 2^n -vectors, \mathbf{C}_2 and \mathbf{L} are $2^n \times 2^n$ matrices.

- For $n = 3$,

$$\gamma_0 = \log \sum \pi = 0$$

$$\gamma_1 = \text{logit } \mu_1 = \log \frac{\pi_{1++}}{\pi_{0++}}$$

$$\gamma_{12} = \log \frac{\pi_{11+} \pi_{00+}}{\pi_{10+} \pi_{01+}}$$

$$\gamma_{123} = \theta_{123}$$

- Similar to log-linear transformation, with the sum $+$ replaces the geometric mean $*$.
- γ_0 is a normalizing constant to ensure $\sum \pi = 1$.
- γ_{12} , γ_{13} , etc, are the log of the *marginal* odds ratios.
- Higher order γ 's can be interpreted as contrasts of log odds ratios.
- As with log-linear model, we can set higher order effects 0 and get a meaning model *and* marginal mean parameter of interest.
- However, $\boldsymbol{\gamma}$ is not variation independent.
- No close form MLE for $\boldsymbol{\gamma}$ and $\boldsymbol{\pi}$.
- The mean and higher-order moment parameters are not orthogonal. If we use $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ to model the means and associations as functions of covariates, the information sub-matrix $\mathcal{I}(\boldsymbol{\beta}, \boldsymbol{\alpha})$ is not zero.
- The interpretations are independent of n .

Hybrid Model

- A compromise is to use the marginal means $\mu_j = E(Y_j)$, and the second- and higher order canonical parameters (Fitzmaurice and Laird, 1993).
- Make the transformation

$$\boldsymbol{\pi} \rightarrow \begin{pmatrix} \boldsymbol{\gamma}^L \\ \boldsymbol{\theta}^{\text{HO}} \end{pmatrix},$$

where $\boldsymbol{\gamma}^L = (\gamma_1, \dots, \gamma_n)$ and $\boldsymbol{\theta}^{\text{HO}} = (\theta_{12}, \theta_{13}, \dots, \theta_{1, \dots, n})$ is $\boldsymbol{\theta}$ without the main effects.

- Given covariate matrices \boldsymbol{X}_i and \boldsymbol{Z}_i for the i th subject, we can write

$$\begin{aligned} \text{logit}(\boldsymbol{\mu}) &= \boldsymbol{\gamma}^L = \boldsymbol{X}_i^T \boldsymbol{\beta} \\ \boldsymbol{\theta}^{\text{HO}} &= \boldsymbol{Z}_i \boldsymbol{\alpha} \end{aligned}$$

- If we set the third- and higher effects to zero, we get a *quadratic exponential family* distribution.
- $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are orthogonal.
- The score equation for $\boldsymbol{\beta}$ has the same form as GEE and we can get a consistent estimate of $\boldsymbol{\beta}$ even if the model for $\boldsymbol{\theta}^{\text{HO}}$ is wrong.
- $(\boldsymbol{\gamma}^L, \boldsymbol{\theta}^{\text{HO}})$ is variation independent.
- Not suitable for unbalanced data.
- Conditional odds ratios are not easily interpreted.

Further Reading

- Chapter 8.2 of DHLZ.

References

- Bishop YMM, Finberg SE, and Holland PW (1975) Discrete multivariate analysis: theory and practice. MIT Press, Cambridge, MA.
- Fitzmaurice GM and Laird NM (1993) A likelihood-based method for analysing longitudinal binary responses. *Biometrika* **80**:141151.
- Laird N (2004) Analysis of longitudinal and cluster-correlated data, vol. 8 of NSF-CBMS Regional Conference Series in Probability and Statistics. Institute of Mathematical Statistics.