Chapter 3. Models for Multi-categorical Responses: Multivariate Extensions of GLM

MAST90084 Statistical Modelling Slides

Dennis Leung

SCHOOL OF MATHEMATICS AND STATISTICS

THE UNIVERSITY OF MELBOURNE

Outline

§3.5 Multivariate models for correlated responses

- So far: The multivariate $\mathbf{y}_i$ really is a surrogate for a univariate response taking multiple categorical values.
- Now: consider a *truly* multivariate non-Gaussian response vector whose components can be **correlated**
- Often happen in longitudinal studies, repeated measurements studies, and grouped (clustered) studies, etc.
- We will explore two approaches
  1. **conditional models**
  2. **marginal models**

§3.5.1 Conditional models: Asymmetric models (1)

Asymmetric models

- In many applications, the components of a response vector are ordered in a way that some components are considered "prior" to the other components, e.g. if they refer to events that take place earlier.
- In general, with m categorical responses $Y_1, \cdots, Y_m$ where Y_j depends on $Y_1, \cdots, Y_{j-1}$ but not on $Y_{j+1}, \cdots, Y_m$, the model has the decomposition

$$P(Y_1, \cdots, Y_m|\mathbf{x}) = P(Y_1|\mathbf{x}) \cdot P(Y_2|Y_1, \mathbf{x}) \cdots P(Y_m|Y_1, \cdots, Y_{m-1}, \mathbf{x}) \tag{1}$$

- Each component in (1) is specified by a GLM:

$$P(Y_j = r|Y_1, \cdots, Y_{j-1}, \mathbf{x}) = h_j(Z_j\beta) \tag{2}$$

where $Z_j = Z(Y_1, \cdots, Y_{j-1}, \mathbf{x})$ is a function of previous components $Y_1, \cdots, Y_{j-1}$ and the explanatory variables $\mathbf{x}$.

§3.5.1 Conditional models: Asymmetric models (2)

- **Markov-type transition models** have the additional assumption $P(Y_j = r | Y_1, \cdots, Y_{j-1}, \mathbf{x}) = P(Y_j = r | Y_{j-1}, \mathbf{x})$.
- A simple model for *binary responses* is

$$\log \frac{P(y_1 = 1|\mathbf{x})}{P(y_1 = 0|\mathbf{x})} = \beta_{01} + \mathbf{z}_1^T \boldsymbol{\beta}_1$$

$$\log \frac{P(y_j = 1|y_1, \cdots, y_{j-1}, \mathbf{x})}{P(y_j = 0|y_1, \cdots, y_{j-1}, \mathbf{x})} = \beta_{0j} + \mathbf{z}_j^T \boldsymbol{\beta}_j + y_{j-1}\gamma_j, \quad j = 2, \cdots, m.$$

§3.5.1 Conditional models: Asymmetric models (3)

- **Regressive logistic model** (Bonney, 1987), for binary responses, has the form

$$\log \frac{P(y_j = 1 | y_1, \cdots, y_{j-1}, \mathbf{x})}{P(y_j = 0 | y_1, \cdots, y_{j-1}, \mathbf{x})} = \beta_0 + \mathbf{z}_j^T \boldsymbol{\beta} + \gamma_1 y_1 + \cdots + \gamma_{j-1} y_{j-1}.$$

- (Markov assumption is not implied)

- If each y_j is multi-categorical, multinomial logit link can be used.

Asymmetric model is a MGLM

- Asymmetric model can be embedded in the multivariate GLM framework.
- Suppose each Y_j takes value in $\{1, \ldots, k_j\}$
- $\mathbf{Y} = (Y_1, \ldots Y_m)$, as a whole, is **identified** with a categorical variable taking possibly $k_1 \times .. \times k_m$ many different values.
- In other words, despite it being multivariate, we treat $\mathbf{Y}$ just like a univariate categorical variable that can take on $k_1 \times .. \times k_m$ possible values.
- So $\mathbf{Y}$ follows the *multinomial* distribution (with number of trials $= 1$) $\Rightarrow$ an exponential family!

Asymmetric model is a MGLM

- We can also "dummy code" $\mathbf{Y}$ to represent it as a random vector of length $k_1 \times .. \times k_m - 1$.
- When we observe n samples $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ of $\mathbf{Y}$, we can take average to give a *scaled* multinomial vector as before.
- The response function and the design matrix are given by (1) and (2). The implied link function generally has a very complicated form that isn't readily available in standard packages.
- However, if the multiplicative factors on the right hand side of

$$P(Y_1, \ldots, Y_m|x) = P(Y_1|x)P(Y_2|Y_1, x), \ldots, P(Y_m|Y_1 \ldots, Y_{m-1}|x)$$

only involve different parts of the β vector without overlapping, one may use standard functions to obtain the MLE factor by factor (as in the next example).

F&T, Example 3.12 (Clogg, 1982)

- **Reported happiness**: Study association between gender (x), years in school (Y_1), and reported happiness (Y_2) .
- Y_1 is modelled to be dependent on x.
- Since x and Y_1 are prior to the statement about happiness, Y_2 is modelled conditionally on Y_1 and x.

