6. Likelihood Models 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 Y, say Y_s also has MVN distribution with mean μ_s and variance Σ_s (the corresponding subset of μ and Σ) (Reproducibility).
  - The MVN theory ensures the consistency of the MLE of μ and Σ even when MVN does not hold (only needs the correct specifications of the mean and variance).
  - The parameters μ and Σ are distinct (and can be estimated orthogonally).

- Multinomial Distribution
  - An $n-$vector of binary variables 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 Y, say Y_s also has a multinomial distribution. The parameters $P(Y_s)$ are sums of the parameters of $P(Y)$.
- The variances are functions of the means.
- To relate covariates to the means μ, a nonlinear link function is typically used (logit, probit).

• <u>Issues with Modeling Repeated Binary Data</u>

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

Log-linear Models

Loglinear models (Bishop et al., 1975) have been popular in study multiple correlated categorical (binary) variables.

- The general form for the log-linear model:

$$\log P(Y=y) = C(\theta) + \sum_{j=1}^{n} \theta_j y_j + \sum_{j_1<j_2} \theta_{j_1 j_2} y_{j_1} y_{j_2} + \cdots + \theta_{1\cdots n} y_1 \cdots y_n$$

  where $y = (y_1, \cdots, y_n)$ and $C(\theta)$ is a normalizing constant.

- θ is a $2^n - 1$ vector of canonical parameters:

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

- θ can be viewed as a loglinear transformation of the multinomial cell probabilities π (an 2^n vector),

$$\theta = C_1^T \log \pi$$

  where C_1 is a $2^n \times (2^n - 1)$ matrix.

- The elements of θ can be partitioned as:

main effects	$\theta_1, \cdots, \theta_n$	n
2-way effects	$\theta_{12}, \theta_{13}, \cdots, \theta_{n-1,n}$	$\binom{n}{2}$
3-way effects	$\theta_{123}, \theta_{124}, \cdots, \theta_{n-2,n-1,n}$	$\binom{n}{3}$
$\vdots$	$\vdots$	$\vdots$
n-way effects	$\theta_{12\cdots n}$	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
  $$\theta_{123} = \log \left\{ \frac{\pi_{111}\pi_{001}}{\pi_{101}\pi_{011}} \Big/ \frac{\pi_{110}\pi_{000}}{\pi_{100}\pi_{010}} \right\}$$
  - 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

$$\begin{aligned}\theta_{123} &= \log\left\{\frac{P(Y_1=1,Y_2=1|Y_3=1)P(Y_1=0,Y_2=0|Y_3=1)}{P(Y_1=1,Y_2=0|Y_3=1)P(Y_1=0,Y_2=1|Y_3=1)}\right\} \\ &- \log\left\{\frac{P(Y_1=1,Y_2=1|Y_3=0)P(Y_1=0,Y_2=0|Y_3=0)}{P(Y_1=1,Y_2=0|Y_3=0)P(Y_1=0,Y_2=1|Y_3=0)}\right\} \\ &= \log OR(Y_1,Y_2|Y_3=1) - \log OR(Y_1,Y_2|Y_3=0).\end{aligned}$$

When $\theta_{123} = 0$, $\theta_{12}, \theta_{13}, \cdots$ can be directly interpreted as log of the conditional odds ratios. That is

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

- PROS and CONS of Loglinear Models
  - By setting higher order parameters to 0, we get reduced parsimonious model that are interpretable.
  - It is easy to characterize and compute the MLE of θ.
  - The range of θ is not constrained. i.e., the log odds ratios do not depend on the marginal means (variation independent).

- The major difficulty is that the "main effects" are not very interesting or meaningful.
- 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 function of θ.
- 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 $\mu_j = E(Y_j)$, $\rho_{jk} = cor(Y_j, Y_k) = E(R_j R_k)$, $\rho_{jkl} = E(R_j R_k R_l)$, $\cdots$, $\rho_{1,\cdots,n} = E(R_1 R_2 \cdots R_n)$

where $R_j = \frac{Y_j - \mu_j}{\left[\mu_j(1-\mu_j)\right]^{\frac{1}{2}}}$.

$$
\begin{aligned}
& P(Y=y) \\
= \quad & \prod_{j=1}^{n} \mu_j^{y_j}(1-\mu_j)^{1-y_j} \\
\times \quad & \left(1 + \sum_{j<k} \rho_{jk} r_j r_k + \sum_{j<k<l} \rho_{jkl} r_j r_k r_l + \cdots + \rho_{1,\cdots,n} r_1 \cdots r_n\right).
\end{aligned}
$$

- Use marginal means (parameters of interest) and correlation (familiar from continuous variables).

- The correlation 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

$$\Gamma = C_2^T \log L\pi,$$

where Γ and π are 2^n-vectors, C_2 and L are $2^n \times 2^n$ matrices.
For $n = 3$,

$$
\begin{aligned}
\gamma_0 &= \log \Sigma\pi = 0, \\
\gamma_1 &= \text{logit}\mu_1 = \log \frac{\pi_{1++}}{\pi_{0++}}, \\
\gamma_{12} &= \log \frac{\pi_{11+}\pi_{00+}}{\pi_{10+}\pi_{01+}}, \\
\gamma_{123} &= \theta_{123}.
\end{aligned}
$$

- Similar to log-linear transformation, with the sum $+$ replaces the geometric mean $*$.

- γ_0 is a normalizing constant to ensure $\Sigma\pi = 1$.

- $\gamma_{12}, \gamma_{13}, \cdots$ 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 effect 0 and get a meaningful model and marginal mean parameter of interest.

- However, γ is not variation independent. No closed forms of the MLE for γ and π as a function of γ are available.

- The mean and higher-order moment parameters are not orthogonal. If we use β and α to model the means and associations as functions of covariates, the information submatrix $I(\beta, \alpha)$ is not zero.

- The multivariate logistic model is reproducible (not dependent on 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

$$\pi \to \begin{pmatrix} \gamma^L \\ \theta^{Ho} \end{pmatrix}$$

where $\gamma^L = (\gamma_1, \cdots, \gamma_n)$ and $\theta^{Ho} = (\theta_{12}, \theta_{13}, \cdots, \theta_{1\cdots n})$ is θ without the main effects.

- Given covariate matrices X_i and Z_i for the ith subject, we can write

$$\begin{aligned} \text{logit}(\mu) \quad &= \quad \gamma^L = X_i^T \beta, \\ &\qquad \theta^{Ho} = Z_i \alpha. \end{aligned}$$

- If we set the third- and higher order effects to zero, we get a <u>quadratic exponential family</u> distribution.

- β and α are orthogonal.

- The score equation for β has the same form as GEE and we can get a consistent estimate of β even if the model for θ^{Ho} is wrong.

- (γ^L, θ^{Ho}) is variation independent.
- Not suitable for unbalanced data.
- Conditional odds ratios are not easily interpreted.

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). (1993). A likelihood-based method for analysing longitudinal binary responses. *Biometrika*, **80**: 141-151.

- 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.