Chapter 6. Models for Multi-categorical Responses: Multivariate Extensions of GLM

MAST90139 Statistical Modelling for Data Science Slides

Guoqi Qian

SCHOOL OF MATHEMATICS AND STATISTICS

THE UNIVERSITY OF MELBOURNE

Outline

1. §6.1 Four examples
    - §6.1.1 Example 1. Caesarian birth study
    - §6.1.2 Example 2. Breathing test results
    - §6.1.3 Example 3. Visual impairment study
    - §6.1.4 Example 4. Ohio children respiratory infection
2. §6.2 Multicategorical response models
    - §6.2.1 Multinomial distribution and Data
    - §6.2.2 The multi-categorical logit model
    - §6.2.3 Multivariate generalised linear models (MGLM)
3. §6.3 Models for nominal responses
    - §6.3.1 Principle of maximum random utility
    - §6.3.2 Modelling explanatory variables: choice of design matrix
4. §6.4 Models for ordinal responses
    - §6.4.1 Cumulative (or threshold) models
    - §6.4.2 Extended version of cumulative models
    - §6.4.3 Link functions and design matrices for cumulative models
    - §6.4.4 Sequential models and other approaches
5. §6.5 Statistical inference for MGLM (optional)
6. §6.6 Multivariate models for correlated responses
    - §6.6.1 Conditional models (optional)
    - §6.6.2 Margin models
    - §6.6.3 Marginal models for longitudinal data

Chapter 6 Outline

In this chapter we consider the following multivariate extensions of GLM:

1. One response variable with $k > 2$ nominal categories;
2. One response variable with $k > 2$ ordinal categories;
3. Multiple correlated binary response variables.

- Cases 1 and 2 refer to the *polychotomous response* which is often modelled by a multinomial distribution.
- Case 3 is often seen in situations involving *repeated measurements* or *longitudinal data*.
- We first present four examples which will be extensively analysed in this chapter.

Example 1. Caesarian birth study (1)

Table 1: Data on classification of 251 births by caesarian section

	Caesarian planned			Caesarian not planned		
	Infection			Infection		
	I	II	non	I	II	non
Antibiotics						
Risk-factors	0	1	17	4	7	87
No risk-factors	0	0	2	0	0	0
No antibiotics						
Risk-factors	11	17	30	10	13	3
No risk-factors	4	4	32	0	0	9

- *Response* variable — `Infection`, with 3 levels (I, II, non)
- Three dichotomous *covariates* (factors)
  1. `CP`: caesarian planned or not;
  2. `RF`: risk factors (e.g. diabetes, over-weight, etc.);
  3. `AB`: antibiotics given as a prophylaxis?

Example 1. Caesarian birth study (2)

- **Aim**: Analysing the effects of the covariates on the risk of infection.
- **Note**:*Response* variable is also called *dependent* variable or *outcome* variable.
- **Note**: *Covariate* is also called *independent* variable, *explanatory* variable, or *predictor*.
- The response variable `Infection` is categorical having > 2 levels. The level values are *nominal*, but may also be treated as `ordinal` which makes the analysis even more difficult.

Example 1. Caesarian birth study (3)

- The data in Table 1 are **grouped data**. They can be converted to **ungrouped data**.

Individual	CP	RF	AB	Infection
1	1	1	1	II
2	1	1	1	non
⋮	⋮	⋮	⋮	⋮
18	1	1	1	non
19	1	0	1	non
20	1	0	1	non
21	1	1	0	I
⋮	⋮	⋮	⋮	⋮
251	0	0	0	non

CP=1 if yes and 0 if no; RF=1 if yes and 0 if no; AB=1 if yes and 0 if no.

Example 2. Breathing test results

- Forthofer & Lehnen (1981) considered the effect of age & smoking on breathing test results (BTRs) for workers in industrial plants in Texas.
- The response variable is `BTR`, ordinal with $K = 3$ levels.
- `Age` (2 levels) and `smoking status` (3 levels) are predictors.

Table 2: BTS of Houston industrial workers

		Breathing test results		
Age	Smoking status	Normal	Borderline	Abnormal
< 40	Never smoked	577	27	7
	Former smoker	192	20	3
	Current smoker	682	46	11
$40 \sim 59$	Never smoked	164	4	0
	Former smoker	145	15	7
	Current smoker	245	47	27

Example 3. Visual impairment study

Table 3: Visual impairment data, from Liang, Zeger & Qaqish (1992)

	White				Black				
Visual	Age								
impairment	40-50	51-60	61-70	70+	40-50	51-60	61-70	70+	Total
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

- Vector binary **response** variable (y_1, y_2), where $y_1 = 1$ if left-eye impaired, 0 otherwise; $y_2 = 1$ if right-eye impaired, 0 otherwise.
- **Covariates**: `Age` (yrs.), `Race` (W or B).
- **Aim**: find the effect of race and age on visual impairment.
- **Complication**: y_1 and y_2 are correlated.
- **Methods**: multivariate models for correlated responses; conditional models; asymmetric models, marginal models, GEE, etc..

Example 4. Respiratory infection in Ohio children (1)

Table 4: Presence and absence of respiratory infection

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

Example 4. Respiratory infection in Ohio children (2)

Data in Table 4 are grouped ones. They can be equivalently expressed as the following ungrouped ones:

individual	mother smoking status	Infection History			
1	0	0	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮
237	0	0	0	0	0
238	0	0	0	0	1
⋮	⋮	⋮	⋮	⋮	⋮
247	0	0	0	0	1
⋮	⋮	⋮	⋮	⋮	⋮
531	1	1	1	1	1
⋮	⋮	⋮	⋮	⋮	⋮
537	1	1	1	1	1

Example 4. Respiratory infection in Ohio children (3)

- Data come from the Harvard Study of Air Pollution and Health (Laird, Beck & Ware, 1984).
- 537 children were exam annually from age 7 to 10 on the presence or absence of R.I.
- **Response variable**: presence/absence of R.I.
- **Repeated measurements** of the response variable for each child, regarded as a short time series, may be referred to as **longitudinal data**. Responses of one child may be correlated.
- **Covariate**: mother's smoking status (regular, non-regular) at the beginning of the study.
- *Aim*: Analysing the effect of mother's smoking on children's R.I.
- *Methods*: Multivariate models for correlated responses, generalised estimating equations (GEEs), generalised linear mixed effects models (GLMM).

For examples 1 to 4, statistical modelling techniques and models are needed for performing data analysis and drawing conclusions.

§6.2.1 Multinomial distribution (1)

Multinomial distribution is an important one for multi-categorical response variables.

- Response variable Y has k levels, labelled $1, 2, \cdots, k$, i.e. $Y \in \{1, 2, \cdots, k\}$.
- Y can be represented by $\mathbf{y} = (y_1, \cdots, y_q)^T$, $q = k - 1$, with binary dummy variable $y_r = \begin{cases} 1 & \text{if } Y = r \\ 0 & \text{otherwise} \end{cases}$, $r = 1, \cdots, q$.
- Hence $Y = r \iff \mathbf{y} = (\underbrace{0, \cdots, 0, 1}_{r}, 0, \cdots, 0)^T$ and $\Pr(Y = r) = \Pr(y_r = 1 \text{ and } y_j = 0 \text{ for } j \neq r)$.
- Suppose we have m independent observations of Y: $Y_1, \cdots, Y_m$. They can be represented by $\mathbf{y}_1, \cdots, \mathbf{y}_m$, with $\mathbf{y}_i = (y_{i1}, \cdots, y_{iq})^T$, $i = 1, \cdots, m$; with $y_{ir} = I(Y_i = r)$.

§6.2.1 Multinomial distribution (2)

- Suppose $\pi_r = \Pr(Y_i = r)$, $r = 1, \cdots, k$, remain constants for all $i = 1, \cdots, m$. Note $\sum_{r=1}^{k} \pi_r = 1$ and $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_q)^T$, $q = k - 1$.
- Let $\tilde{\mathbf{y}} = \sum_{i=1}^{m} \mathbf{y}_i = \left(\sum_{i=1}^{m} y_{i1}, \cdots, \sum_{i=1}^{m} y_{iq} \right)^T \overset{denoted}{=} (\tilde{y}_1, \cdots, \tilde{y}_q)^T$ which is a vector of frequencies of $Y_i = r$, $r = 1, \cdots, q$. Then $\tilde{\mathbf{y}}$ follows a multinomial distribution $M(m, \boldsymbol{\pi})$ with pmf

$$\Pr(\tilde{\mathbf{y}} = (m_1, \cdots, m_q)) = \frac{m!}{m_1! \cdots m_q! (m - m_1 - \cdots - m_q)!} \pi_1^{m_1} \cdots \pi_q^{m_q} (1 - \pi_1 - \cdots - \pi_q)^{m - m_1 - \cdots - m_q}$$

- It can be shown that $E(\tilde{\mathbf{y}}) = (m\pi_1, \cdots, m\pi_q)^T = m\boldsymbol{\pi}$.

§6.2.1 Multinomial distribution (3)

- It can also be shown that

$$\mathrm{Cov}(\tilde{\mathbf{y}}) = m\left[\mathrm{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^T\right] = \begin{bmatrix} m\pi_1(1-\pi_1) & -m\pi_1\pi_2 & \cdots & -m\pi_1\pi_q \\ -m\pi_2\pi_1 & m\pi_2(1-\pi_2) & \cdots & -m\pi_2\pi_q \\ \vdots & \vdots & \ddots & \vdots \\ -m\pi_q\pi_1 & -m\pi_q\pi_2 & \cdots & m\pi_q(1-\pi_q) \end{bmatrix}$$

- Let $\bar{\mathbf{y}} = \tilde{\mathbf{y}}/m$ be the vector of relevant frequencies.
- It can be shown that $E(\bar{\mathbf{y}}) = (\pi_1, \cdots, \pi_q)^T = \boldsymbol{\pi}$ and

$$\mathrm{Cov}(\bar{\mathbf{y}}) = \frac{1}{m}\left[\mathrm{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^T\right] = \frac{1}{m^2}\mathrm{Cov}(\tilde{\mathbf{y}}).$$

§6.2.1 Data (1)

Data under analysis are usually of two forms:

Ungrouped data

$$\begin{array}{cccc}
 & \text{Response variable obs.} & & \text{Explanatory variables obs.} \\
\begin{array}{c} \text{Unit } 1 \\ \vdots \\ \text{Unit } i \\ \vdots \\ \text{Unit } n \end{array} &
\begin{bmatrix} y_{11} & \cdots & y_{1q} \\ \vdots & \ldots & \vdots \\ y_{i1} & \cdots & y_{iq} \\ \vdots & \ldots & \vdots \\ y_{n1} & \cdots & y_{nq} \end{bmatrix} = \begin{bmatrix} \mathbf{y}_1^T \\ \vdots \\ \mathbf{y}_i^T \\ \vdots \\ \mathbf{y}_n^T \end{bmatrix}_{n\times q} & , &
\begin{bmatrix} x_{11} & \cdots & x_{1L} \\ \vdots & \ldots & \vdots \\ x_{i1} & \cdots & x_{iL} \\ \vdots & \ldots & \vdots \\ x_{n1} & \cdots & x_{nL} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_i^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix}_{n\times L}
\end{array}$$

§6.2.1 Data (2)

Grouped data

- Suppose $\mathbf{y}_i^{(1)}, \cdots, \mathbf{y}_i^{(n_i)}$ are response observations with fixed covariate observations $\mathbf{x}_i = (x_{i1}, \cdots, x_{iL})^T$.
- Let $\bar{\mathbf{y}}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{y}_i^{(j)}$ be the mean vector over the group-i units, giving the relative frequencies of $(Y_i = 1, \cdots, Y_i = q)$.
- Let $\tilde{\mathbf{y}}_i = \sum_{j=1}^{n_i} \mathbf{y}_i^{(j)}$ be the total vector over the group-i units, giving the frequencies of $(Y_i = 1, \cdots, Y_i = q)$.

