Ch.13 Clustered Categorical Data: Random Effects Models

Jaehyoung Hong

Generalized linear mixed model

- Let y_{it} : observation t in cluster i / $\boldsymbol{x}_{it}$: explanatory variables / $\boldsymbol{u}_i$: random effects for cluster i / $\mu_{it} = E(Y_{it}|\boldsymbol{u}_i)$

$$g(\mu_{it}) = \boldsymbol{x}_{it}^T\boldsymbol{\beta} + \boldsymbol{z}_{it}^T\boldsymbol{u}_i$$

g : Link function, $\boldsymbol{\beta}$: fixed effect model parameters, $\boldsymbol{u}_i \sim N(\boldsymbol{0}, \boldsymbol{\Sigma})$

✓ Random effect enters the model on the same scale as the predictor terms

✓ Sometimes, random effect represents heterogeneity caused by omitting certain explanatory variables

$$g(\mu_{it}) = \boldsymbol{x}_{it}^T\boldsymbol{\beta} + u_i^*\sigma$$

Logistic GLMM with random intercept for binary matched pairs

- Let cluster i consists of the responses (y_{i1}, y_{i2}), $y_{it} = 1\ (success)\ or\ 0\ (failure)$

$$(11.2.2)\ logit[P(Y_{it} = 1)] = \alpha_i + \beta x_t : x_1 = 0\ and\ x_2 = 1$$

$$logit[P(Y_{i1} = 1|u_i)] = \alpha + u_i,\ logit[P(Y_{i2} = 1|\mathrm{u_i})] = \alpha + \beta + u_i : u_i = \alpha_i - \alpha \sim \mathrm{N}(0, \sigma^2)$$

- ✓ Special case of GLMM : Random intercept model
- ✓ If σ is large, Y_1 and Y_2 has greater association ($Y_1 = \sum y_{i1}$)
- ✓ If $\sigma = 0$, Y_1 and Y_2 are independent

Logistic GLMM with random intercept for binary matched pairs

TABLE 12.1 Rating of Performance of Prime Minister

First Survey	Second Survey: Approve	Second Survey: Disapprove	Total
Approve	794	150	944
Disapprove	86	570	656
Total	880	720	1600

- $\hat{\beta} = \log\left(\frac{\hat{\mu}_{21}}{\hat{\mu}_{12}}\right) = \log\left(\frac{n_{21}}{n_{12}}\right) = \log\left(\frac{86}{150}\right) = -0.556$

- $\hat{\sigma} = 5.16$: strong association between the two response

Rasch model

- $T > 2$ observations in each cluster

$$logit[P(Y_{it} = 1|\mathrm{u_i})] = \alpha + \beta_t + u_i \ : u_i \sim \mathrm{N}(0, \sigma^2)$$

e.g) Response to a battery of T questions on an exam

Marginal model : $logit[P(Y_{it} = 1)] = \alpha + \beta_t$

- $T \times 2$ observation-by-outcome table
- $\beta_s - \beta_t = logit[P(Y_{hs} = 1)] - logit[P(Y_{it} = 1)]$

Rasch model : $logit[P(Y_{it} = 1|u_i)] = \alpha + \beta_t + u_i$

- $T \times 2 \times n$ observation-by-outcome-by subject table
- $\beta_s - \beta_t = logit[P(Y_{is} = 1|u_i)] - logit[P(Y_{it} = 1|u_i)]$

Logistic normal model

- Logistic normal model

$$logit[P(Y_{it} = 1|\mathrm{u_i})] = \boldsymbol{x}_{it}^T\boldsymbol{\beta} + u_i \ : u_i \sim \mathrm{N}(0, \sigma^2)$$

- If link function is arbitrary inverse cdf Φ, for $s \neq t$

$$\begin{aligned}\mathrm{cov}(Y_{is}, Y_{it}) &= E\big[\mathrm{cov}(Y_{is}, Y_{it}\,|\,u_i)\big] + \mathrm{cov}\big[E(Y_{is}\,|\,u_i), E(Y_{it}\,|\,u_i)\big] \\ &= 0 + \mathrm{cov}\big[\Phi(\mathbf{x}'_{is}\boldsymbol{\beta} + u_i), \Phi(\mathbf{x}'_{it}\boldsymbol{\beta} + u_i)\big].\end{aligned}$$