Table 1: Cross classification of gender, reported happiness, and years of schooling

		Years of school completed			
Gender	Reported happiness	< 12	12	13-16	≥ 17
Male	Not too happy	40	21	14	3
	Pretty happy	131	116	112	27
	Very happy	82	61	55	27
Female	Not too happy	62	26	12	3
	Pretty happy	155	156	95	15
	Very happy	87	127	76	15

Variables:

- $Y_1 =$ Years of School (4 ordinal levels: "< 12", "12","$13 - 16$","≥ 17",)
- $Y_2 =$ Happiness, (3 ordinal levels: "Not too happy", "pretty happy", "very happy")
- $X = Sex$, (2 levels, Male or Female)

Proposed asymmetric model in F & T:

- $P(Y_1 \leq r|x) = F(\theta_r + x'\beta_r^{(1)})$, $r = 1, 2, 3$
- $P(Y_2 \leq s|Y_1 = r, x) = F(\theta_{rs} + x'\beta_s^{(2)})$, $r = 1, 2, 3, 4$, $s = 1, 2$
- F simply taken to be the logistic function.
- Note: $\beta_r^{(1)}$'s are different for different r's, and so are $\beta_s^{(2)}$ for different s. $\Rightarrow$ We have **category-specific coefficient** for x. `polr()` from the `MASS` package doesn't handle this. But `vglm()` from the package *VGAM* can.
- F & T regressed the above model under the further restriction that $\beta_s^{(2)} = 0$ for all s.

	Estimate	Standard deviation	p-value
θ_1	-0.545	0.053	0.000
θ_2	0.841	0.056	0.000
θ_3	2.794	0.112	0.000
$\beta_1^{(1)}$	0.001	0.053	0.984
$\beta_2^{(1)}$	-0.201	0.056	0.000
$\beta_3^{(1)}$	-0.388	0.112	0.000
θ_{11}	-1.495	0.109	0.000
θ_{12}	0.831	0.092	0.000
θ_{21}	-2.281	0.153	0.000
θ_{22}	0.528	0.091	0.000
θ_{31}	-2.564	0.203	0.000
θ_{32}	0.575	0.109	0.000
θ_{41}	-2.639	0.422	0.000
θ_{42}	0.133	0.211	0.527

- F&T claims that this regression gives a deviance of 13.27 on $8 = 22 - 14$ degree of freedom. $14 = 6 + 8$ is the # parameters under the full GLM model; $22 = 2 \times 11$ is the # parameters for the saturated model because there are two (male and female) different samples of multinomial data with $12 = 3 \times 4$ categories.

- Strategy to compute the deviance:

  1. Compute the log-likelihood of the saturated multinomial model
  2. Compute the same for the asymmetric model.
  3. Take the difference, and multiply with 2.

- Step 2 involves two regressions using `vglm`: One for $P(Y_1 \leq r|x)$, another for $P(Y_2 \leq s|Y_1 = r, x)$. One can then add the log-likelihoods resulting from these two sub-regressions.

- See `Happiness.R`.

Symmetric models

- Response vector: $\mathbf{Y} = (y_1, \cdots, y_m)$.
- Assume, for simplicity, all $y_1, \cdots, y_m$ are **binary**. (Multicategorical cases can be handled similarly.)
- A symmetric model specifies:

$$P(y_j = 1 | y_k, k \neq j; \mathbf{x}_j), \quad j = 1, \cdots, m \tag{3}$$

- Defining feature: no natural ordering of the components of the response vector.

§3.5.1 Example 3.13: Visual impairment study

Table 2: Visual impairment data, from Liang, Zeger & Qaqish (1992)

Visual impairment	White				Black				Total
	Age								
	40-50	51-60	61-70	70+	40-50	51-60	61-70	70+	
Left eye									
Yes	15	24	42	139	29	38	50	85	422
No	617	557	789	673	750	574	473	344	4777
Right eye									
Yes	19	25	48	146	31	37	49	93	448
No	613	556	783	666	748	575	474	336	4751

- Binary **response** variables in the vector (y_1, y_2), where $y_1 = 1$ if left-eye impaired, 0 otherwise; $y_2 = 1$ if right-eye impaired, 0 otherwise. (y_1 and y_2 are correlated with no natural ordering)
- **Covariates**: `Age` (yrs., 4 levels), `Race` (W or B).
- **Aim**: find the effect of race and age on visual impairment.
- (Unfortunately, this dataset in the `Fahrmeir` R package is corrupted; we won't reproduce this example from the book)

§3.5.1 Conditional models: Symmetric models (2)

- Qu, Williams, Beck & Goormastic (1987) considers **logistic models** of the form:

$$\pi_j = P(y_j = 1 | y_k, k \neq j; \mathbf{x}_j) = h(\alpha(w_j; \boldsymbol{\theta}) + \mathbf{x}_j^T \boldsymbol{\beta}_j), \quad j = 1, \cdots, m \tag{4}$$

where $h(t) = \dfrac{e^t}{1+e^t}$ is the logistic cdf; and $\alpha(\cdot)$ is some function of a parameter θ and $w_j = \sum_{k \neq j} y_k$.