Then the **grouped data** have the following form

$$\begin{array}{ccccc}
\text{Group} & \text{Size} & \text{Response variable obs.} & & \text{Explanatory variables obs.} \\
\begin{matrix} 1 \\ \vdots \\ i \\ \vdots \\ g \end{matrix} &
\begin{matrix} n_1 \\ \vdots \\ n_i \\ \vdots \\ n_g \end{matrix} &
\begin{bmatrix} y^*_{11} & \cdots & y^*_{1q} \\ \vdots & \cdots & \vdots \\ y^*_{i1} & \cdots & y^*_{iq} \\ \vdots & \cdots & \vdots \\ y^*_{g1} & \cdots & y^*_{gq} \end{bmatrix} = \begin{bmatrix} \mathbf{y}_1^{*T} \\ \vdots \\ \mathbf{y}_i^{*T} \\ \vdots \\ \mathbf{y}_g^{*T} \end{bmatrix}_{g \times q} & , &
\begin{bmatrix} x_{11} & \cdots & x_{1L} \\ \vdots & \cdots & \vdots \\ x_{i1} & \cdots & x_{iL} \\ \vdots & \cdots & \vdots \\ x_{g1} & \cdots & x_{gL} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1^{T} \\ \vdots \\ \mathbf{x}_i^{T} \\ \vdots \\ \mathbf{x}_g^{T} \end{bmatrix}_{g \times L}
\end{array}$$

where $\mathbf{y}_i^*$ is either the group mean $\bar{\mathbf{y}}_i$ or the group total $\tilde{\mathbf{y}}_i$, $i = 1, \cdots, g$.

§6.2.2 The multi-categorical logit model (1)

- In the multinomial response case $\boldsymbol{\pi}_i = \boldsymbol{\mu}_i = E(\mathbf{y}_i|\mathbf{x}_i)$ is a $q \times 1$ vector. $\boldsymbol{\pi}_i = (\pi_{i1}, \cdots, \pi_{iq})^T$.
- The structure part of the GLM has the form $\boldsymbol{\pi}_i = \mathbf{h}(Z_i\boldsymbol{\beta})$, where
  - $\mathbf{h}(\cdot)$ is a $q \times 1$ vector-valued **response function** (also called **inverse link function**);
  - Z_i is a $q \times p$ design matrix constructed from $\mathbf{x}_i$;
  - $\boldsymbol{\beta}$ is a $p \times 1$ unknown vector parameter.
- The $q \times 1$ *vector linear predictor* for unit i is $\boldsymbol{\eta}_i = Z_i\boldsymbol{\beta}$.
- And $\left[\boldsymbol{\eta}_1^T, , \cdots, \boldsymbol{\eta}_n^T\right]^T = \text{vec}(\boldsymbol{\eta}_1, \cdots, \boldsymbol{\eta}_n)$ is the $nq \times 1$ **vector linear predictor for all n units**.

§6.2.2 The multi-categorical logit model (2)

A **multi-categorical logit model** is expressed as

$$\pi_{ir} = P(Y_i = r) = \frac{\exp\left(\beta_{r0} + \mathbf{z}_i^T \boldsymbol{\beta}_r\right)}{1 + \sum_{s=1}^{q} \exp\left(\beta_{s0} + \mathbf{z}_i^T \boldsymbol{\beta}_s\right)}, \quad r = 1, 2, \cdots, q$$

where $\mathbf{z}_i$ is an $(a-1) \times 1$ vector from $\mathbf{x}_i$. Equivalently,

$$\log \frac{\pi_{ir}}{\pi_{ik}} = \log \frac{P(Y_i = r)}{P(Y_i = k)} = \beta_{r0} + \mathbf{z}_i^T \boldsymbol{\beta}_r, \quad r = 1, 2, \cdots, q.$$

Note $\pi_{ik} = 1 - \sum_{s=1}^{q} \pi_{is}$ for all $i = 1, \cdots, n$.

- We see for this model, $\mathbf{h} = (h_1, \cdots, h_q)^T$ is a $q \times 1$ vector response function, where for each unit i,

  $$h_r = h_r(\boldsymbol{\eta}_i) = h_r(Z_i \boldsymbol{\beta}) = \frac{\exp(\eta_{ir})}{1 + \sum_{s=1}^{q} \exp(\eta_{is})}; \quad r = 1, 2, \cdots, q; \ i = 1, \cdots, n$$

  where $\eta_{ir} = \beta_{r0} + \mathbf{z}_i^T \boldsymbol{\beta}_r$ and $\boldsymbol{\eta}_i = (\eta_{i1}, \cdots, \eta_{iq})^T$.

- We also see for this model, the design matrix for each unit i is

$$Z_i = \begin{bmatrix} 1 & \mathbf{z}_i^T & 0 & \mathbf{0}^T & \cdots & 0 & \mathbf{0}^T \\ 0 & \mathbf{0}^T & 1 & \mathbf{z}_i^T & \cdots & 0 & \mathbf{0}^T \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & \mathbf{0}^T & 0 & \mathbf{0}^T & \cdots & 1 & \mathbf{z}_i^T \end{bmatrix}_{q\times(aq)}$$

- $\boldsymbol{\beta} = (\beta_{10}, \boldsymbol{\beta}_1^T, \cdots, \beta_{q0}, \boldsymbol{\beta}_q^T)^T$ is an $(aq) \times 1$ vector parameter.

$$\begin{aligned} \begin{bmatrix} \boldsymbol{\eta}_1 \\ \vdots \\ \boldsymbol{\eta}_n \end{bmatrix} &= \left[\boldsymbol{\eta}_1^T, \cdots, \boldsymbol{\eta}_n^T\right]^T = [\eta_{11}, \cdots, \eta_{1q}, \cdots, \eta_{n1}, \cdots, \eta_{nq}]^T \\ &= \left[(Z_1\boldsymbol{\beta})^T, \cdots, (Z_n\boldsymbol{\beta})^T\right]^T = \begin{bmatrix} Z_1\boldsymbol{\beta} \\ \vdots \\ Z_n\boldsymbol{\beta} \end{bmatrix} = \begin{bmatrix} Z_1 \\ \vdots \\ Z_n \end{bmatrix}_{(nq)\times(aq)} \boldsymbol{\beta} \end{aligned}$$

§6.2.2 The multi-categorical logit model (4)

- Note that the vector link function of this multi-categorical logit model is $\mathbf{g} = (g_1, \cdots, g_q)^T$, where for each unit i,

$$g_r = g_r(\pi_{i1}, \cdots, \pi_{iq}) = \log \frac{\pi_{ir}}{\pi_{ik}} = \log \frac{\pi_{iir}}{1 - (\pi_{i1} + \cdots + \pi_{iq})};$$

  $r = 1, \cdots, q$ and $i = 1, \cdots, n$.
- It shows that a multivariate GLM with multinomial responses is characterised by two specifications:
  - the vector response function $\mathbf{h}$ (or the vector link function $\mathbf{g} = \mathbf{h}^{-1}$);
  - the design matrix that depends on the covariates and the model.

§6.2.2 Example: Caesarian birth study (1)

- Consider the data on infection following Caesarian birth given in §6.1.
- Now we want to distinguish between the two different types of infection.
- Therefore the response variable Y has 3 possible outcomes: 1 for type I infection, 2 for type II infection, and 3 for no infection.
- Represent Y_i by 2 dummy variables (y_{i1}, y_{i2}), where

$$
\begin{aligned}
y_{i1} &= I(\text{the } i\text{th Caesarian was followed by infection I}) \\
y_{i2} &= I(\text{the } i\text{th Caesarian was followed by infection II})
\end{aligned}
$$

- There are 3 binary covariates: `NoPlan`, `Antib`, `RiskF`, coded by 3 dummy variables:

$$
\begin{aligned}
\texttt{NoPlan} &= I(\text{the Caesarian has not been planned}) \\
\texttt{Antib} &= I(\text{antibiotics were given as a prophylaxis}) \\
\texttt{RiskF} &= I(\text{risk factors were present})
\end{aligned}
$$

§6.2.2 Example: Caesarian birth study (2)

The data can be condensed in the grouped data (group total) form.

$$
\begin{array}{lccc}
 & \text{Size} & \text{Response variable obs.} & \text{Explanatory variables obs.} \\
 & & (\tilde{y}_{i1}, \tilde{y}_{i2}) & \texttt{NoPlan}, \texttt{Antib}, \texttt{RiskF} \\
\begin{array}{l} \text{Group 1} \\ \text{Group 2} \\ \text{Group 3} \\ \text{Group 4} \\ \text{Group 5} \\ \text{Group 6} \\ \text{Group 7} \end{array}
&
\begin{array}{c} n_1 = 40 \\ n_2 = 58 \\ n_3 = 2 \\ n_4 = 18 \\ n_5 = 9 \\ n_6 = 26 \\ n_7 = 98 \end{array}
&
\left[\begin{array}{cc} 4 & 4 \\ 11 & 17 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \\ 10 & 13 \\ 4 & 7 \end{array}\right]
&
\left[\begin{array}{ccc} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{array}\right]
\end{array}
$$

§6.2.2 Example: Caesarian birth study (3)

- We fit the following multi-categorical logit model to the data

$$\begin{aligned}
\log \frac{\pi_{i1}}{\pi_{i3}} &= \log \frac{P(\text{infection type I})}{P(\text{no infection})} = \beta_{10} + \beta_{1N}\texttt{NoPlan} + \beta_{1A}\texttt{Antib} + \beta_{1R}\texttt{RiskF} \\
\log \frac{\pi_{i2}}{\pi_{i3}} &= \log \frac{P(\text{infection type II})}{P(\text{no infection})} = \beta_{20} + \beta_{2N}\texttt{NoPlan} + \beta_{2A}\texttt{Antib} + \beta_{2R}\texttt{RiskF}
\end{aligned}$$

- Equivalently,

$$\frac{P(\text{infection type } j)}{P(\text{no infection})} = e^{\beta_{j0}} e^{\beta_{jN}\texttt{NoPlan}} e^{\beta_{jA}\texttt{Antib}} e^{\beta_{jR}\texttt{RiskF}}, \quad j = \text{I, II}.$$

- Thus, the exponential of the parameter gives the factorial contribution to the odds if the corresponding explanatory variable takes value 1 instead of 0.
- Alternatively, the exponential of the parameter may be seen as odds ratio between odds for value 1 and odds for value 0.

§6.2.2 Example: Caesarian birth study (4)

- For example, for NoPlan one obtains

$$e^{\beta_{jN}} = \frac{\frac{P(\text{infection type } j|\texttt{NoPlan=1,Antib,RiskF})}{P(\text{no infection}|\texttt{NoPlan=1,Antib,RiskF})}}{\frac{P(\text{infection type } j|\texttt{NoPlan=0,Antib,RiskF})}{P(\text{no infection}|\texttt{NoPlan=0,Antib,RiskF})}}, \quad j = \text{I}, \text{II}.$$

  where Antib and RiskF can take any fixed values.

- Note the vector-valued link function for this model is

$$\mathbf{g}(\pi_{i1}, \pi_{i2}) = \left(\log \frac{\pi_{i1}}{1 - \pi_{i1} - \pi_{i2}}, \ \log \frac{\pi_{i2}}{1 - \pi_{i1} - \pi_{i2}}\right)^T$$

§6.2.2 Example: Caesarian birth study (5)

- The estimates of the parameters and their exponential are given in the following table

Table 5: Parameter estimates in Caesarian birth study

	β	$\exp(\beta)$		β	$\exp(\beta)$
constant	-2.621	0.072	constant	-2.560	0.077
`NoPlan (I)`	1.174	3.235	`NoPlan (II)`	0.996	2.707
`Antib (I)`	-3.520	0.030	`Antib (II)`	-3.087	0.046
`RiskF (I)`	1.829	6.228	`RiskF (II)`	2.195	8.980

- `NoPlan` of 1 vs. 0 increases the type I infection odds by a factor of 3.235.
- Taking `Antib` decreases the type I infection OR to 0.030 (from 1).
- `RiskF` increases the type I infection odds ratio to 6.228.
- Similar things can be said of the type II infection odds ratio.