✓ Both monotone increasing in u_i, and hence are nonnegatively correlated

✓ Usually, the main focus in using a GLMM is inference about the fixed effect

✓ The random effect represents, how the positive correlation occurs between observations *within cluster*

Marginal effect is smaller than the subject-specific effect as $\boldsymbol{\sigma}$ becomes larger

- GLMM

$$E(Y_{it}|\boldsymbol{u}_i) = g^{-1}(\boldsymbol{x}_{it}^T\boldsymbol{\beta} + \boldsymbol{z}_{it}^T\boldsymbol{u}_i)$$

$$E(Y_{it}) = E[E(Y_{it}|\boldsymbol{u}_i)] = \int g^{-1}\left(\boldsymbol{x}_{it}^T\boldsymbol{\beta} + \boldsymbol{z}_{it}^T\boldsymbol{u}_i\right)f(\boldsymbol{u}_i;\boldsymbol{\Sigma})d\boldsymbol{u}_i$$

- If identity link

$$E(Y_{it}) = \int \left(\boldsymbol{x}_{it}^T\beta + \boldsymbol{z}_{it}^T\boldsymbol{u}_i\right)f(\boldsymbol{u}_i;\boldsymbol{\Sigma})d\boldsymbol{u}_i = \boldsymbol{x}_{it}^T\boldsymbol{\beta}$$

✓ Marginal model has the same model form and effects $\boldsymbol{\beta}$

Marginal effect is smaller than the subject-specific effect as σ becomes larger

- Logistic normal model

$$E(Y_{it}) = E[\frac{\exp(\boldsymbol{x}_{it}^T\boldsymbol{\beta} + u_i)}{1 + \exp(\boldsymbol{x}_{it}^T\boldsymbol{\beta} + u_i)}]$$

✓ Not have same form except when u_i has a degenerate distribution ($\sigma = 0$)