- When $m = 2$, a simple choice is

$$\pi_j = P(y_j = 1 | y_k, k \neq j; \mathbf{x}_j) = h(\theta_0 + \theta_1 y_k + \mathbf{x}_j^T \boldsymbol{\beta}_j), \quad j, k = 1, 2. \tag{5}$$

§3.5.1 Conditional models: Symmetric models (3)

- The joint density $P(y_1, \dots, y_m | x_1, \dots, x_m)$ derived from (4) or (5) involves a normalizing constant that is a complicated function in θ and β, making MLE-type full likelihood estimation computationally cumbersome. (Prentice 1988)
- Quasi-likelihood approach (Conolly and Liang, 1988): use an "*independent working*" quasi-likelihood and quasi-score function for each cluster $i \in \{1, \cdots, n\}$:

$$
\begin{aligned}
L_i(\boldsymbol{\beta}, \boldsymbol{\theta}) &= \prod_{j=1}^{m} \pi_{ij}^{y_{ij}} (1 - \pi_{ij})^{1 - y_{ij}} \\
\mathbf{S}_i(\boldsymbol{\beta}, \boldsymbol{\theta}) &= \sum_{j=1}^{m} \frac{\partial \pi_{ij}}{\partial (\boldsymbol{\beta}^T, \boldsymbol{\theta}^T)^T} \sigma_{ij}^{-2} (y_{ij} - \pi_{ij}(\boldsymbol{\beta}, \boldsymbol{\theta}))
\end{aligned}
$$

where $\mathbf{y}_i = (y_{i1}, \cdots, y_{ij}, \cdots, y_{im})^T$ are the responses for each i, $\pi_{ij}(\boldsymbol{\beta}, \boldsymbol{\theta}) = P(y_{ij} = 1 | \cdot)$ is defined by (4), and $\sigma_{ij}^2 = \pi_{ij}(\boldsymbol{\beta}, \boldsymbol{\theta})(1 - \pi_{ij}(\boldsymbol{\beta}, \boldsymbol{\theta}))$.

§3.5.1 Conditional models: Symmetric models (4)

- Denoting

$$M_i = \text{diag}\left\{\frac{\partial \pi_{i1}}{\partial (\boldsymbol{\beta})^T, \boldsymbol{\theta}^T)^T}, \cdots, \frac{\partial \pi_{im}}{\partial (\boldsymbol{\beta})^T, \boldsymbol{\theta}^T)^T}\right\}$$

$$\Sigma_i = \text{diag}\{\sigma_{i1}^2, \cdots, \sigma_{im}^2\}$$

$$\boldsymbol{\pi} = (\pi_{i1}, \cdots, \pi_{im})^T$$

we can rewrite $\mathbf{S}_i(\boldsymbol{\beta}, \boldsymbol{\theta})$ in matrix form

$$\mathbf{S}_i(\boldsymbol{\beta}, \boldsymbol{\theta}) = M_i \Sigma_i^{-1}(\mathbf{y}_i - \boldsymbol{\pi}_i),$$

a multivariate extension of the quasi-score.

- $\Rightarrow$ generalised estimating equation (GEE)

$$\mathbf{S}(\boldsymbol{\beta}, \boldsymbol{\theta}) = \sum_{i=1}^{n} \mathbf{S}_i(\boldsymbol{\beta}, \boldsymbol{\theta}) = \mathbf{0}$$

§3.5.1 Conditional models: Symmetric models (5)

- Roots $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})$ of the resulting generalised estimating equation (GEE)

$$\mathbf{S}(\boldsymbol{\beta}, \boldsymbol{\theta}) = \sum_{i=1}^{n} \mathbf{S}_i(\boldsymbol{\beta}, \boldsymbol{\theta}) = \mathbf{0}$$

  are consistent & asymptotically normal under regularity assumptions:

$$(\hat{\boldsymbol{\beta}}^T, \hat{\boldsymbol{\theta}}^T)^T \overset{a}{\sim} N((\boldsymbol{\beta}^T, \boldsymbol{\theta}^T)^T, \hat{F}^{-1}\hat{V}\hat{F}^{-1})$$

  with $\hat{F} = \sum_{i=1}^{n} \hat{M}_i \hat{\Sigma}_i^{-1} \hat{M}_i$ and $\hat{V} = \sum_{i=1}^{n} \hat{\mathbf{S}}_i \hat{\mathbf{S}}_i^T$.

- See Varin, Reid and Firth(2011)'s review article on "composite likelihood" for a modern treatment on this type of quasi-likelihood inference.

§3.5.2 Marginal models

- A potential drawback of conditional models: Measure the effect of $\mathbf{x}$ on a binary component y_j *conditional on the effects of other responses* y_k, $k \neq j \Rightarrow$ not able to provide prediction based on $\mathbf{x}$ alone.
- Marginal models: Analyse the **marginal mean** of the responses given the covariates. The association between the responses is of secondary interest.
- Proposed by Liang & Zeger (1986) and Zeger & Liang (1986) in the context of longitudinal data with many short time series.

Marginal models: Setup

- $\mathbf{y}_i = (y_{i1}, \cdots, y_{im_i})^T$ and $\mathbf{x}_i^T = (\mathbf{x}_{i1}^T, \cdots, \mathbf{x}_{im_i}^T)$ are respectively the vector of responses and the vectors of covariates for each sample $i \in \{1, \ldots, n\}$.
- Each i is often called a "cluster" to emphasize the components of $\mathbf{y}_i$ are correlated observations on the same type of variable.
- m_i is known as the "cluster size", and may vary with i.
- Within i, $y_{i1}, \cdots, y_{im_i}$ are correlated, but $\mathbf{y}_1, \ldots, \mathbf{y}_n$ are independent.
- The *marginal means* refer to the means of the components of $\mathbf{y}_i$, i.e. $\mu_{i1}, \ldots, \mu_{im_i}$.
- The effects of covariates on responses and the association between responses are modelled separately

Marginal models: mean specifications

- The **marginal means** of y_{ij}, $j = 1, \cdots, m_i$, are assumed **correctly specified** by common univariate response models:

$$\mu_{ij}(\beta) = E(y_{ij}|\mathbf{x}_{ij}) = h(\mathbf{z}_{ij}^T \beta) \tag{6}$$

where $h(\cdot)$ is a response function, e.g. a logistic function, and $\mathbf{z}_{ij}$ is an appropriate design vector.