§6.2.2 Example: Caesarian birth study (6)

```
Caes.dat <- read.csv("D:/MAST90139/Example6-1Caes.csv")
is.data.frame(Caes.dat)   #TRUE
Caes.dat
```

```
  size noInf Inf1 Inf2 NoPlan Antib RiskF
1   40    32    4    4      0     0     0
2   58    30   11   17      0     0     1
3    2     2    0    0      0     1     0
4   18    17    0    1      0     1     1
5    9     9    0    0      1     0     0
6   26     3   10   13      1     0     1
7   98    87    4    7      1     1     1
```

```
library(nnet)
    ##package nnet is used for fitting multinomial log-linear model
    ##or multi-categorical logit model.

Caes1=multinom(as.matrix(Caes.dat[,2:4])~NoPlan+Antib+RiskF,data=Caes.dat)
```

```
# weights:  15 (8 variable)
initial  value 275.751684
iter  10 value 161.068578
final  value 160.937147
converged
```

§6.2.2 Example: Caesarian birth study (7)

```
Caes1=multinom(as.matrix(Caes.dat[,2:4])~NoPlan+Antib+RiskF,data=Caes.dat)
summary(Caes1)
Call:
multinom(formula=as.matrix(Caes.dat[,2:4]) ~ NoPlan + Antib + RiskF, data=Caes.dat)
Coefficients:
     (Intercept)    NoPlan     Antib    RiskF
Inf1   -2.621012 1.1742513 -3.520249 1.829241
Inf2   -2.559917 0.9959762 -3.087160 2.195458
Std. Errors:
     (Intercept)    NoPlan     Antib     RiskF
Inf1   0.5567209 0.5213014 0.6717416 0.6023322
Inf2   0.5462995 0.4813634 0.5498675 0.5869601

Residual Deviance: 321.8743     AIC: 337.8743
attributes(Caes1)
$names
 [1] "n"               "nunits"          "nconn"           "conn"            "nsunits"
 [6] "decay"           "entropy"         "softmax"         "censored"        "value"
[11] "wts"             "convergence"     "fitted.values"   "residuals"       "call"
[16] "terms"           "weights"         "deviance"        "rank"            "lab"
[21] "coefnames"       "vcoefnames"      "xlevels"         "edf"             "AIC"
$class
[1] "multinom" "nnet"
```

§6.2.2 Example: Caesarian birth study (8)

```
Caes1$fitted    ##fitted category probabilities

      noInf        Inf1        Inf2
1 0.8695347 0.063240568 0.067224753
2 0.4656329 0.210951192 0.323415876
3 0.9943521 0.002140051 0.003507889
4 0.9568456 0.012827892 0.030326540
5 0.6922135 0.162899513 0.144886971
6 0.2300766 0.337273035 0.432650355
7 0.8855925 0.038416539 0.075990954

Caes1$residuals
 #resid(Caes1) does the same thing, = category means - fitted category probabilities

         noInf         Inf1         Inf2
1 -0.069534679  0.036759432  0.032775247
2  0.051608447 -0.021296019 -0.030312428
3  0.005647940 -0.002140051 -0.003507889
4 -0.012401124 -0.012827892  0.025229016
5  0.307786484 -0.162899513 -0.144886971
6 -0.114691994  0.047342349  0.067349645
7  0.002162595  0.002399788 -0.004562382
```

§6.2.2 Example: Caesarian birth study (9)

```
Caes1$deviance    #residual deviance of the model, also given by deviance(Caes1)

[1] 321.8743

Caes1$edf

[1] 8

predict(Caes1, data.frame(NoPlan=1, Antib=1, RiskF=0), type="probs")

      noInf       Inf1        Inf2
0.983753294 0.006850796 0.009395909

predict(Caes1, data.frame(NoPlan=1, Antib=1, RiskF=0), type="class")

 [1] noInf
Levels: noInf Inf1 Inf2

predict(Caes1, data.frame(NoPlan=1, Antib=1, RiskF=0), type="probs", se.fit=TRUE)

      noInf       Inf1        Inf2
0.983753294 0.006850796 0.009395909
```

§6.2.2 Example: Caesarian birth study (10)

```
step(Caes1)

Start:  AIC=337.87
as.matrix(Caes.dat[, 2:4]) ~ NoPlan + Antib + RiskF

trying - NoPlan
# weights:  12 (6 variable)
...................................................................
converged
         Df      AIC
<none>    8 337.8743
- NoPlan  6 340.8649
- RiskF   6 358.3103
- Antib   6 397.4663

###The model ~ NoPlan + Antib + RiskF has the smallest AIC and is the best model.
```

§6.2.2 Example: Caesarian birth study (11)

```
Caes2=multinom(as.matrix(Caes.dat[,2:4])~NoPlan+Antib,data=Caes.dat)
summary(Caes2)
```

```
Coefficients:
     (Intercept)   NoPlan     Antib
Inf1   -1.431678 1.255595 -2.962218
Inf2   -1.058278 1.086508 -2.484319

Std. Errors:
     (Intercept)    NoPlan     Antib
Inf1   0.2870129 0.4954738 0.6408664
Inf2   0.2470012 0.4475669 0.5136334

Residual Deviance: 346.3103
AIC: 358.3103
```

```
Caes2$edf
```

```
[1] 6
```

```
Caes2$deviance-Caes1$deviance
```

```
[1] 24.43605
```

```
1-pchisq(24.43605,2)
```

```
[1] 4.940594e-06
## Hence RiskF has significant effect on the response variable.
```

§6.2.3 Extending Multicategory logit model to MGLM (1)

- Consider a multivariate $q \times 1$ response variable $\mathbf{y}_i$ for unit i. Let
$$\boldsymbol{\mu}_i = E(\mathbf{y}_i|\mathbf{x}_i), \quad i = 1, \cdots, n.$$
- **Distributional assumption:** The $\mathbf{y}_i$'s given respective $\mathbf{x}_i$'s are independent. Each $\mathbf{y}_i$ has a distribution belonging to a simple (multivariate) exponential family with pdf
$$f(\mathbf{y}_i|\boldsymbol{\theta}_i, \phi, \omega_i) = \exp\left\{\frac{\mathbf{y}_i^T\boldsymbol{\theta}_i - b(\boldsymbol{\theta}_i)}{\phi} \cdot \omega_i + c(\mathbf{y}_i, \phi, \omega_i)\right\}$$
- **Structural assumption:** $\boldsymbol{\mu}_i$'s depends on the linear predictor $\boldsymbol{\eta}_i = Z_i\boldsymbol{\beta}$ in the form
$$\boldsymbol{\mu}_i = \mathbf{h}(\boldsymbol{\eta}_i) = \mathbf{h}(Z_i\boldsymbol{\beta}) = \left[h_1(Z_i\boldsymbol{\beta}), \cdots, h_q(Z_i\boldsymbol{\beta})\right]^T, \quad i = 1, \cdots, n$$
where
    - the response function $\mathbf{h} : S \to M$ is defined on $S \subset \mathbb{R}^q$, taking values in $M \subset \mathbb{R}^q$;
    - Z_i is a $q \times p$ design matrix for unit i;
    - $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T$ is an unknown parameter vector from $B \subset \mathbb{R}^p$.

§6.2.3 Extending Multicategory logit model to MGLM (2)

- For the case of a multi-categorical response, one has to consider the multinomial distribution which may be embedded into the framework of a simple multivariate exponential family.
- For $\mathbf{y}_i = (y_{i1}, \cdots, y_{iq})^T \sim \text{Multinomial}(n_i, \boldsymbol{\pi}_i)$, the pmf of $\bar{\mathbf{y}}_i = n_i^{-1}\mathbf{y}_i$ has the form

$$f(\bar{\mathbf{y}}_i|\boldsymbol{\theta}_i, \phi, \omega_i) = \exp\left\{\frac{\bar{\mathbf{y}}_i^T\boldsymbol{\theta}_i - b(\boldsymbol{\theta}_i)}{\phi}\cdot\omega_i + c(\bar{\mathbf{y}}_i, \phi, \omega_i)\right\}, \ i = 1, \cdots, g$$

where the natural parameter $\boldsymbol{\theta}_i$ is given by

$$\boldsymbol{\theta}_i = \left[\log\frac{\pi_{i1}}{\pi_{ik}}, \cdots, \log\frac{\pi_{iq}}{\pi_{ik}}\right]^T; \quad \pi_{ik} = 1 - \sum_{j=1}^{q}\pi_{ij}; \text{ and}$$

$b(\boldsymbol{\theta}_i) = -\log(1-\pi_{i1}-\cdots-\pi_{iq}) = -\log\pi_{ik} = \log\left(1+e^{\theta_{i1}}+\cdots+e^{\theta_{iq}}\right)$;
$c(\mathbf{y}_i, \phi, \omega_i) = \log\frac{n_i!}{y_{i1}!\cdots y_{iq}!(n_i-y_{i1}-\cdots-y_{iq})!}$; and $\omega_i = n_i$.

§6.3.1 Principle of maximum random utility (1)

- Models for response variables with unordered categories may be motivated from a consideration of latent variables.
- In probabilistic choice theory, it is often assumed that an unobserved utility U_r is associated with category r of the response variable Y.
- Let U_r be a latent variable associated with category r. Now assume $U_r = u_r + \varepsilon_r$, with u_r being a fixed value and $\varepsilon_1, \cdots, \varepsilon_k$ i.i.d. having continuous cdf F.
- Following the **principle of maximum random utility**

$$Y = r \quad \Leftrightarrow \quad U_r = \max\{U_1, \cdots, U_k\}, \quad r = 1, \cdots, k.$$

Now it follows that

$$\begin{aligned} P(Y = r) &= P(U_r - U_1 \geq 0, \cdots, U_r - U_k \geq 0) \\ &= P(\varepsilon_1 \leq u_r - u_1 + \varepsilon_r, \cdots, \varepsilon_k \leq u_r - u_k + \varepsilon_r) \\ &= \int_{-\infty}^{+\infty} \prod_{s \neq r} F(u_r - u_s + \varepsilon) \cdot f(\varepsilon) d\varepsilon, \ \text{ where } f = F' \text{ is the pdf of } \varepsilon_r. \end{aligned}$$

§6.3.1 Principle of maximum random utility (2)

- Different distributional assumption for ε_r's yields different models.
- If ε_r's are i.i.d. $N(0,1)$, one gets independent **probit model** for Y. (This is true if $k=2$. But I am not able to prove it when $k>2$.)
- Try the case $k=3$.
  - Then $U_i \sim N(u_i, 1)$, $i=1,2,3$; and U_1, U_2, U_3 are independent.
  - Let $W_1 = U_1 - U_2$ and $W_2 = U_1 - U_3$. Then

$$\begin{bmatrix} W_1 \\ W_2 \end{bmatrix} \sim N\left(\begin{bmatrix} u_1 - u_2 \\ u_1 - u_3 \end{bmatrix}, \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \right), \quad \text{with joint pdf}$$

$$\begin{aligned} f(w_1,w_2) &= \frac{1}{2\sqrt{3}\pi} \exp\left\{ -\frac{1}{6} \begin{bmatrix} w_1-(u_1-u_2) \\ w_2-(u_1-u_3) \end{bmatrix}^T \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} w_1-(u_1-u_2) \\ w_2-(u_1-u_3) \end{bmatrix} \right\} \\ &= \frac{1}{2\sqrt{3}\pi} e^{\left\{-\frac{1}{3}\left[(w_1-(u_1-u_2))^2-(w_1-(u_1-u_2))(w_2-(u_1-u_3))+(w_2-(u_1-u_3))^2\right]\right\}} \end{aligned}$$

  - Then e.g. $P(Y=1)=P(U_1-U_2\geq 0, U_1-U_3\geq 0)=P(W_1\geq 0, W_2\geq 0)$
$=\int_0^\infty\int_0^\infty f(w_1,w_2)dw_1dw_2=\int_0^\infty \frac{1}{4\sqrt{\pi}} e^{-\frac{1}{4}(w_2-(u_1-u_3))^2}\left[\text{erf}\left(\frac{1}{2\sqrt{3}}[w_2-(u_1-u_3)+2(u_1-u_2)]\right)+1\right]dw_2$
where $\text{erf}(x)=\frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt = 2[\Phi(\sqrt{2}x)-\frac{1}{2}]$. This does not seem to result in a probit model.