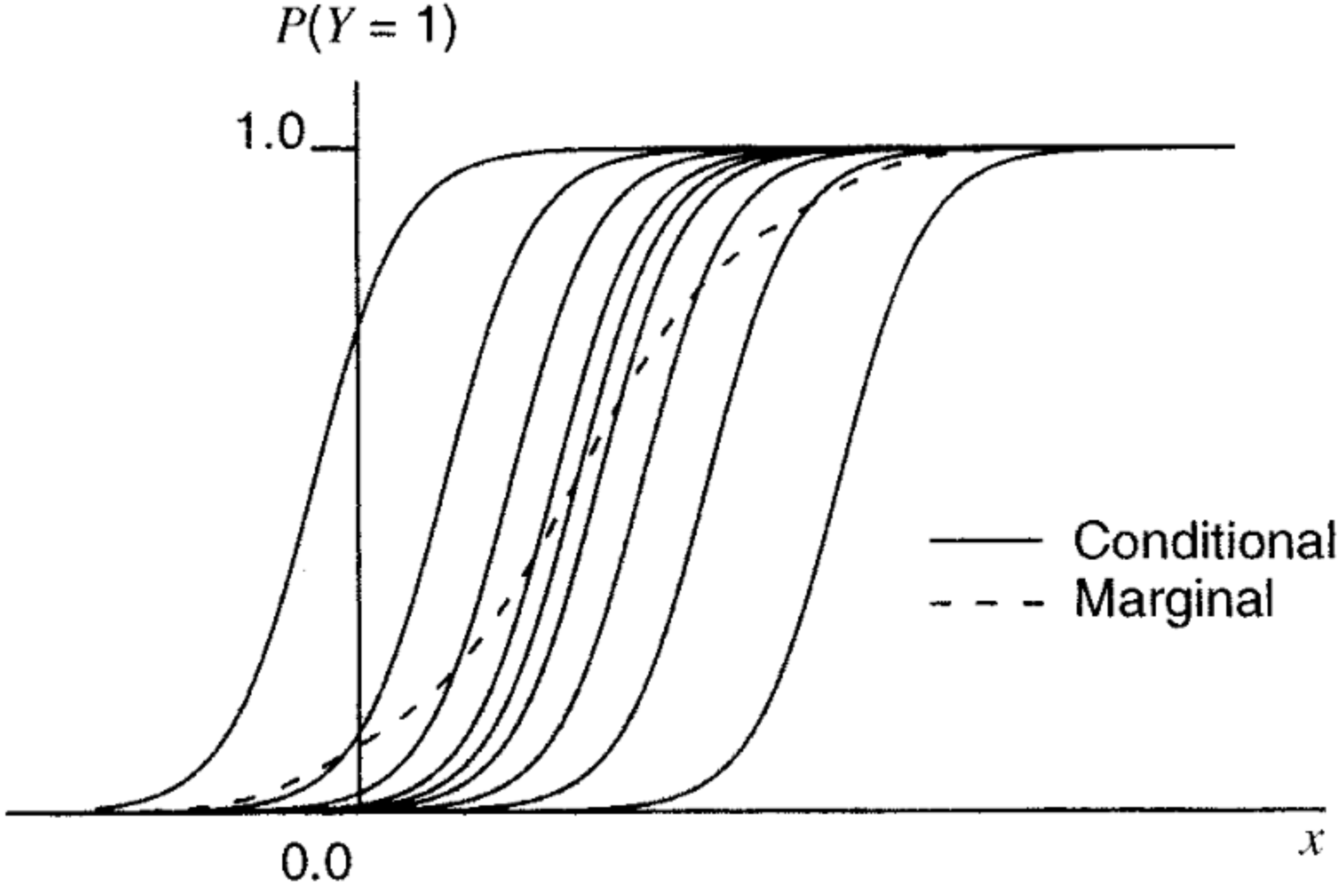
- Approximation

$$E(Y_{it}) = \frac{\exp(c\boldsymbol{x}_{it}^T\boldsymbol{\beta})}{1 + \exp(c\boldsymbol{x}_{it}^T\boldsymbol{\beta})}$$

✓ Where $c = [1 + 0.346\sigma^2]^{-0.5}$

✓ As σ increases, marginal effect decreases

Marginal effect is smaller than the subject-specific effect as σ becomes larger



- $P(Y_{it} = 1|u_i)$ has considerable heterogeneity (i.e. σ is large)

Modeling repeated binary responses

TABLE 10.13 Support for Legalizing Abortion in Three Situations, by Gender

	Sequence of Responses on the Three Items[a]							
Gender	(1, 1, 1)	(1, 1, 2)	(2, 1, 1)	(2, 1, 2)	(1, 2, 1)	(1, 2, 2)	(2, 2, 1)	(2, 2, 2)
Male	342	26	6	21	11	32	19	356
Female	440	25	14	18	14	47	22	457

- $logit[P(Y_{it} = 1|u_i)] = \alpha + \beta_t + \gamma x_i + u_i$: $x_i = 1$ for females and $x_i = 0$ for males; $u_i \sim N(0, \sigma^2)$

TABLE 12.3 Summary of ML Estimates for Random Effects Model (12.10) and ML and GEE Estimates for Corresponding Marginal Model

		GLMM ML		Marginal Model ML		Marginal Model GEE	
Effect	Parameter	Estimate	SE	Estimate	SE	Estimate	SE
Abortion	$\beta_1 - \beta_3$	0.83	0.16	0.148	0.030	0.149	0.030
	$\beta_1 - \beta_2$	0.54	0.16	0.098	0.027	0.097	0.028
	$\beta_2 - \beta_3$	0.29	0.16	0.049	0.027	0.052	0.027
Gender	γ	0.01	0.48	0.005	0.088	0.003	0.088
$\sqrt{\text{var}(u_i)}$	σ	8.6	0.54				

- ✓ $\hat{\gamma} = 0.013$: Low gender effect
- ✓ $\hat{\beta}$ shows situation1 gets more yes
- ✓ $\hat{\sigma}$ shows heterogenous subjects
  : strong association for three situations
- ✓ Marginal effect is small

Cumulative logit model with random intercept

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

- $logit[P(Y_{it} \leq j|\boldsymbol{u}_i)] = \alpha_j + \boldsymbol{x}_{it}^T\boldsymbol{\beta} + \boldsymbol{z}_{it}^T\boldsymbol{u}_i$
- $logit[P(Y_t \leq j] = \alpha_j + \beta_1 t + \beta_2 x + \beta_3(t\times x)$: x (treatment), t (initial / follow-up)

TABLE 12.7 Fits of Cumulative Logit Models to Table 11.4[a]

Effect	Marginal ML	Marginal GEE	Random Effects (GLMM) ML
Treatment	0.046 (0.236)	0.034 (0.238)	0.058 (0.366)
Occasion	1.074 (0.162)	1.038 (0.168)	1.602 (0.283)
Treatment × occasion	0.662 (0.244)	0.708 (0.244)	1.081 (0.380)

[a] Values in parentheses represent standard errors.