Marginal models: variance/correlation specifications

- The **marginal variance** of each y_{ij} is specified as a function of μ_{ij}:

$$\sigma_{ij}^2 = \mathrm{var}(y_{ij}|\mathbf{x}_{ij}) = v(\mu_{ij})\phi \tag{7}$$

  where $v(\cdot)$ is a known **variance function**.

- The **correlation** between y_{ij} and y_{ik} is

$$\mathrm{corr}(y_{ij}, y_{ik}) = c(\mu_{ij}, \mu_{ik}, \boldsymbol{\alpha}) \tag{8}$$

  with a known $c(\cdot,\cdot,\cdot)$; so it is a function of $\mu_{ij} = \mu_{ij}(\boldsymbol{\beta})$, $\mu_{ik} = \mu_{ik}(\boldsymbol{\beta})$, and perhaps additional **association parameters** $\boldsymbol{\alpha}$:

- (7) and (8) $\Rightarrow$ **working covariance matrix**

$$\mathrm{cov}(\mathbf{y}_i) = \Sigma_i(\boldsymbol{\beta}, \boldsymbol{\alpha})$$

  (the dependence on ϕ is notationally suppressed).

Remarks:

- Apparently, this is the multivariate extension of the quasi-likelihood models in §2.3.1.

- The parameters $(\boldsymbol{\beta}, \boldsymbol{\alpha})$ are the same for all clusters $\Rightarrow$ marginal models analyze **population-averaged** effects.

- Marginal effects $\boldsymbol{\beta}$, which is the primary scientific objective, can be consistently estimated even if both $v(\mu_{ij})\phi$ and $c(\mu_{ij}, \mu_{ik}, \boldsymbol{\alpha})$ are just **working** (i.e. potentially misspecified) variance/correlation of y_{ij} and y_{ik}.

- However, when the correlation function is incorrectly specified, efficiency of $\hat{\boldsymbol{\beta}}$ can be compromise, as expected from our previous discussion in §2.3.1.

Specify a **working correlation matrix** $R_i(\boldsymbol{\alpha})$ to give the working covariance matrix

$$\Sigma_i(\boldsymbol{\beta}, \boldsymbol{\alpha}) = C_i^{1/2}(\boldsymbol{\beta}) R_i(\boldsymbol{\alpha}) C_i^{1/2}(\boldsymbol{\beta}),$$

where $C_i(\boldsymbol{\beta}) = \text{diag}\left[\text{var}(y_{ij}|x_{ij})\right] = \text{diag}\{\sigma_{i1}^2, \cdots, \sigma_{im_i}^2\}$. Common choices for $R_i(\boldsymbol{\alpha})$:

1. *working independence model*: $R_i(\boldsymbol{\alpha}) = I$, the identity matrix.
2. *equicorrelation* (or *exchangeable*) model: $\text{corr}(y_{ij}, y_{ik}) = \alpha$ for all $j \neq k$, i.e. $\boldsymbol{\alpha}$ reduces to be a scalar, and $R_i(\boldsymbol{\alpha}) = \begin{bmatrix} 1 & \alpha & \cdots & \alpha \\ \alpha & 1 & \cdots & \alpha \\ \vdots & \vdots & \ddots & \vdots \\ \alpha & \alpha & \cdots & 1 \end{bmatrix}$.
3. If enough data: $R_i(\boldsymbol{\alpha})$ *completely unspecified*, except being positive definite, i.e. $\alpha_{jk} = \text{corr}(y_{ij}, y_{ik})$, $j < k$.

Specifying association structure: Second approach

For binary responses, specifying the **odds ratios**:

- The odds ratio for y_{ij}, y_{ik}, $1 \leq j \neq k \leq m_i$, is defined by

$$\gamma_{ijk} = \frac{P(y_{ij} = 1, y_{ik} = 1)/P(y_{ij} = 0, y_{ik} = 1)}{P(y_{ij} = 1, y_{ik} = 0)/P(y_{ij} = 0, y_{ik} = 0)}$$

- Let $\pi_{ij} := P(y_{ij} = 1)$. It can be shown that
$P(y_{ij} = y_{ik} = 1) = E(y_{ij} y_{ik})$

$$= \begin{cases} \frac{1-(\pi_{ij}+\pi_{ik})(1-\gamma_{ijk})-s(\pi_{ij},\pi_{ik},\gamma_{ijk})}{2(\gamma_{ijk}-1)} & \text{if } \gamma_{ijk} \neq 1 \\ \pi_{ij}\pi_{ik} & \text{if } \gamma_{ijk} = 1 \end{cases} \quad (9)$$

with
$s(\pi_{ij}, \pi_{ik}, \gamma_{ijk}) = \left([1-(\pi_{ij}+\pi_{ik})(1-\gamma_{ijk})]^2 - 4(\gamma_{ijk}-1)\gamma_{ijk}\pi_{ij}\pi_{ik} \right)^{1/2}.$
(Lipstiz, Laird and Harrington, 1991)

- Hence, $\text{Cov}(\mathbf{y}_i)$ is expressible as a function in $\{\pi_{ij}, \pi_{ik}, \gamma_{ijk}\}_{1 \leq j,k \leq m_i}$, since $Cov(y_{ik}, y_{ij}) = E(y_{ij} y_{ik}) - \pi_{ij}\pi_{ik}$.