§6.3.1 Principle of maximum random utility (3)

- If $\varepsilon_1, \cdots, \varepsilon_k$ i.i.d. following the **extreme-value distribution**

  with cdf $F(x) = \exp\{-e^{-x}\}$ and pdf $f(x) = \exp\{-e^{-x}\}\exp(-x)$,

  then we get the **multinomial logit model**

$$P(Y=r) = \frac{e^{u_r}}{\sum_{s=1}^{k} e^{u_s}} = \frac{e^{u_r-u_k}}{1+\sum_{s=1}^{q} e^{u_s-u_k}} = \frac{e^{\tilde{u}_r}}{1+\sum_{s=1}^{q} e^{\tilde{u}_s}}, \text{ with } \tilde{u}_r = u_r - u_k.$$

- We see a specific cdf F determines the link or response function of the model.
- **Proof.**

$$\begin{aligned}
P(Y=r) &= \int_{-\infty}^{+\infty} \prod_{s\neq r} \exp\{-e^{-u_r+u_s-\varepsilon}\} \cdot \exp\{-e^{-\varepsilon}\} \cdot \exp\{-\varepsilon\} d\varepsilon \\
&= \int_{-\infty}^{+\infty} e^{-\sum_{s\neq r} e^{u_s-u_r-\varepsilon} - e^{-\varepsilon}} e^{-\varepsilon} d\varepsilon = \int_{-\infty}^{+\infty} e^{-\left[\sum_{s\neq r} e^{u_s-u_r}+1\right]e^{-\varepsilon}} e^{-\varepsilon} d\varepsilon \\
&= \left.\frac{e^{-\left[\sum_{s\neq r} e^{u_s-u_r}+1\right]e^{-\varepsilon}}}{\sum_{s\neq r} e^{u_s-u_r}+1}\right|_{-\infty}^{+\infty} = \frac{1}{\sum_{s\neq r} e^{u_s-u_r}+1} = \frac{e^{u_r}}{\sum_{s=1}^{k} e^{u_s}}.
\end{aligned}$$

§6.3.2 Choice of design matrix (1)

- Explanatory variables and their coefficients in a linear predictor function $\boldsymbol{\eta}_i = Z_i\boldsymbol{\beta}$ are the terms influencing the response variable.
- Assume the response variable has multiple categories. *Explanatory variables* can be classified into 3 types:
  1. **global**: the ones whose values depend on individuals but not the response categories;
  2. **alternative-specific**: the ones whose values depend on the response categories but not the individuals;
  3. neither (1) or (2); or both (1) and (2).
- The *coefficient parameters* (which always do not depend on individuals) are classified into
  1. **global**: whose values don't depend on response categories;
  2. **category-specific**: depend on categories.

§6.3.2 Choice of design matrix (2)

Example.

- Suppose the mean utility for individual i is u_{ir}, satisfying

$$u_{ir} = \alpha_{r0} + \mathbf{z}_i^T \boldsymbol{\alpha}_r, \quad r = 1, \cdots, k; \; i = 1, \cdots, n.$$

  Then $\mathbf{z}_i$ is global and $\boldsymbol{\alpha}_r$ is category-specific.

- The multinomial logit model becomes

$$P(Y_i = r | \mathbf{z}_i) = \frac{e^{\beta_{r0} + \mathbf{z}_i^T \boldsymbol{\beta}_r}}{1 + \sum_{s=1}^{q} e^{\beta_{s0} + \mathbf{z}_i^T \boldsymbol{\beta}_s}}$$

  where $\beta_{r0} + \mathbf{z}_i^T \boldsymbol{\beta}_r = \tilde{u}_{ir} = u_{ir} - u_{ik} = (\alpha_{r0} - \alpha_{k0}) + \mathbf{z}_i^T(\boldsymbol{\alpha}_r - \boldsymbol{\alpha}_k)$.

§6.3.2 Choice of design matrix (3)

Example (continued 1)

- The associated GLM is

$$\begin{bmatrix} \pi_{i1} \\ \vdots \\ \pi_{iq} \end{bmatrix} = \begin{bmatrix} P(Y_i = 1|\mathbf{z}_i) \\ \vdots \\ P(Y_i = q|\mathbf{z}_i) \end{bmatrix} = \begin{bmatrix} \frac{e^{\eta_{i1}}}{1+\sum_{s=1}^{q} e^{\eta_{is}}} \\ \vdots \\ \frac{e^{\eta_{iq}}}{1+\sum_{s=1}^{q} e^{\eta_{is}}} \end{bmatrix} \overset{denoted}{=} \mathbf{h}(\boldsymbol{\eta}_i)$$

where the linear predictor has the form $\boldsymbol{\eta}_i = Z_i\boldsymbol{\beta}$, with the design matrix for individual i being

$$Z_i = \begin{bmatrix} 1 & \mathbf{z}_i^T & 0 & \mathbf{0}^T & \cdots & 0 & \mathbf{0}^T \\ 0 & \mathbf{0}^T & 1 & \mathbf{z}_i^T & \cdots & 0 & \mathbf{0}^T \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & \mathbf{0}^T & 0 & \mathbf{0}^T & \cdots & 1 & \mathbf{z}_i^T \end{bmatrix}_{q\times p}$$

and $\boldsymbol{\beta} = (\beta_{10}, \boldsymbol{\beta}_1^T, \cdots, \beta_{q0}, \boldsymbol{\beta}_q^T)^T$ is a $p \times 1$ parameter vector.

§6.3.2 Choice of design matrix (4)

Example (continued 2)

- If u_{ir} also depends on some alternative-specific explanatory variables, then u_{ir} may be assumed of the form

$$u_{ir} = \alpha_{r0} + \mathbf{z}_i^T \boldsymbol{\alpha}_r + \mathbf{w}_r^T \gamma, \quad r = 1, \cdots, k; \; i = 1, \cdots, n$$

  where $\mathbf{w}_r$ is an alternative-specific explanatory vector variable and γ is a global parameter.

- Then $\tilde{u}_{ir} = u_{ir} - u_{ik} = \beta_{r0} + \mathbf{z}_i^T \boldsymbol{\beta}_r + (\mathbf{w}_r - \mathbf{w}_k)^T \gamma$.
- The resultant multinomial logit model will have the following design matrix for individual i

$$Z_i = \begin{bmatrix} 1 & \mathbf{z}_i^T & 0 & \mathbf{0}^T & \cdots & 0 & \mathbf{0}^T & (\mathbf{w}_1 - \mathbf{w}_k)^T \\ 0 & \mathbf{0}^T & 1 & \mathbf{z}_i^T & \cdots & 0 & \mathbf{0}^T & (\mathbf{w}_2 - \mathbf{w}_k)^T \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \mathbf{0}^T & 0 & \mathbf{0}^T & \cdots & 1 & \mathbf{z}_i^T & (\mathbf{w}_q - \mathbf{w}_k)^T \end{bmatrix}_{q \times p}$$

  and $\boldsymbol{\beta} = (\beta_{10}, \boldsymbol{\beta}_1^T, \cdots, \beta_{q0}, \boldsymbol{\beta}_q^T, \gamma^T)^T$ is a $p \times 1$ parameter vector.

§6.4 Models for ordinal responses

- Response variable having $k > 2$ categories may be of ordinal nature.
    - **Example**: Breathing test results
- Ordinal variables may come from quite different mechanisms. Two distinguished types are: **grouped continuous** and **assessed ordered** categorical variables.
    - **Grouped continuous** — merely a categorized continuous variable
    - **Assessed ordered** — obtained after an assessor's judgment.

§6.4.1 Cumulative models: the threshold approach

- Suppose there is an unobserved continuous variable U for the response variable Y having k categories.
- Suppose Y is determined by U in the following way

$$Y = r \quad \Leftrightarrow \quad \theta_{r-1} < U \leq \theta_r; \quad r = 1, \cdots, k$$

  where $-\infty = \theta_0 < \theta_1 < \cdots < \theta_k = +\infty$ unknown.
  That means Y is a coarser (categorized) version of U determined by the thresholds $\theta_1, \cdots, \theta_{k-1}$.
- Suppose U depends on a $p \times 1$ vector explanatory variable $\mathbf{x}$ in the following way

$$U = -\mathbf{x}^T \boldsymbol{\gamma} + \varepsilon$$

  where $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_p)^T$ is an unknown vector parameter and ε is random with cdf F.
- It follows that

$$P(Y \leq r|\mathbf{x}) = P(U \leq \theta_r) = F(\theta_r + \mathbf{x}^T \gamma).$$

  This is called a **cumulative model** or **threshold model** with cdf F.

§6.4.1 Cumulative logistic (or proportional odds) model (1)

- If choose $F(x) = \left[1 + e^{-x}\right]^{-1}$ which is the cdf of a logistic distribution, then

$$P(Y \le r|\mathbf{x}) = \frac{e^{\theta_r + \mathbf{x}^T\gamma}}{1 + e^{\theta_r + \mathbf{x}^T\gamma}}, \quad r = 1, \cdots, q = k - 1. \tag{1}$$

- It is equivalent to

$$\log \frac{P(Y \le r|\mathbf{x})}{P(Y > r|\mathbf{x})} = \theta_r + \mathbf{x}^T\gamma \qquad \text{or} \tag{2}$$

$$\frac{P(Y \le r|\mathbf{x})}{P(Y > r|\mathbf{x})} = \exp\{\theta_r + \mathbf{x}^T\gamma\} \tag{3}$$

The model given by (1), (2) or (3) is called the **cumulative logistic model**.

§6.4.1 Cumulative logistic (or proportional odds) model (2)

- The cumulative logistic model is also called a **proportional odds model** due to following interpretation:
- Suppose there are 2 populations (groups) identified by $\mathbf{x}_1$ and $\mathbf{x}_2$ respectively. Then

$$\frac{P(Y \le r|\mathbf{x}_1)}{P(Y > r|\mathbf{x}_1)} = e^{\theta_r + \mathbf{x}_1^T \gamma} \quad \text{and} \quad \frac{P(Y \le r|\mathbf{x}_2)}{P(Y > r|\mathbf{x}_2)} = e^{\theta_r + \mathbf{x}_2^T \gamma}.$$

It follows that

$$\text{cumulative odds ratio} = \frac{P(Y \le r|\mathbf{x}_1)/P(Y > r|\mathbf{x}_1)}{P(Y \le r|\mathbf{x}_2)/P(Y > r|\mathbf{x}_2)} = e^{(\mathbf{x}_1 - \mathbf{x}_2)^T \gamma}$$

which is the same for all $r = 1, \cdots, q$.

§6.4.1 Grouped Cox (or proportional hazards) model

- If choose F as the cdf of the extreme minimal-value distribution, i.e. $F(x) = 1 - \exp\{-e^x\}$, then
$$\begin{aligned} P(Y \leq r|\mathbf{x}) &= 1 - \exp\{-e^{\theta_r + \mathbf{x}^T\gamma}\} \qquad (4) \\ \Longleftrightarrow \quad \underbrace{\log\left[-\log P(Y > r|\mathbf{x})\right]}_{\text{complementary log-log link}} &= \theta_r + \mathbf{x}^T\gamma, \quad r = 1, \cdots, q \end{aligned}$$
Model (4) is referred to as the **grouped Cox model** or **proportional hazards model** due to the following interpretation:
- For 2 populations (groups) identified by $\mathbf{x}_1$ and $\mathbf{x}_2$ respectively
$$\log P(Y > r|\mathbf{x}_1) = \log P(Y > r|\mathbf{x}_2) \cdot e^{(\mathbf{x}_1 - \mathbf{x}_2)^T\gamma}, \quad r = 1, \cdots, q$$
Differentiating w.r.t. r on both sides results in
$$\lambda(r|\mathbf{x}_1) = \lambda(r|\mathbf{x}_2) \cdot e^{(\mathbf{x}_1 - \mathbf{x}_2)^T\gamma}, \quad \Leftrightarrow \quad \frac{\lambda(r|\mathbf{x}_1)}{\lambda(r|\mathbf{x}_2)} = e^{(\mathbf{x}_1 - \mathbf{x}_2)^T\gamma},$$
which is the same for all $r = 1, \cdots, q$. Here $\lambda(r|\cdot) = -\frac{d}{dr}\log P(Y > r|\cdot)$ is the **hazard function**.

