Clustered Categorical Data

Jinhwan Suk

Department of Mathematical Science, KAIST

September 15, 2020

- Repeated measurements provides a *multivariate response* $(\boldsymbol{Y}_1, \boldsymbol{Y}_2, ..., \boldsymbol{Y}_T)$
- Consider marginal models for the $\{Y_t\}$ = mean of $\{Y_{it}\}$

Marginal Modeling : ML Approach

Longitudinal Study of Mental Depression

TABLE 11.2 Cross-Classification of Responses on Depression at Three Times by Diagnosis and Treatment

		Response at Three Times[a]							
Diagnosis	Treatment	NNN	NNA	NAN	NAA	ANN	ANA	AAN	AAA
Mild	Standard	16	13	9	3	14	4	15	6
	New drug	31	0	6	0	22	2	9	0
Severe	Standard	2	2	8	9	9	15	27	28
	New drug	7	2	5	2	31	5	32	6

[a] N, normal; A, abnormal.

Source: Reprinted with permission from the Biometric Society (Koch et al. 1977).

TABLE 11.3 Sample Marginal Proportions of Normal Response for Depression Data of Table 11.2

		Sample Proportion		
Diagnosis	Treatment	Week 1	Week 2	Week 4
Mild	Standard	0.51	0.59	0.68
	New drug	0.53	0.79	0.97
Severe	Standard	0.21	0.28	0.46
	New drug	0.18	0.50	0.83

- The marginal logistic model

$$\text{logit} P(Y_t = 1) = \alpha + \beta_1 s + \beta_2 d + \beta_3 t$$

  (time effect is the same for each group)
- $df = 12 - 4 = 8,\ G^2 = 34.6$
- A more realistic model permits the time effect to differ by drug,

$$\text{logit} P(Y_t = 1) = \alpha + \beta_1 s + \beta_2 d + \beta_3 t + \beta_4 (d \times t)$$

- $df = 12 - 5 = 7,\ G^2 = 4.2$
- When modeling multinomial response??

- At observation t, the marginal response distribution has $I - 1$ logits.

$$\text{logit}_j(t) = \alpha_j + \beta_j^T x_t$$

- For a **nominal** response, we can use a baseline-category logit,

$$\text{logit}_j(t) = \log \frac{P(Y_t = j)}{P(Y_t = I)}$$

- For **ordinal** responses, we can use the cumulative logit,

$$\text{logit}_j(t) = \text{logit}\left[P(Y_t \leq j)\right]$$

TABLE 11.4 Time to Falling Asleep, by Treatment and Occasion

	Time to Falling Asleep				
		Follow-up			
Treatment	Initial	< 20	20–30	30–60	> 60
Active	< 20	7	4	1	0
	20–30	11	5	2	2
	30–60	13	23	3	1
	> 60	9	17	13	8
Placebo	< 20	7	4	2	1
	20–30	14	5	1	0
	30–60	6	9	18	2
	> 60	4	11	14	22

Source: From S. F. Francom, C.Chuang-Stein, and J. R. Landis, *Statist. Med.* **8**: 571–582 (1989). Reprinted with permission from John Wiley & Sons Ltd.

TABLE 11.5 Sample Marginal Distributions of Table 11.4

		Response			
Treatment	Occasion	< 20	20–30	30–60	> 60
Active	Initial	0.101	0.168	0.336	0.395
	Follow-up	0.336	0.412	0.160	0.092
Placebo	Initial	0.117	0.167	0.292	0.425
	Follow-up	0.258	0.242	0.292	0.208

- The cumulative logit model,

$$\text{logit}\left[P(Y_t \leq j)\right] = \alpha_j + \beta_1 t + \beta_2 x + \beta_3 (t \times x)$$

- $df = 4 \cdot 3 - 3 - 1 - 1 - 1 = 6$
- The ML estimates are $\hat{\beta}_1 = 1.074$, $\hat{\beta}_2 = 0.046$, and $\hat{\beta}_3 = 0.662$.
- At the initial observation, estimated odds is $\exp(0.046) = 1.04$
- At the follow-up observation, the effect is $\exp(0.046 + 0.662) = 2.03$

- For T observations on an I-category response, at each setting of predictions the likelihood refers to I^T multinomial joint probabilities.
- $\boldsymbol{\pi}$: I^T-multinomial distribution parameter.
- Marginal logistic models have the form

$$\underbrace{\boldsymbol{C}\log(A\boldsymbol{\pi})}_{\text{logit}} = \boldsymbol{X}\beta$$

$$\begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix} \log \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \pi_{11} \\ \pi_{12} \\ \pi_{21} \\ \pi_{22} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \alpha,$$