Specifying association structure: Second approach

- In light of the inequality

$$
\begin{aligned}
P(y_{ij} = y_{ik} = 1) &= P(y_{ij} = 1) + P(y_{ik} = 1) - P(y_{ij} = 1 \text{ or } y_{ik} = 1) \\
&\geq \pi_{ij} + \pi_{ik} - 1,
\end{aligned}
$$

  the intersection probability $P(y_{ij} = y_{ik} = 1)$ is constrained by

$$
\max(0, \pi_{ij} + \pi_{ik} - 1) \leq P(y_{ij} = y_{ik} = 1) \leq \min(\pi_{ij}, \pi_{ik}),
$$

  known as Fréchet inequaliity.

- Since

$$
\text{corr}(y_{ij}, y_{ik}) = \frac{P(y_{ij} = y_{ik} = 1) - \pi_{ij}\pi_{ik}}{\sqrt{\pi_{ij}(1 - \pi_{ij})\pi_{ik}(1 - \pi_{ik})}},
$$

  the correlations are constrained by the marginal means. This may narrow the range of admissible correlations if one models association structures with them.

Specifying association structure: Second approach

- In comparison, modeling with odds ratios has the advantage of not being constrained by the means.
- When modeling with odds ratios, one may further parametrize $\gamma_{ijk} = \gamma_{ijk}(\boldsymbol{\alpha})$ by $\boldsymbol{\alpha}$ to reduce the number of parameters (for parsimony).
- Common choices of $\gamma_{ijk}(\boldsymbol{\alpha})$:
  1. $\gamma_{ijk} = \gamma$, for all i, j, k.
  2. $\log \gamma_{ijk} = \boldsymbol{\alpha}^T w_{ijk}$, for some covariate w_{ijk}.
- Via (9), the working $\text{cov}(\mathbf{y}_i) = \Sigma_i(\boldsymbol{\beta}, \boldsymbol{\alpha})$ is a function in $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$.

Some examples of marginal models

We've focused on binary responses so far; apparently marginal models also apply to other data type:

I Continuous responses:

$$\mu_{ij}(\boldsymbol{\beta}) = E(y_{ij}|\mathbf{x}_{ij}) = \mathbf{z}_{ij}^T\boldsymbol{\beta}; \quad \mathrm{var}(y_{ij}|\mathbf{x}_{ij}) = \phi = \sigma^2; \quad \mathrm{corr}(y_{ij}, y_{ik}) = \alpha_{jk}.$$

II Binary responses:

$$\mu_{ij}(\boldsymbol{\beta}) = \pi_{ij}(\boldsymbol{\beta}) = P(y_{ij} = 1|\mathbf{x}_{ij}), \quad \log\frac{\pi_{ij}(\boldsymbol{\beta})}{1-\pi_{ij}(\boldsymbol{\beta})} = \mathbf{z}_{ij}^T\boldsymbol{\beta};$$

$$\mathrm{var}(y_{ij}|\mathbf{x}_{ij}) = \pi_{ij}(\boldsymbol{\beta})(1-\pi_{ij}(\boldsymbol{\beta}));$$

$$\mathrm{corr}(y_{ij}, y_{ik}) = 0 \text{ (independence struc.)} \quad \text{or} \quad \gamma_{ijk} = \alpha \text{ (equal odds ratio).}$$

III Count data:

$$\begin{aligned}
\log\mu_{ij}(\boldsymbol{\beta}) &= \log E(y_{ij}|\mathbf{x}_{ij}) = \mathbf{z}_{ij}^T\boldsymbol{\beta}; \\
\mathrm{var}(y_{ij}|\mathbf{x}_{ij}) &= \mu_{ij}(\boldsymbol{\beta})\phi; \\
\mathrm{corr}(y_{ij}, y_{ik}) &= \alpha \quad \text{(equicorrelation).}
\end{aligned}$$

Multivariate GEE

- The **generalised estimating equation** (GEE) for effect $\boldsymbol{\beta}$ is

$$\mathbf{S}_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{n} Z_i^T D_i(\boldsymbol{\beta}) \Sigma_i^{-1}(\boldsymbol{\beta}, \boldsymbol{\alpha})(\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})) = 0, \qquad (10)$$

  with $Z_i^T = (\mathbf{z}_{i1}, \cdots, \mathbf{z}_{im_i})$ and diagonal matrices $D_i(\boldsymbol{\beta}) = \text{diag}\{D_{ij}(\boldsymbol{\beta})\}$, $D_{ij}(\boldsymbol{\beta}) = \frac{\partial h}{\partial \eta_{ij}}$ evaluated at $\eta_{ij} = \mathbf{z}_{ij}^T \boldsymbol{\beta}$.
- This is a multivariate extension of the GEE in §2.3.1 with a correctly specified mean structure and a possibly misspecified covariance structure.
- $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$ (and possibly ϕ) are unknown and have to be estimated.
- General estimation strategy: Iterate between estimation of $\boldsymbol{\beta}$ given $(\phi, \boldsymbol{\alpha})$ and estimation of $(\phi, \boldsymbol{\alpha})$ given $\boldsymbol{\beta}$, until convergence.