§6.4.1 Cumulative model based on extreme maximal-value CDF

- If choose F as the cdf of the extreme maximal-value distribution, i.e. $F(x) = \exp\{-e^{-x}\}$, then

$$\begin{aligned} P(Y \le r|\mathbf{x}) &= \exp\{-e^{-(\theta_r + \mathbf{x}^T\gamma)}\} \qquad (5) \\ \Longleftrightarrow \quad \underbrace{\log\left[-\log P(Y \le r|\mathbf{x})\right]}_{\text{log-log link}} &= -(\theta_r + \mathbf{x}^T\gamma), \; r = 1, \cdots, q. \qquad (6) \end{aligned}$$

Model (5) or (6) is referred to as the **extreme maximal-value distribution model**.

S6.4.2 Extended version of cumulative models (1)

- Recall that an ordinal categorical response Y with k categories is determined by a latent variable U as

  $$Y = r \;\Leftrightarrow\; \theta_{r-1} < U \leq \theta_r; \quad r = 1, \cdots, k; -\infty = \theta_0 < \theta_1 < \cdots < \theta_k = +\infty$$

  and $U = -\mathbf{x}^T\gamma + \varepsilon$, with ε having cdf F.
- Then a **cumulative model** or **threshold model** with cdf F is

  $$P(Y \leq r|\mathbf{x}) = P(U \leq \theta_r) = F(\theta_r + \mathbf{x}^T\gamma). \tag{7}$$

- One may assume a linear form for thresholds $\theta_1, \cdots, \theta_q$,

  $$\theta_r = \beta_{r0} + \mathbf{w}^T\boldsymbol{\beta}_r$$

  where $\mathbf{w}$ is a vector explanatory variable and $\boldsymbol{\beta}_r = (\beta_{r1}, \cdots, \beta_{rm})^T$ is a *category-specific* parameter vector.
- Then an **extended cumulative model** is obtained:

  $$P(Y \leq r|\mathbf{x}, \mathbf{w}) = F(\beta_{r0} + \mathbf{w}^T\boldsymbol{\beta}_r + \mathbf{x}^T\gamma). \tag{8}$$

S6.4.2 Extended version of cumulative models (2)

Remarks

1. Model (8) becomes unidentifiable if $\mathbf{w}_i = \mathbf{x}_i$ for each individual $i = 1, \cdots, n$; i.e., it will not be possible to separate γ from each $\boldsymbol{\beta}_r$ if $\mathbf{w}_i = \mathbf{x}_i$.
2. So $\mathbf{w}_i$ and $\mathbf{x}_i$ must not be the same.
3. $\boldsymbol{\beta}_r$ can be global, while γ can be replaced by a category-specific parameter.
4. $\mathbf{w}_i$ are called **threshold variables**, while $\mathbf{x}_i$ are called **shift variables**.

§6.4.3 Link functions for cumulative models

- For a cumulative GLM, the link function $\mathbf{g}=(g_1,\cdots,g_q)$, as a vector function of $(\pi_1 = P(Y=1),\cdots,\pi_q = P(Y=q))$, is given by

$$g_r(\pi_1,\cdots,\pi_q) = F^{-1}(\pi_1+\cdots+\pi_r), \quad r=1,\cdots,q,$$

  where $F^{-1}(\cdot)$ is the inverse function for F that specifies the cumulative model

$$\pi_1+\cdots+\pi_r = P(Y\le r|\mathbf{x},\mathbf{w}) = F(\beta_{r0}+\mathbf{w}^T\boldsymbol{\beta}_r+\mathbf{x}^T\gamma).$$

- F being the $N(0,1)$ cdf gives the **probit link**

$$g_r(\pi_1,\cdots,\pi_q) = \Phi^{-1}(\pi_1+\cdots+\pi_r), \quad r=1,\cdots,q.$$

- For proportional odds model, the link is the **cumulative logit**

$$g_r(\pi_1,\cdots,\pi_q) = \log\frac{\pi_1+\cdots+\pi_r}{1-(\pi_1+\cdots+\pi_r)}, \quad r=1,\cdots,q.$$

- For proportional hazards model, the link is the **complement log-log**

$$g_r(\pi_1,\cdots,\pi_q) = \log\{-\log(1-(\pi_1+\cdots+\pi_r))\}, \quad r=1,\cdots,q.$$

§6.4.3 Design matrices for cumulative models (1)

- For the simple cumulative model

$$P(Y_i \leq r|\mathbf{x}_i) = F(\theta_r + \mathbf{x}_i^T \gamma),$$

the design matrix Z_i in the linear predictor $\boldsymbol{\eta}_i = Z_i \boldsymbol{\beta}$ is given by

$$Z_i = \begin{bmatrix} 1 & & & & \mathbf{x}_i^T \\ & 1 & & & \mathbf{x}_i^T \\ & & \ddots & & \vdots \\ & & & 1 & \mathbf{x}_i^T \end{bmatrix}_{q\times(q+\dim(\mathbf{x}_i))}$$

and

$$\boldsymbol{\beta} = (\theta_1, \cdots, \theta_q, \boldsymbol{\gamma}^T)^T.$$

Here $\mathbf{x}_i$ should not contain a constant component.

§6.4.3 Design matrices for cumulative models (2)

- For the extended cumulative model

$$P(Y_i \leq r|\mathbf{x}_i, \mathbf{w}_i) = F(\beta_{r0} + \mathbf{w}_i^T \boldsymbol{\beta}_r + \mathbf{x}_i^T \gamma),$$

the design matrix Z_i in the linear predictor $\boldsymbol{\eta}_i = Z_i \boldsymbol{\beta}$ is given by

$$Z_i = \begin{bmatrix} 1 & \mathbf{w}_i^T & & & & & \mathbf{x}_i^T \\ & & 1 & \mathbf{w}_i^T & & & \mathbf{x}_i^T \\ & & & \ddots & \ddots & & \vdots \\ & & & & 1 & \mathbf{w}_i^T & \mathbf{x}_i^T \end{bmatrix}_{q\times(q(1+\dim(\mathbf{w}_i))+\dim(\mathbf{x}_i))}$$

and

$$\boldsymbol{\beta} = (\beta_{10}, \boldsymbol{\beta}_1^T, \cdots, \beta_{q0}, \boldsymbol{\beta}_q^T, \gamma^T)^T.$$

Here neither $\mathbf{w}_i$ nor $\mathbf{x}_i$ should contain a constant component.

§6.4.3 Alternative formulation/parameterisation in cumulative models (1)

- The threshold parameters $\theta_1, \cdots, \theta_q$ in cumulative models $P(Y \leq r|\mathbf{x}) = F(\theta_r + \mathbf{x}^T\gamma)$ are restricted by

$$\theta_1 < \theta_2 < \cdots < \theta_q$$

- The corresponding restriction for the extended cumulative model is

$$\beta_{10} + \mathbf{w}^T\boldsymbol{\beta}_1 < \beta_{20} + \mathbf{w}^T\boldsymbol{\beta}_2 < \cdots < \beta_{q0} + \mathbf{w}^T\boldsymbol{\beta}_q$$

  which is more complex.

- If such constraints are not explicitly accounted for in estimation, the iterative estimation procedure may fail due to fitting inadmissible parameters.

§6.4.3 Alternative formulation/parameterisation in cumulative models (2)

- For the simple cumulative model, the problem can be avoided by using an alternative formulation or reparameterisation:

$$\begin{aligned} & \alpha_1 = \theta_1, \quad \alpha_r = \log(\theta_r - \theta_{r-1}), \quad r = 2, \cdots, q \\ \Longleftrightarrow \quad & \theta_1 = \alpha_1, \quad \theta_r = \theta_1 + \sum_{i=2}^{r} e^{\alpha_i}, \quad r = 2, \cdots, q. \end{aligned}$$

  The new parameters $\alpha_1, \cdots, \alpha_q$ are unconstrained.
- The cumulative model under the new formulation has a linear structure of the following form

$$\begin{aligned} F^{-1}\left(P(Y = 1|\mathbf{x})\right) &= \alpha_1 + \mathbf{x}^T\gamma \\ \log\left[F^{-1}\left(P(Y \le r|\mathbf{x})\right) - F^{-1}\left(P(Y \le r-1|\mathbf{x})\right)\right] &= \alpha_r, \quad r = 2, \cdots, q. \end{aligned}$$

- The link function corresponding to the new formulation can be determined accordingly.

§6.4.3 Alternative formulation/parameterisation in cumulative models (3)

- For the special case of the *cumulative logistic model*, one obtains the following link function (**log-logit link**)

$$\begin{aligned} g_1(\pi_1, \cdots, \pi_q) &= \log\left(\frac{\pi_1}{1-\pi_1}\right) \\ g_r(\pi_1, \cdots, \pi_q) &= \log\left[\log\left\{\frac{\pi_1+\cdots+\pi_r}{1-\pi_1-\cdots-\pi_r}\right\} - \log\left\{\frac{\pi_1+\cdots+\pi_{r-1}}{1-\pi_1-\cdots-\pi_{r-1}}\right\}\right], \\ & q = 2, \cdots, q. \end{aligned}$$

- If this alternative link function is used, the design matrix Z_i has to be adapted. Now

$$Z_i = \begin{bmatrix} 1 & & & & \mathbf{x}_i^T \\ & 1 & & & \mathbf{0} \\ & & \ddots & & \vdots \\ & & & 1 & \mathbf{0} \end{bmatrix}_{q\times(q+\dim(\mathbf{x}_i))}$$

and the parameter vector becomes $\boldsymbol{\beta} = (\alpha_1, \cdots, \alpha_q, \gamma^T)^T$.

Here $\mathbf{x}_i$ should not contain a constant component.

§6.4.3 Computing packages and an example

- The `polr()` function in the R **MASS** package can be used to fit cumulative models (but not extended cumulative models).
- More generally, the **ordinal** package in R can be used for fitting ordinal regression models.

Example 2 in §6.1. Breathing test results.

- The response variable is `BTR`, ordinal with $k = 3$ levels.
- `Age` (2 levels) and `Smoking` status (3 levels) are predictors.
- The aim is to see how `Age` and `Smoking` are associated with `BTR`.
- Command `polr` in R `MASS` package is used in the analysis.