✓ Estimates are small in marginal which reflects relatively large heterogeneity ($\hat{\sigma} = 1.90$)

Multilevel modeling

 $\subseteq$  $\subseteq$ 

Student $\subseteq$ School $\subseteq$ Country

Hierarchical Structure

- ✓ GLMMs for data having hierarchical structure are called *multilevel modeling*

- ✓ Student / School / Country can be treated as random effects

Two-level model

- Let $y_{i(j)t}$ denote the response for student i in school j on test t (1=pass / 0=fail)

$$logit[P(Y_{i(j)t} = 1] = \boldsymbol{x}_{i(j)t}^{T}\boldsymbol{\beta} + u_j + \nu_{i(j)}$$

✓ Two random effects : $\{\nu_{i(j)}\}$ for students and $\{u_j\}$ for schools

✓ Two random effects are independent with $N(0, \sigma_u^2)$ and $N(0, \sigma_\nu^2)$

✓ $\{\nu_{i(j)}\}$: variability among students (large σ_ν : correlated result for test in each students)

✓ $\{u_j\}$: variability among schools

Two-level model

- Let $y_{i(j)t}$ denote the response for student i in school j on test t (1=pass / 0=fail)

$$logit[P(Y_{i(j)t} = 1] = \boldsymbol{x}_{i(j)t}^T\boldsymbol{\beta} + u_j + \nu_{i(j)}$$

- Latent variable model with $y^*_{i(j)t}$

$$y^*_{i(j)t} = \boldsymbol{x}_{i(j)t}^T\boldsymbol{\beta} + u_j + \nu_{i(j)} + \epsilon_{i(j)t}$$

✓ Latent model implies above model

✓ Random effects enters at two levels but actually three levels

✓ Total unexplained variability : $var(u_j) + var(\nu_{i(j)}) + var(\epsilon_{i(j)t})$

Two-level model

518

Table 13.13 ML Estimates and *SE* Values for … Adult Child Cares for Her Unmarried Elderly Mother

Effect	Estimate	*SE*
Intercept	−2.027	0.317
Ethnicity (vs. White)		
Black	0.162	0.157
Hispanic	−0.165	0.207
Other	0.459	0.498
Year (vs. 1998)		
2000	−0.152	0.084
2002	0.019	0.092
2004	0.072	0.106
Mother's characteristics		
Health (vs. Excellent)		
Very good	−0.105	0.173
Good	0.420	0.169
Fair	0.701	0.173
Poor	0.867	0.182
Age (vs. 75–79)		
70–74	−0.552	0.177
80–84	0.482	0.096
85–89	0.928	0.123
90+	1.213	0.156
Assets (dollars) (vs. 100,000–249,000)		
Negative	−0.336	0.258
0	0.004	0.151
<25,000	0.070	0.118
25,000–49,999	0.234	0.128
50,000–99,999	0.171	0.111
250,000+	−0.184	0.137
Final illness	1.411	0.088

Effect	Estimate	*SE*
Child characteristics		
Sex (Male = 1)	−1.435	0.118
Married (Yes = 1)	−0.179	0.119
Stepchild (Yes = 1)	−3.574	0.503
Children (Yes = 1)	−0.414	0.154
College (Yes = 1)	0.183	0.142
Parent raised child	0.154	0.250
Parent finan. help	−0.205	0.184
Family characteristics		
Family size (vs. 1)		
2	−1.052	0.181
3	−1.538	0.187
4	−1.967	0.201
5–6	−2.508	0.207
7+	−2.521	0.224
% Children		
Male	0.946	0.203
Married	−0.051	0.202
Stepchild	0.940	0.478
Have children	0.464	0.236
Attended college	−0.136	0.192
Family got help (vs. No)		
Yes	0.595	0.187
Missing	1.300	0.290

Source: Results taken from Table 2 in J. Henretta et al., *J. Marriage & Family*, **73**: 383–395, 2011. Reprinted with permission of J. Wiley & Sons.

- Child is 1-level, family is 2-level
- $\widehat{v_{i(j)}} = 4.38$ and $\widehat{u_j} = 1.2$
- Indicates 50% variability caused by within-child correlation