Parameter estimation: Estimating β given (α, ϕ)

- Given current estimates $\hat{\alpha}$ (and $\hat{\phi}$), the GEE (10) for $\hat{\beta}$ is solved by the iterations

$$\hat{\beta}^{(k+1)} = \hat{\beta}^{(k)} + (\hat{F}^{(k)})^{-1}\hat{\mathbf{S}}_{\beta}(\hat{\beta}^{(k)}, \hat{\alpha}), \quad k = 0, 1, 2, \cdots,$$

  with

$$\hat{F}^{(k)} = \sum_{i=1}^{n} Z_i^T D_i(\hat{\beta}^{(k)})\Sigma_i^{-1}(\hat{\beta}^{(k)}, \hat{\alpha})D_i(\hat{\beta}^{(k)})Z_i$$

  being the observed quasi-information matrix.

- This is, again, a modified Fisher scoring algorithm. Like the univariate case in §2.3.1, ϕ isn't really involved due to cancellations.

Parameter estimation: Estimating (α, ϕ) given β

- Given the current estimate $\hat{\beta}$, Liang and Zeger (1986) suggest **method of moments** estimators for (α, ϕ) based on the Pearson residuals
$$\hat{r}_{ij} = \frac{y_{ij} - \hat{\mu}_{ij}}{\sqrt{v(\hat{\mu}_{ij})}}.$$
- The dispersion ϕ is estimated by $\hat{\phi} = \frac{1}{N-p} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \hat{r}_{ij}^2$, with
$N = \sum_{i=1}^{n} m_i$ and $p = \dim(\beta)$.

Parameter estimation: Estimating (α, ϕ) given β

- Estimation of $\boldsymbol{\alpha}$ depends on the choice of $R_i(\boldsymbol{\alpha})$. For exchangeable (equicorrelation) correlation matrix $R_i(\alpha)$ with $\dim(\alpha) = 1$,

$$\hat{\alpha} = \left[\hat{\phi}\left\{\sum_{i=1}^{n} \frac{1}{2} m_i(m_i - 1) - p\right\}\right]^{-1} \sum_{i=1}^{n} \sum_{k>j} \hat{r}_{ik}\hat{r}_{ij}.$$

- An unspecified working correlation matrix R can be estimated by

$$\hat{R} = \frac{1}{n} \sum_{i=1}^{n} \hat{C}_i^{-\frac{1}{2}} (\mathbf{y}_i - \hat{\boldsymbol{\mu}}_i)(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_i)^T \hat{C}_i^{-\frac{1}{2}}$$

if all cluster sizes $m_i = m$ and $m << n$, where

$$C_i(\boldsymbol{\beta}) = \mathrm{diag}\left[\mathrm{var}(y_{ij}|x_{ij})\right] = \mathrm{diag}\{\sigma_{i1}^2, \cdots, \sigma_{im_i}^2\}.$$

(there seems to be a typo in the formula of $\hat{R}$ in F & T)

- Cycling between Fisher scoring steps for $\boldsymbol{\beta}$ and estimation of $(\boldsymbol{\alpha}, \phi)$ leads to a consistent estimation of $\boldsymbol{\beta}$.
- (Remember our mean structure is *assumed* correctly specified. So using previous reasoning as in the quasi-likelihood inference for univariate *y*, we still have consistency for β. On that note, since the specified covariance structure isn't necessarily true, so $(\hat{\alpha}, \hat{\phi})$ may converge to **some** (α, ϕ), but not a true one, since there isn't a truth here anyway!)
- Alternatively, $\boldsymbol{\alpha}$ (and possible ϕ) can be estimated by simultaneously solving an additional estimating equation (Prentice, 1988). Details are not pursued here.

Statistical inference

- We also have the approximation

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \sim_a F^{-1}(\boldsymbol{\beta})\mathbf{S}_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \alpha)$$

  with $F = \sum_{i=1}^{n} Z_i^T D_i \Sigma_i^{-1} D_i Z_i$.
- The quasi-score $\mathbf{S}_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \alpha)$ is approximately $N(0, V)$ by central limit theorem, where

$$V = \text{cov}(\mathbf{S}_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\alpha})) = \sum_{i=1}^{n} Z_i^T D_i \Sigma_i^{-1} S_i \Sigma_i^{-1} D_i Z_i,$$

  and $S_i = \text{Cov}(\mathbf{y}_i)$.

Statistical inference: Making sandwich again

- Under regularity conditions, the GEE estimator $\hat{\boldsymbol{\beta}}$ satisfies

$$\hat{\boldsymbol{\beta}} \overset{a}{\sim} N(\boldsymbol{\beta}, F^{-1}VF^{-1}).$$

- $\text{cov}(\hat{\boldsymbol{\beta}})$ is approximated by the "sandwich matrix":

$$\hat{A} = \hat{F}^{-1} \underbrace{\left\{ \sum_{i=1}^{n} Z_i^T \hat{D}_i \hat{\Sigma}_i^{-1} (\mathbf{y}_i - \hat{\boldsymbol{\mu}}_i)(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_i)^T \hat{\Sigma}_i^{-1} \hat{D}_i Z_i \right\}}_{\hat{V}} \hat{F}^{-1}$$

- However, unlike the univariate case in Ch.2, this robust sandwich estimator may still require a dispersion estimate $\hat{\phi}$ to be constructed; if the working correlation model is the exchangeable model, the association parameter estimate $\hat{\alpha}$ (which cannot be cancelled) may need $\hat{\phi}$ to be constructed.