§6.4.3 Breathing testing example (1)

Usage of `polr`:

```
polr(formula, data, weights, start, ..., subset, na.action,
     contrasts = NULL, Hess = FALSE, model = TRUE,
     method = c("logistic", "probit", "loglog", "cloglog", "cauchit"))
```

Note **the estimates of γ are those from** `polr` **multiplied by** -1 due to the model used in `polr` being $P(Y \leq r|\mathbf{x}) = F(\theta_r - \mathbf{x}^T\gamma)$.

```
> library(MASS); BTR.dat <- read.csv("D:/MAST90139/Example6-2BTR.csv"); BTR.dat

       BTR    Age  Smoking Freq
1  1Normal    <40   1Never  577
2  2Border    <40   1Never   27
3  3Abnorm    <40   1Never    7
4  1Normal    <40  2Former  192
5  2Border    <40  2Former   20
6  3Abnorm    <40  2Former    3
7  1Normal    <40 3Current  682
8  2Border    <40 3Current   46
9  3Abnorm    <40 3Current   11
10 1Normal 40to59   1Never  164
11 2Border 40to59   1Never    4
12 3Abnorm 40to59   1Never    0
13 1Normal 40to59  2Former  145
14 2Border 40to59  2Former   15
15 3Abnorm 40to59  2Former    7
16 1Normal 40to59 3Current  245
17 2Border 40to59 3Current   47
18 3Abnorm 40to59 3Current   27
```

§6.4.3 Breathing testing example (2)

```
> is.factor(BTR.dat$BTR)    #TRUE;     > is.ordered(BTR.dat$BTR)    #FALSE
> is.factor(BTR.dat$Age)    #TRUE;     > is.factor(BTR.dat$Smoking)  #TRUE
> is.factor(BTR.dat$Freq) #FALSE;
> BTR.dat$BTR=as.ordered(BTR.dat$BTR)   #Set BTR as an ordinal factor
> help(polr)
> BTR.logit=polr(BTR~Age+Smoking+Age:Smoking, data=BTR.dat, weights=Freq,
                Hess=T,method="logistic")
> summary(BTR.logit)
```

```
Coefficients:
                            Value Std. Error t value
Age40to59                 -0.8857     0.5359  -1.653
Smoking2Former             0.6973     0.2822   2.471
Smoking3Current            0.3472     0.2238   1.551
Age40to59:Smoking2Former   1.1458     0.6228   1.840
Age40to59:Smoking3Current  2.2007     0.5689   3.868

Intercepts:
                Value   Std. Error t value
1Normal|2Border  2.8334  0.1764     16.0603
2Border|3Abnorm  4.3078  0.2130     20.2210

Residual Deviance: 1564.968
AIC: 1578.968
```

§6.4.3 Breathing testing example (3)

```
BTR.logit1=polr(BTR~Age+Smoking,data=BTR.dat, weights=Freq,Hess=T,method="logistic")
summary(BTR.logit1)
```

```
Coefficients:
                 Value Std. Error t value
Age40to59       0.7772     0.1482   5.243
Smoking2Former  0.7815     0.2333   3.350
Smoking3Current 0.9607     0.1918   5.010

Intercepts:
                Value   Std. Error t value
1Normal|2Border  3.1927  0.1749    18.2544
2Border|3Abnorm  4.6543  0.2125    21.9076

Residual Deviance: 1589.744
AIC: 1599.744
```

```
anova(BTR.logit, BTR.logit1)
```

```
Likelihood ratio tests of ordinal regression models

Response: BTR
                        Model Resid. df Resid. Dev   Test    Df LR stat.   Pr(Chi)
1               Age + Smoking      2214   1589.744
2 Age + Smoking + Age:Smoking      2212   1564.968 1 vs 2     2 24.77585 4.168e-06
```

§6.4.3 Breathing testing example (4)

```
> BTR.logit$fitted
     1Normal    2Border     3Abnorm
1  0.9444565 0.04225961 0.013283873
2  0.9444565 0.04225961 0.013283873
3  0.9444565 0.04225961 0.013283873
4  0.8943685 0.07930628 0.026325236
5  0.8943685 0.07930628 0.026325236
6  0.8943685 0.07930628 0.026325236
7  0.9231700 0.05813458 0.018695369
8  0.9231700 0.05813458 0.018695369
9  0.9231700 0.05813458 0.018695369
10 0.9763207 0.01815786 0.005521451
11 0.9763207 0.01815786 0.005521451
12 0.9763207 0.01815786 0.005521451
13 0.8671630 0.09895796 0.033879045
14 0.8671630 0.09895796 0.033879045
15 0.8671630 0.09895796 0.033879045
16 0.7633793 0.17036521 0.066255462
17 0.7633793 0.17036521 0.066255462
18 0.7633793 0.17036521 0.066255462
> BTR.logit$residual  #does not exist.
> BTR.logit$lp  #linear predictor values
         1          2          3          4          5          6          7          8          9
 0.0000000  0.0000000  0.0000000  0.6972823  0.6972823  0.6972823  0.3472245  0.3472245  0.3472245
        10         11         12         13         14         15         16         17         18
-0.8857462 -0.8857462 -0.8857462  0.9573393  0.9573393  0.9573393  1.6621466  1.6621466  1.6621466
> attributes(BTR.logit)
 [1] "coefficients"  "zeta"          "deviance"      "fitted.values" "lev"           "terms"
 [7] "df.residual"   "edf"           "n"             "nobs"          "call"          "method"
[13] "convergence"   "niter"         "lp"            "Hessian"       "model"         "contrasts"
[19] "xlevels"
```

§6.4.3 Breathing testing example (5)

- On the previous 3 slides we fit two cumulative logistic models to `BTR` using `Age` and `Smoking`: the first (`BTR.logit`) includes the `Age:Smokings` interaction-effect terms as well as the main-effect terms; while the second (`BTR.logit1`) includes main-effect terms only.
- Dummy coding (i.e. `contr.treatment` coding in R) is used for representing `Age` and `Smoking`:

$$\begin{aligned}
\texttt{Age40to59} &= \begin{cases} 1, & \text{if age is } 40 \sim 59 \\ 0 & \text{if age} < 40 \end{cases} \\
\texttt{Smoking2Former} &= \begin{cases} 1, & \text{former smoker} \\ 0 & \text{else} \end{cases} , \\
\texttt{Smoking3Current} &= \begin{cases} 1, & \text{current smoker} \\ 0 & \text{else} \end{cases}
\end{aligned}$$

§6.4.3 Breathing testing example (6)

- The first model can be written as

$$\log \frac{P(\texttt{BTR} = \texttt{Normal})}{P(\texttt{BTR} > \texttt{Normal})} = \theta_1 + \gamma_1 \texttt{Age40to59} + \gamma_2 \texttt{Smoking2Former} + \gamma_3 \texttt{Smoking3Current}$$
$$+ \gamma_4 \texttt{Age40to59} \cdot \texttt{Smoking2Former} + \gamma_5 \texttt{Age40to59} \cdot \texttt{Smoking3Current}$$
$$\log \frac{P(\texttt{BTR} \le \texttt{Bordering})}{P(\texttt{BTR} > \texttt{Bordering})} = \theta_2 + \gamma_1 \texttt{Age40to59} + \gamma_2 \texttt{Smoking2Former} + \gamma_3 \texttt{Smoking3Current}$$
$$+ \gamma_4 \texttt{Age40to59} \cdot \texttt{Smoking2Former} + \gamma_5 \texttt{Age40to59} \cdot \texttt{Smoking3Current}$$

- MLEs of the parameters in the first model are

$$\hat{\theta}_1 = 2.833, \hat{\theta}_2 = 4.308, \hat{\gamma}_1 = 0.886, \hat{\gamma}_2 = -0.697, \hat{\gamma}_3 = -0.347, \hat{\gamma}_4 = -1.146, \hat{\gamma}_5 = -2.201$$

with their standard errors given in the R output.

- Estimated values of the γ parameters can be interpreted as log cumulative odds ratios. **For example**, the odds of having a normal test result (or normal or bordering result) for a worker aged 40 to 59 who never smoked is estimated to be $e^{-\hat{\gamma}_2 - \hat{\gamma}_4} = e^{0.697+1.146} = 6.32$ times of that for a worker aged 40 to 59 who is a former worker.

§6.4.3 Breathing testing example (7)

- The second model can be written as

$$\log \frac{P(\texttt{BTR} = \texttt{Normal})}{P(\texttt{BTR} > \texttt{Normal})} = \theta_1 + \gamma_1 \texttt{Age40to59} + \gamma_2 \texttt{Smoking2Former} + \gamma_3 \texttt{Smoking3Current}$$

$$\log \frac{P(\texttt{BTR} \leq \texttt{Bordering})}{P(\texttt{BTR} > \texttt{Bordering})} = \theta_2 + \gamma_1 \texttt{Age40to59} + \gamma_2 \texttt{Smoking2Former} + \gamma_3 \texttt{Smoking3Current}$$

- MLEs of the parameters in the first model are

$$\hat{\theta}_1 = 3.193,\ \hat{\theta}_2 = 4.654,\ \hat{\gamma}_1 = -0.777,\ \hat{\gamma}_2 = -0.782,\ \hat{\gamma}_3 = -0.961$$

  with their standard errors given in the R output.
- "$H_0 : \gamma_4 = \gamma_5 = 0$ vs. H_1 : not H_0" is the hypothesis about the `Age:Smoking` interaction effects on `BTR`.
- A χ^2 analysis of deviance test comparing models `BTR.logit` and `BTR.logit1` gives an LR statistic of 22.776 and p-value of 4.168×10^{-6}. Therefore, there is very strong evidence to reject H_0. We conclude that `Age` and `Smoking` have very strong interaction effects on `BTR`.

§6.4.3 Breathing testing example (8)

- Linear predictor of the first model is

$$\begin{aligned}\eta_1 &= \theta_1 + \gamma_1 \texttt{Age40to59} + \gamma_2 \texttt{Smoking2Former} + \gamma_3 \texttt{Smoking3Current} \\ &\quad + \gamma_4 \texttt{Age40to59} \cdot \texttt{Smoking2Former} + \gamma_5 \texttt{Age40to59} \cdot \texttt{Smoking3Current} \\ \eta_2 &= \theta_2 + \gamma_1 \texttt{Age40to59} + \gamma_2 \texttt{Smoking2Former} + \gamma_3 \texttt{Smoking3Current} \\ &\quad + \gamma_4 \texttt{Age40to59} \cdot \texttt{Smoking2Former} + \gamma_5 \texttt{Age40to59} \cdot \texttt{Smoking3Current}\end{aligned}$$

- However, the linear predictor returned from `polr` is just $\mathbf{x}^T\boldsymbol{\gamma}$, i.e.

$$\gamma_1 \texttt{Age40to59} + \gamma_2 \texttt{Smoking2Former} + \gamma_3 \texttt{Smoking3Current} \\ + \gamma_4 \texttt{Age40to59} \cdot \texttt{Smoking2Former} + \gamma_5 \texttt{Age40to59} \cdot \texttt{Smoking3Current}$$

  thus has just 6 different values knowing that there are 6 combinations of `Age` and `Smoking`.

- Knowing these tricks, there should be no difficulty to reveal how the fitted values returned by `BTR.logit$fitted` are obtained.

§6.4.3 Breathing testing example (9)

- Effect coding (i.e. `contr.sum` coding in R, -1 for last category) may also be used for representing `Age` and `Smoking`:

$$\texttt{Age1} = \begin{cases} 1, & \text{if age} < 40 \\ -1, & \text{if age is } 40 \sim 59 \end{cases}$$

$$\texttt{Smoking1} = \begin{cases} 1, & \text{never smoked} \\ 0, & \text{former smoker} \\ -1, & \text{current smoker} \end{cases}, \quad \texttt{Smoking2} = \begin{cases} 0, & \text{never smoked} \\ 1, & \text{former smoker} \\ -1, & \text{current smoker} \end{cases}$$

- With this coding, the first model can be written as

$$\begin{aligned} \log \frac{P(\texttt{BTR} = \texttt{Normal})}{P(\texttt{BTR} > \texttt{Normal})} &= \theta_1 + \gamma_1 \texttt{Age1} + \gamma_2 \texttt{Smoking1} + \gamma_3 \texttt{Smoking2} \\ &\quad + \gamma_4 \texttt{Age1} \cdot \texttt{Smoking1} + \gamma_5 \texttt{Age1} \cdot \texttt{Smoking2} \\ \log \frac{P(\texttt{BTR} \le \texttt{Bordering})}{P(\texttt{BTR} > \texttt{Bordering})} &= \theta_2 + \gamma_1 \texttt{Age1} + \gamma_2 \texttt{Smoking1} + \gamma_3 \texttt{Smoking2} \\ &\quad + \gamma_4 \texttt{Age1} \cdot \texttt{Smoking1} + \gamma_5 \texttt{Age1} \cdot \texttt{Smoking2} \end{aligned}$$

- Although the parameter estimates will be different now, other results from the model will be the same as that using dummy coding.