- For *ML fitting*, # of parameters increases dramatically as T increases.
- An alternative to ML fitting uses a multivariate generalization of **quasi-likelihood**
- Recall (univariate) quasi-likelihood method

$$\sum_{i=1}^{N} \frac{(y_i - \mu_i)x_{ij}}{\nu(\mu_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right) = 0$$

- GEEs(1986) also requires the correlation structure among $\{Y_t\}$
  - *exchangeable* : $corr(Y_s, Y_t)$ identical for all s and t.
  - *autoregressive* : $corr(Y_s, Y_t) = \alpha^{|t-s|}$
  - *independence*, *unstructured*, ...
  - **Misspecified covariance doesn't affect consistency of GEE**

- For link function g, $\eta_i = g(\mu_i)$
- QL estimates $\hat{\boldsymbol{\beta}}$ are solutions of

$$\boldsymbol{u}(\boldsymbol{\beta}) = \sum_{i=1}^{N} \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \right)^T \nu(\mu_i)^{-1}(y_i - \mu_i) = \mathbf{0}$$

where $\mu_i = g^{-1}(\boldsymbol{x}_i^T \boldsymbol{\beta})$

- $\mathbf{E}\left[\boldsymbol{u}(\boldsymbol{\beta})\right] = 0$

Quasi-likelihood and GEE : Details

Properties of Quasi-likelihood Estimators

- QL estimators have properties similar to ML estimators.
- QL estimators are **asymptotically efficient** among estimators that are locally linear in $\{y_i\}$
- The QL estimators $\hat{\beta}$ are **asymptotically normal** with covariance matrix approximate by

$$\boldsymbol{V} = \left[\sum_{i=1}^{N} \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}}\right)^T v(\mu_i)^{-1} \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}}\right)\right]^{-1}$$

- $\hat{\beta}$ is **consistent** for β even if the variance function is misspecified.

- If we assume that $Var(Y_i) = \nu(\mu_i)$ but the true $Var(Y_i) \neq \nu(\mu_i)$, then the asymptotic covariance of $\hat{\beta}_{QL}$ is

$$\boldsymbol{V}\left[\sum_{i=1}^{n}\left(\frac{\partial\mu_i}{\partial\beta}\right)^T\left[\nu(\mu_i)\right]^{-1} Var(Y_i)\left[\nu(\mu_i)\right]^{-1}\left(\frac{\partial\mu_i}{\partial\beta}\right)\right]\boldsymbol{V} \tag{1}$$

- A consistent estimator of (1) : $\mu_i \to \hat{\mu}_i$ and $Var(Y_i) \to (y_i - \hat{\mu}_i)^2$

Quasi-likelihood and GEE : Details

- Let $\boldsymbol{y}_i = (y_{i1}, ..., y_{iT_i})^T$ and $\boldsymbol{\mu}_i = (\mu_{i1}, ..., \mu_{iT_i})^T$, $\mathbf{E}\boldsymbol{Y} = \boldsymbol{\mu}$
- GLM model : $\eta_{it} = g(\mu_{it}) = \boldsymbol{x}_{it}^T \boldsymbol{\beta}$
- Assume that y_{it} has probability mass function of form

$$f(y_{it};\ \theta_{it}, \phi) = \exp\{[y_{it}\theta_{it} - b(\theta_{it})]/\phi + c(y_{it}, \phi)\}$$

- From Section 4.4.2,

$$\mu_{it} = b'(\theta_{it}), \quad \nu(\mu_{it}) = b''(\theta_{it})\phi$$

- Assume a working correlation matrix $\boldsymbol{R}(\alpha)$ for $\boldsymbol{Y_i}$. Then covariance matrix is

$$\boldsymbol{V}_i = \boldsymbol{B}_i^{1/2} \boldsymbol{R}(\alpha) \boldsymbol{B}_i^{1/2} \phi$$

where $\boldsymbol{B}_i = \mathrm{diag}(\boldsymbol{b}''(\theta))$

Quasi-likelihood and GEE : Details

- Assume a working correlation matrix $\boldsymbol{R}(\alpha)$ for $\boldsymbol{Y_i}$.
- Then the working covariance matrix is

$$\boldsymbol{V}_i = \boldsymbol{B}_i^{1/2} \boldsymbol{R}(\alpha) \boldsymbol{B}_i^{1/2} \phi$$

  where $\boldsymbol{B}_i = \text{diag}(\boldsymbol{b}_i''(\theta))$
- $\boldsymbol{\Delta}_i = \text{diag}(\partial\theta_{it} / \partial\eta_{it})$
- $\boldsymbol{D}_i = \partial\boldsymbol{\mu}_i / \partial\boldsymbol{\beta} = \boldsymbol{B}_i \boldsymbol{\Delta}_i \boldsymbol{X}_i$
- Generalized estimating equations :

$$\sum_{i=1}^{n} \boldsymbol{D}_i^T \boldsymbol{V}_i^{-1} [\boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})] = \boldsymbol{0}$$