Marginal models for correlated responses having k categories

- Suppose categorical responses Y_{ij}, $j = 1, \cdots, m_i$, are observed in cluster i, $i = 1, \cdots, n$.

- For simplicity, each Y_{ij} has the k categories and is dummy coded by

$$\mathbf{y}_{ij} = (y_{ij1}, \cdots, y_{ijq})^T, \quad q = k - 1$$

- Let $\mathbf{y}_i^T = (\mathbf{y}_{i1}^T, \cdots, \mathbf{y}_{im_i}^T)$ be observations of $\mathbf{y}_{ij}$ in cluster i; $\mathbf{x}_i^T = (\mathbf{x}_{i1}^T, \cdots, \mathbf{x}_{im_i}^T)$ be the corresponding covariate observations.

- For data involving categorical responses, a marginal categorical response model can be defined for each response variable, and then supplemented by a working association model to relate the responses with each other within a cluster.

Marginal models for correlated responses having k categories

(i) The vector of marginal means or categorical probabilities of Y_{ij} is assumed being correctly specified by an *response model*:

$$\boldsymbol{\pi}_{ij}(\boldsymbol{\beta}) = (\pi_{ij1}(\boldsymbol{\beta}), \cdots, \pi_{ijq}(\boldsymbol{\beta}))^T = \mathbf{h}(Z_{ij}\boldsymbol{\beta})$$

with $\pi_{ijr} = P(Y_{ij} = r|\mathbf{x}_{ij}) = P(y_{ijr} = 1|\mathbf{x}_{ij})$, and the response function $\mathbf{h}(\cdot)$ and design matrix Z_{ij}. h can follow a nominal or ordinal response model, depending on whether the response categories can be ordered.

(ii) The marginal covariance function of $\mathbf{y}_{ij}$ is given by

$$\Sigma_{ij} = \text{cov}(\mathbf{y}_{ij}|\mathbf{x}_{ij}) = \text{diag}(\boldsymbol{\pi}_{ij}) - \boldsymbol{\pi}_{ij}\boldsymbol{\pi}_{ij}^T$$

i.e. the covariance matrix of a multinomial random variable.

Marginal models for correlated responses having k categories

(iii) Association between Y_{ij} and Y_{ik} can be modeled by a **working correlation matrix** R_i.

e.g.: the working matrix of exchangeable correlations is

$$R_i(\boldsymbol{\alpha}) = \begin{bmatrix} I & Q & \cdots & Q \\ Q^T & I & \cdots & Q \\ \vdots & \vdots & \ddots & \vdots \\ Q^T & Q^T & \cdots & I \end{bmatrix},$$

where the $q \times q$ matrix Q contains $\boldsymbol{\alpha}$ to be estimated by a method of moments.

If a working correlation matrix R_i is specified, then the working covariance structure is

$$\Sigma_i(\boldsymbol{\beta}, \boldsymbol{\alpha}) = C_i^{1/2}(\boldsymbol{\beta}) R_i(\boldsymbol{\alpha}) C_i^{1/2}(\boldsymbol{\beta}),$$

where $C_i(\boldsymbol{\beta}) = \mathrm{diag}(\Sigma_{i1}, \ldots, \Sigma_{im_i})$ is block-diagonal.

Marginal models for correlated responses having k categories

(iii) Alternatively, association can be modeled by odds ratios.

- For ordinal categories, the **global cross-ratios** (GCR) can be used.

  For a pair of categories ℓ and m of Y_{ij} and Y_{ik}, GCR is defined as

$$\gamma_{ijk}(\ell, m) = \frac{P(Y_{ij} \leq \ell, Y_{ik} \leq m)P(Y_{ij} > \ell, Y_{ik} > m)}{P(Y_{ij} > \ell, Y_{ik} \leq m)P(Y_{ij} \leq \ell, Y_{ik} > m)}.$$

  GCR can be modelled log-linearly, i.e.

$$\log(\gamma_{ijk}(\ell, m)) = \alpha_{\ell m}$$

  or by a regression model including covariate effects. The off-diagonal blocks of Σ_i can still be computed to construct the score equations; refer to Dale (1986), Fahrmeir & Pritscher (1996) and Gieger (1998).

- For nominal categories, local odds ratios can be used.

Marginal models for correlated responses having k categories

- The involved regression and association parameters can be estimated by a multivariate GEE approach. Details not pursued here.
- R packages `multgee`, `geepack` and `repolr` may be used to fit the above models.
- For instance, `ordgee()` in `geepack` implement the approach based on GCR by Heaberty and Zeger (1996).
- `multgee` implements the approach based on local odds ratio by Touloumis, Agresti, Kateri (2013).

Likelihood-based inference for marginal models

- The GEE approach is not likelihood-based
  $\Rightarrow$ doesn't require a full specification of the joint distribution of multivariate response vector $\mathbf{y}_i$.

- For example, for $\mathbf{y}_i = (y_{i1}, \ldots, y_{im})$, where each of y_{ij} is binary taking 0 or 1, the fully parametrized distribution has $2^m - 1$ parameters. For m marginal mean models only account for m parameters and the remaining $2^m - 1 - m$ can be specified some other ways.

- Difficulty with the likelihood-based inference is due to the difficulty in formulating this joint distribution, as well as computations; refer to the technical papers such as Glonek and McCullagh (1995) .

Marginal models for longitudinal data (§6.2.2 in F&T)

- Longitudinal data (LD) is a specific case of data with correlated responses, where short time series data

$$(y_{it}, x_{it}), \quad t = 1, \ldots, T_i$$

are available for each individual/unit/cluster $i = 1, \ldots, n$. Essentially, we simply use the notation T_i instead of m_i to emphasize repeated observations over time.