§6.4.3 Breathing testing example (10)

```
> options(contrasts = c("contr.sum", "contr.poly"))
> BTR.logit=polr(BTR~Age+Smoking+Age:Smoking, data=BTR.dat, weights=Freq, Hess=T,method="logistic")
> summary(BTR.logit)

Coefficients:
                 Value Std. Error t value
Age1          -0.11487     0.1086 -1.0580
Smoking1      -0.90596     0.1890 -4.7933
Smoking2       0.36428     0.1421  2.5644
Age1:Smoking1  0.55777     0.1890  2.9512
Age1:Smoking2 -0.01517     0.1421 -0.1068

Intercepts:
                Value   Std. Error t value
1Normal|2Border  2.3704  0.1086    21.8209
2Border|3Abnorm  3.8447  0.1599    24.0486

Residual Deviance: 1564.968
AIC: 1578.968

> BTR.logit1=polr(BTR~Age+Smoking, data=BTR.dat, weights=Freq, Hess=T,method="logistic")
> anova(BTR.logit, BTR.logit1)

Likelihood ratio tests of ordinal regression models

Response: BTR
                          Model Resid. df Resid. Dev   Test    Df LR stat.      Pr(Chi)
1                Age + Smoking      2214   1589.744
2 Age + Smoking + Age:Smoking      2212   1564.968 1 vs 2     2 24.77585 4.168618e-06
```

§6.4.3 Breathing testing example (11)

```
> attributes(BTR.logit)

$names
 [1] "coefficients"  "zeta"          "deviance"      "fitted.values" "lev"           "terms"         "df.residu
 [8] "edf"           "n"             "nobs"          "call"          "method"        "convergence"   "niter"
[15] "lp"            "Hessian"       "model"         "contrasts"     "xlevels"

> BTR.logit$fitted

     1Normal    2Border     3Abnorm
1  0.9444565 0.04225911 0.013284361
2  0.9444565 0.04225911 0.013284361
3  0.9444565 0.04225911 0.013284361
4  0.8943660 0.07930714 0.026326874
5  0.8943660 0.07930714 0.026326874
6  0.8943660 0.07930714 0.026326874
7  0.9231672 0.05813603 0.018696801
8  0.9231672 0.05813603 0.018696801
9  0.9231672 0.05813603 0.018696801
10 0.9763222 0.01815653 0.005521309
11 0.9763222 0.01815653 0.005521309
12 0.9763222 0.01815653 0.005521309
13 0.8671585 0.09895993 0.033881540
14 0.8671585 0.09895993 0.033881540
15 0.8671585 0.09895993 0.033881540
16 0.7633676 0.17037058 0.066261785
17 0.7633676 0.17037058 0.066261785
18 0.7633676 0.17037058 0.066261785

> BTR.logit$lp    #linear predictor values

         1          2          3          4          5          6          7          8          9         10
-0.4630592 -0.4630592 -0.4630592  0.2342498  0.2342498  0.2342498 -0.1157938 -0.1157938 -0.1157938 -1.3488684
        11         12         13         14         15         16         17         18
-1.3488684 -1.3488684  0.4943191  0.4943191  0.4943191  1.1991525  1.1991525  1.1991525
```

§6.4.3 Breathing testing example (12)

We also fit a grouped Cox or proportional hazards model.

```
> BTR.ph=polr(BTR~Age+Smoking+Age:Smoking,data=BTR.dat, weights=Freq,
                      Hess=T, method="cloglog")
> summary(BTR.ph)

Coefficients:
                            Value Std. Error t value
Age40to59                 -0.2865    0.14264  -2.008
Smoking2Former             0.2125    0.10054   2.114
Smoking3Current            0.1088    0.07301   1.490
Age40to59:Smoking2Former   0.4333    0.18941   2.288
Age40to59:Smoking3Current  0.8379    0.16357   5.122

Intercepts:
                Value   Std. Error t value
1Normal|2Border  1.0477  0.0558    18.7612
2Border|3Abnorm  1.5528  0.0629    24.6916

Residual Deviance: 1559.949
AIC: 1573.949
```

§6.4.3 Breathing testing example (13)

We further fit a cumulative model with the extreme maximal-value distribution.

```
> BTR.extm=polr(BTR~Age+Smoking+Age:Smoking,data=BTR.dat,weights=Freq,
                              Hess=T,method="loglog")
> summary(BTR.extm)

Coefficients:
                            Value Std. Error t value
Age40to59                 -0.8672     0.5286  -1.640
Smoking2Former             0.6755     0.2700   2.501
Smoking3Current            0.3368     0.2167   1.554
Age40to59:Smoking2Former   1.1003     0.6070   1.813
Age40to59:Smoking3Current  2.0724     0.5573   3.719

Intercepts:
                Value   Std. Error t value
1Normal|2Border  2.8614  0.1715    16.6838
2Border|3Abnorm  4.2761  0.2080    20.5605

Residual Deviance: 1566.335
AIC: 1580.335
```

§6.4.3 Breathing testing example (14)

We fit a grouped Cox or proportional hazards model using the effect coding.

```
> options(contrasts = c("contr.sum", "contr.poly"))
> BTR.ph=polr(BTR~Age+Smoking+Age:Smoking,data=BTR.dat, weights=Freq,
                           Hess=T, method="cloglog")
> summary(BTR.ph)

Coefficients:
                 Value Std. Error  t value
Age1          -0.06862    0.03421 -2.00575
Smoking1      -0.31898    0.05366 -5.94408
Smoking2       0.11022    0.04961  2.22168
Age1:Smoking1  0.21188    0.05361  3.95205
Age1:Smoking2 -0.00481    0.04960 -0.09699

Intercepts:
                Value   Std. Error t value
1Normal|2Border  0.8720  0.0353    24.7072
2Border|3Abnorm  1.3771  0.0441    31.2213

Residual Deviance: 1559.949
AIC: 1573.949
```

§6.4.3 Breathing testing example (15)

We finally fit a cumulative model with the extreme maximal-value distribution using the effect coding.

```
> options(contrasts = c("contr.sum", "contr.poly"))
> BTR.extm=polr(BTR~Age+Smoking+Age:Smoking,data=BTR.dat,weights=Freq,
                               Hess=T,method="loglog")
> summary(BTR.extm)

Coefficients:
                Value Std. Error t value
Age1         -0.09523     0.1053  -0.904
Smoking1     -0.86618     0.1854  -4.671
Smoking2      0.35943     0.1361   2.642
Age1:Smoking1  0.52877     0.1854   2.852
Age1:Smoking2 -0.02136     0.1361  -0.157

Intercepts:
                Value   Std. Error t value
1Normal|2Border  2.4288  0.1054    23.0536
2Border|3Abnorm  3.8434  0.1573    24.4373

Residual Deviance: 1566.335
AIC: 1580.335
```

§6.4.4 Motivating example for sequential models

- In many applications the ordering of the response categories is due to a sequential mechanism. The categories are ordered since they can be reached only successively.

Example: Tonsil size. Children have been classified according to their tonsil size and whether or not they are carriers of Streptococcus pyogenes.

Table 6: Tonsil size and Streptococcus pyogenes (Holmes & Williams, 1954)

	Present but not enlarged	Enlarged	Greatly enlarged
Carriers	19	29	24
Noncarriers	497	560	269

- It may be assumed that tonsil size always starts in the normal state "present but not enlarged" (*category 1*). If the tonsils grow abnormally, they may become "enlarged" (*category 2*); if the process does not stop, tonsils may become "greatly enlarged" (*category 3*). But in order to get greatly enlarged tonsils, they first have to be enlarged for the duration of the intermediate state "enlarged."

§6.4.4 Sequential models (1)

- Let latent variables U_r, $r = 1, \cdots, q$ with $q = k - 1$ have the linear form
$$U_r = -\mathbf{x}^T \gamma + \varepsilon_r, \quad \text{where } \varepsilon_r \text{ has cdf } F.$$
- We assume the ordinal response variable Y is determined by U_r's in the following sequential way:
$$\begin{aligned} Y = 1 \quad &\Leftrightarrow \quad U_1 \leq \theta_1 \\ Y = 2 \text{ given } Y \geq 2 \quad &\Leftrightarrow \quad U_2 \leq \theta_2; \quad \text{and in general} \\ Y = r \text{ given } Y \geq r \quad &\Leftrightarrow \quad U_r \leq \theta_r; \quad \text{or equivalently} \\ Y > r \text{ given } Y \geq r \quad &\Leftrightarrow \quad U_r > \theta_r, \quad r = 1, \cdots, q \end{aligned} \tag{9}$$
- The sequential mechanism (9) models the transition from category r to category $r + 1$ given that category r is reached.
- The main difference to the threshold approach used in the cumulative model is the conditional modelling of the transitions. The sequential mechanism assumes a binary decision in each step.

§6.4.4 Sequential models (2)

- The sequential response mechanism (9) combined with the linear form of the latent variable

  $$U_r = -\mathbf{x}^T\gamma + \varepsilon_r, \quad \text{where } \varepsilon_r \text{ has cdf } F,$$

  leads to the **sequential model** with cdf F

  $$P(Y = r | Y \geq r, \mathbf{x}) = P(U_r \leq \theta_r) = F(\theta_r + \mathbf{x}^T\gamma), \quad r = 1, \cdots, k, \tag{10}$$

  where $\theta_k = +\infty$. It follows that

  $$\begin{aligned} P(Y = r|\mathbf{x}) &= P(Y = r|Y \geq r, \mathbf{x}) \cdot P(Y \geq r|\mathbf{x}) \\ &= F(\theta_r + \mathbf{x}^T\gamma) \cdot P(Y > r-1|Y \geq r-1, \mathbf{x}) \cdot \\ &\qquad P(Y > r-2|Y \geq r-2, \mathbf{x}) \cdots P(Y > 1|Y \geq 1, \mathbf{x}) \\ &= F(\theta_r + \mathbf{x}^T\gamma)[1 - F(\theta_{r-1} + \mathbf{x}^T\gamma)] \cdots [1 - F(\theta_1 + \mathbf{x}^T\gamma)] \\ &= F(\theta_r + \mathbf{x}^T\gamma) \prod_{i=1}^{r-1}[1 - F(\theta_i + \mathbf{x}^T\gamma)], \quad r = 1, \cdots, k; \quad \text{Note } \textstyle\prod_{i=1}^{0}[\cdot] = 0 \end{aligned}$$

- Model (10) is also called the **continuation ratio model**.

§6.4.4 Sequential models (3)

- Note that there is no need to impose any order restriction on $\theta_1, \cdots, \theta_q$ in (10).
- Now specify a cdf F will give a specific sequential model.

1. Choosing $F(x) = (1 + e^{-x})^{-1}$ gives the **sequential logit model**

$$P(Y = r|Y \geq r, \mathbf{x}) = \frac{\exp(\theta_r + \mathbf{x}^T\gamma)}{1 + \exp(\theta_r + \mathbf{x}^T\gamma)}, \quad r = 1, \cdots, k$$

which is equivalent to $\log\left\{\frac{P(Y = r|\mathbf{x})}{P(Y > r|\mathbf{x})}\right\} = \theta_r + \mathbf{x}^T\gamma.$

2. Choosing $F(x) = 1 - \exp(-e^x)$, the sequential model has the form

$$P(Y = r|Y \geq r, \mathbf{x}) = 1 - \exp\left(-e^{\theta_r + \mathbf{x}^T\gamma}\right), \quad r = 1, \cdots, k \qquad (11)$$

which is equivalent to

$$\log\left[-\log\left\{\frac{P(Y > r|\mathbf{x})}{P(Y \geq r|\mathbf{x})}\right\}\right] = \theta_r + \mathbf{x}^T\gamma, \quad r = 1, \cdots, k. \qquad (12)$$

§6.4.4 Sequential models (4)

- It is worth to note that (12) is equivalent to the cumulative model with $F(x) = 1 - \exp(-e^x)$. This means (12) is a special parametric form of the grouped Cox model. This equivalence follows by the reparameterisation:

$$\theta_r = \log\{e^{\tilde{\theta}_r} - e^{\tilde{\theta}_{r-1}}\}, \quad r = 1, \cdots, k-1.$$

Note $\tilde{\theta}_0 = -\infty$ and $\theta_1 = \tilde{\theta}_1$. Further $\tilde{\theta}_r = \log\left(\sum_{i=1}^{r} e^{\theta_i}\right)$.

Substituting $\theta_r = \log\{e^{\tilde{\theta}_r} - e^{\tilde{\theta}_{r-1}}\}$ into (12), one obtains

$$-\log P(Y > r|\mathbf{x}) + \log P(Y > r-1|\mathbf{x}) = e^{\tilde{\theta}_r + \mathbf{x}^T\gamma} - e^{\tilde{\theta}_{r-1} + \mathbf{x}^T\gamma}, \quad r = 1, \cdots, q$$

$$\Rightarrow \quad -\log P(Y > r|\mathbf{x}) = e^{\tilde{\theta}_r + \mathbf{x}^T\gamma}, \quad r = 1, \cdots, q$$

$\Rightarrow \quad P(Y \le r|\mathbf{x}) = 1 - \exp\left(-e^{\tilde{\theta}_r + \mathbf{x}^T\gamma}\right)$ which is a **grouped Cox model**.