- Marginal models for LD have the exact same theory based on GEE; choices of the working association structure may borrow ideas from the times series literature. For example, the working correlation $R_i(\alpha)$ for $(y_{i1}, \ldots, y_{iT_i})$ may take the autocorrelation form

$$(R_i(\alpha))_{st} = \alpha^{|t-s|} \text{ for } s, t = 1, \ldots, T_i.$$

Example 6.4. Respiratory infection (RI) in Ohio children

Table 3: Presence and absence of **respiratory infection** (RI)

Mother did not smoke					Mother smoked				
Age of child				Frequency	Age of child				Frequency
7	8	9	10		7	8	9	10	
0	0	0	0	237	0	0	0	0	118
0	0	0	1	10	0	0	0	1	6
0	0	1	0	15	0	0	1	0	8
0	0	1	1	4	0	0	1	1	2
0	1	0	0	16	0	1	0	0	11
0	1	0	1	2	0	1	0	1	1
0	1	1	0	7	0	1	1	0	6
0	1	1	1	3	0	1	1	1	4
1	0	0	0	24	1	0	0	0	7
1	0	0	1	3	1	0	0	1	3
1	0	1	0	3	1	0	1	0	3
1	0	1	1	2	1	0	1	1	1
1	1	0	0	6	1	1	0	0	4
1	1	0	1	2	1	1	0	1	2
1	1	1	0	5	1	1	1	0	4
1	1	1	1	1	1	1	1	1	7

- Data reported for 537 children in Ohio annually from age 7 to 10. ($n = 537$ and $T = 4$)
- Analyze influence of mother's smoking status and age on the presence (1) and absence (0) of respiratory infection, using the logit model:

$$\log \frac{P(infection)}{P(noinfection)} = \beta_0 + \beta_S x_S + \beta_{A1} x_{A1} + \beta_{A2} x_{A2} + \beta_{A3} x_{A3} + \beta_{S,A1} x_S x_{A1} + \beta_{S,A2} x_S x_{A2} + \beta_{S,A3} x_S x_{A3}.$$

- Mother's smoking status is "effect-coded" as $x_S = 1$ for smoking, $x_S = -1$ for non-smoking
- Age is effect-coded with three dummies x_{A1} (Age 7) , x_{A2} (Age 8), x_{A3} (Age 9), with x_{A4} (Age 10) reserved as a reference level for the value -1.
- All three working correlation structures will be used :
    - $R = I$ (independence)
    - $R_{st} = \alpha$ for all $s \neq t$ (equicorrelation)
    - unspecified "free" R

Ohio Children: fit with all interactions

- For all three working correlations, estimates are almost identical for the first relevant digits, so only one column is given for the points estimates and the robust (sandwich) standard deviations.
- The "naive" column shows standard errors computed based on the independence correlation structure.

Table 6.8. Marginal logit model fits for Ohio children data

		Standard Deviation	
Parameter	Effect	Robust	Naive
$\hat{\beta}_0$	−1.696	0.090	0.062
$\hat{\beta}_S$	0.136	0.090	0.062
$\hat{\beta}_{A1}$	0.059	0.088	0.107
$\hat{\beta}_{A2}$	0.156	0.081	0.104
$\hat{\beta}_{A3}$	0.066	0.082	0.106
$\hat{\beta}_{SA1}$	−0.115	0.088	0.107
$\hat{\beta}_{SA2}$	0.069	0.081	0.104
$\hat{\beta}_{SA3}$	0.025	0.082	0.106

Ohio Children: fit with all interactions

- The book also report $\hat{\beta}_{A4} = -\hat{\beta}_{A1} - \hat{\beta}_{A2} - \hat{\beta}_{A3} = -0.28$ with standard dev 0.094

Table 6.8. Marginal logit model fits for Ohio children data

		Standard Deviation	
Parameter	Effect	Robust	Naive
$\hat{\beta}_0$	−1.696	0.090	0.062
$\hat{\beta}_S$	0.136	0.090	0.062
$\hat{\beta}_{A1}$	0.059	0.088	0.107
$\hat{\beta}_{A2}$	0.156	0.081	0.104
$\hat{\beta}_{A3}$	0.066	0.082	0.106
$\hat{\beta}_{SA1}$	−0.115	0.088	0.107
$\hat{\beta}_{SA2}$	0.069	0.081	0.104
$\hat{\beta}_{SA3}$	0.025	0.082	0.106

Ohio Children: fit with main effects only

- The "naive" column shows standard errors computed based on the independence correlation structure.

Table 6.9. Main effects model fits for Ohio children data

	Effect		Standard Deviation	
Parameter	Independent	Exchangeable/Unspecified	Robust	Naive
$\hat{\beta}_0$	-1.695	-1.696	0.090	0.062
$\hat{\beta}_S$	0.136	0.130	0.089	0.062
$\hat{\beta}_{A1}$	0.087	0.087	0.086	0.103
$\hat{\beta}_{A2}$	0.141	0.141	0.079	0.102
$\hat{\beta}_{A3}$	0.060	0.060	0.080	0.103

Example in Faraway §13.5

- I have also done a simple analysis with the dataset `ctsib` in section 13.5 of the applied textbook "Extending the Linear Model with R" by Julian Faraway.
- See `gee_ctsib.R`.