§6.4.4 Sequential models (5)

3. Choosing exponential cdf $F(x) = 1 - e^{-x}$, the **exponential sequential model** becomes

$$\begin{aligned} & P(Y = r|Y \geq r, \mathbf{x}) = 1 - e^{-(\theta_r + \mathbf{x}^T\gamma)} \\ \Leftrightarrow \quad & -\log\left\{\frac{P(Y>r|\mathbf{x})}{P(Y\geq r|\mathbf{x})}\right\} = \theta_r + \mathbf{x}^T\gamma, \quad r = 1, \cdots, q. \end{aligned}$$

Generalised sequential models:
If assuming $\theta_r = \delta_{r0} + \mathbf{z}^T\boldsymbol{\delta}_r$, where $\boldsymbol{\delta}_r = (\delta_{r1}, \cdots, \delta_{rm})^T$ is a category-specific vector parameter, then one gets the *Generalised sequential model*:

$$P(Y = r|Y \geq r, \mathbf{x}, \mathbf{z}) = F(\delta_{r0} + \mathbf{z}^T\boldsymbol{\delta}_r + \mathbf{x}^T\gamma) \tag{13}$$

Remark: The effect of **z** is non-homogeneous over the categories, while the effect of **x** is homogeneous.

§6.4.4 Link functions and design matrices in sequential models

- Response and link functions may be derived directly from model (10).
- The link function $\mathbf{g} = (g_1, \cdots, g_q)^T$ is given by

$$g_r(\pi_1, \cdots, \pi_q) = F^{-1}\left(\frac{\pi_r}{1 - \pi_1 - \cdots - \pi_{r-1}}\right), \quad r = 1, \cdots, q$$

  and the response function $\mathbf{h} = (h_1, \cdots, h_q)^T$ has the form

$$h_r(\eta_1, \cdots, \eta_q) = F(\eta_r) \prod_{i=1}^{r-1} (1 - F(\eta_i)), \quad r = 1, \cdots, q.$$

- For the sequential logit model, with $r = 1, \cdots, q$

$$g_r(\pi_1, \cdots, \pi_q) = \log \frac{\pi_r}{1 - \pi_1 - \cdots - \pi_{r-1}} \quad \text{and} \quad h_r(\eta_1, \cdots, \eta_q) = e^{\eta_r} \prod_{i=1}^{r-1} (1 + e^{\eta_i})^{-1}.$$

- In regard to the design matrices there is no difference between sequential and cumulative models. Thus the design matrices from §6.3.3 apply.

§6.4.4 Strict stochastic ordering (optional)

- **Strict stochastic ordering** is a concept generalizing the *proportional odds* for simple cumulative model (7)

$$\Delta_c(\mathbf{x}_1, \mathbf{x}_2) = F^{-1}\left(P(Y \le r|\mathbf{x}_1)\right) - F^{-1}\left(P(Y \le r|\mathbf{x}_2)\right).$$

  If $\Delta_c(\mathbf{x}_1, \mathbf{x}_2) = \gamma^T(\mathbf{x}_1 - \mathbf{x}_2)$, we say strict stochastic ordering holds.
- Then for general cumulative model (8) where $\mathbf{w} = \mathbf{x}$ is set and $\mathbf{x}^T\gamma$ is omitted, one can test strict stochastic ordering by testing $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \cdots = \boldsymbol{\beta}_q$.
- The concept can be generalized to the sequential model (10) as well. Let

$$\Delta_s(\mathbf{x}_1, \mathbf{x}_2) = F^{-1}\left(P(Y = r|Y \ge r, \mathbf{x}_1)\right) - F^{-1}\left(P(Y = r|Y \ge r, \mathbf{x}_2)\right).$$

  If $\Delta_s(\mathbf{x}_1, \mathbf{x}_2) = \gamma^T(\mathbf{x}_1 - \mathbf{x}_2)$, we say strict stochastic ordering holds for the sequential model.

§6.4.4 Two-step models (optional) (1)

- It may be more appropriate to divide the ordered categories into sets of categories with very homogeneous responses in the first step, then model the categories within each set separately in the second step. This is the **two-step modelling approach**.
- **Example: Rheumatoid arthritis.** Mehta, Patel & Tsiatis (1984) analysed data of patients with acute rheumatoid arthritis. A new agent was compared with an active control, and each patient was evaluated on a 5-point assessment scale.

Table 7: Clinical trial of a new agent and an active control

	Global assessment				
Drug	Much improved	Improved	No change	Worse	Much worse
New agent	24	37	21	19	6
Active control	11	51	22	21	7

Global assessment may be subdivided into“improvement”, “no change” and “worse” first. Then “improvement” is split up into “much improved” and “improved” and the “worse” category is split into “worse” and “much worse.”

§6.4.4 Two-step models (optional) (2)

- Specifically, let the categories $1, \cdots, k$ be subdivided into t basic sets $S_1, \cdots, S_t$, where $S_j = \{m_{j-1}+1, \cdots, m_j\}$, $m_0 = 0$ and $m_t = k$.

  In Step 1: $Y \in S_j \quad \Leftrightarrow \quad \theta_{j-1} < U_0 \le \theta_j, \;\; j = 1, \cdots, t.$
  Assume $U_0 = -\mathbf{x}^T\gamma_0 + \varepsilon$.

  In Step 2: $Y = R \mid Y \in S_j \quad \Leftrightarrow \quad \theta_{j,r-1} < U_j \le \theta_{jr}.$
  Assume $U_j = -\mathbf{x}^T\gamma_j + \varepsilon_j$. Also $\varepsilon, \varepsilon_j$'s are iid with cdf F.

- Then the two-step model becomes

$$P(Y \in T_j|\mathbf{x}) = F(\theta_j + \mathbf{x}^T\gamma_0), \quad P(Y \le r|Y \in S_j, \mathbf{x}) = F(\theta_{jr} + \mathbf{x}^T\gamma_j) \tag{14}$$

  where $T_j = S_1 \cup \cdots \cup S_j, \quad \theta_1 < \cdots < \theta_{t-1}, \quad \theta_t = \infty;$
  $\theta_{j,m_{j-1}+1} < \cdots < \theta_{j,m_{j-1}}, \quad \theta_{j,m_j} = \infty, \quad j = 1, \cdots, t.$

- Since cumulative mechanisms are used in both steps, (14) is called a **two-step cumulative model**. **Two-step sequential model** can be similarly defined.

§6.4.4 Alternative approaches (optional)

1. Anderson (1984).

$$P(Y = r|\mathbf{x}) = \frac{e^{\beta_{r0} - \phi_r \boldsymbol{\beta}^T \mathbf{x}}}{1 + \sum_{i=1}^{q} e^{\beta_{i0} - \phi_i \boldsymbol{\beta}^T \mathbf{x}}}, \quad r = 1, \cdots, q;$$

   where $1 = \phi_1 > \cdots > \phi_k = 0$.

2. Williams and Grizzle (1972). $\displaystyle\sum_{r=1}^{k} s_r P(Y = r|\mathbf{x}) = \mathbf{x}^T \boldsymbol{\beta}$

   where $s_1, \cdots, s_k$ are some given scores for the categories.

3. **Adjacent categories logit model** (Agresti 1984).

$$\log \frac{P(Y = r|\mathbf{x})}{P(Y = r-1|\mathbf{x})} = \mathbf{x}^T \boldsymbol{\beta}_r$$
$$\Leftrightarrow \quad P(Y = r|Y \in \{r, r+1\}, \mathbf{x}) = F(\mathbf{x}^T \boldsymbol{\beta}_r)$$

   where F is the logistic cdf.

§6.5 Statistical inference for MGLM overview

1. MLE
2. Testing of linear hypotheses
3. Goodness of fit (model adequacy) statistics
4. Power-divergence statistics

§6.5.1 Maximum likelihood estimation in MGLM (1)

- An extension of the exponential family is the **multivariate exponential family**, having the pdf

$$f(\mathbf{y}_i|\boldsymbol{\theta}_i, \phi, \omega_i) = \exp\left\{\frac{\mathbf{y}_i^T\boldsymbol{\theta}_i - b(\boldsymbol{\theta}_i)}{\phi}\cdot\omega_i + c(\mathbf{y}_i, \phi, \omega_i)\right\}$$

- For $q \times 1$ vector observations $\mathbf{y}_1, \cdots, \mathbf{y}_n$ which are suppose to be independent of each other, let $\boldsymbol{\mu}_i = E(\mathbf{y}_i|\mathbf{x}_i)$, $i = 1, \cdots, n$, with $\mathbf{x}_i$ being the covariate vector.
- The **structural assumption** of MGLM is that $\boldsymbol{\mu}_i$'s depend on the linear predictor $\boldsymbol{\eta}_i = Z_i\boldsymbol{\beta}$ in the form

$$\boldsymbol{\mu}_i = \mathbf{h}(\boldsymbol{\eta}_i) = \mathbf{h}(Z_i\boldsymbol{\beta}) = \left[h_1(Z_i\boldsymbol{\beta}), \cdots, h_1(Z_i\boldsymbol{\beta})\right]^T, \quad i = 1, \cdots, n$$

  where
  - the response function $\mathbf{h} : S \to M$ is defined on $S \subset \mathbb{R}^q$, taking values in $M \subset \mathbb{R}^q$;
  - Z_i is a $q \times p$ design matrix for unit i;
  - $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T$ is an unknown parameter vector from $B \subset \mathbb{R}^p$.

§6.5.1 Maximum likelihood estimation in MGLM (2)

- The MLE of $\boldsymbol{\beta}$ may be derived in analogy to the one-dimensional case.
- The log-likelihood kernel for observation $\mathbf{y}_i$ is

$$\ell_i(\boldsymbol{\mu}_i) = \frac{\mathbf{y}_i^T \boldsymbol{\theta}_i - b(\boldsymbol{\theta}_i)}{\phi} \cdot \omega_i, \quad \text{where } \boldsymbol{\theta}_i = \boldsymbol{\theta}(\boldsymbol{\mu}_i) \tag{15}$$

- Using the relation $\boldsymbol{\mu}_i = \mathbf{h}(\boldsymbol{\eta}_i) = \mathbf{h}(Z_i\boldsymbol{\beta})$, it can be found that the score function is $\mathbf{s}(\boldsymbol{\beta}) = \dfrac{\partial \ell}{\partial \boldsymbol{\beta}} = \displaystyle\sum_{i=1}^{n} \mathbf{s}_i(\boldsymbol{\beta})$, with

$$\mathbf{s}_i(\boldsymbol{\beta}) = Z_i^T D_i(\boldsymbol{\beta}) \Sigma_i^{-1}(\boldsymbol{\beta})[\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})], \tag{16}$$

where $D_i(\boldsymbol{\beta}) = \dfrac{\partial \mathbf{h}^T(\boldsymbol{\eta}_i)}{\partial \boldsymbol{\eta}_i}$ is the derivative matrix of $\mathbf{h}^T(\boldsymbol{\eta}_i)$ evaluated at $\boldsymbol{\eta}_i = Z_i\boldsymbol{\beta}$, which is also called the Jacobian (matrix). And $\Sigma_i(\boldsymbol{\beta}) = \text{cov}(\mathbf{y}_i)$.

§6.5.1 Maximum likelihood estimation in MGLM (3)

- An alternative form for $\mathbf{s}_i(\boldsymbol{\beta})$ is

$$\mathbf{s}_i(\boldsymbol{\beta}) = Z_i^T W_i(\boldsymbol{\beta}) \frac{\partial \mathbf{g}(\boldsymbol{\mu}_i)}{\partial \boldsymbol{\mu}_i^T} [\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})], \quad \text{where}$$

$$W_i(\boldsymbol{\beta}) = D_i(\boldsymbol{\beta}) \Sigma_i^{-1}(\boldsymbol{\beta}) D_i(\boldsymbol{\beta}) = \left\{ \frac{\partial \mathbf{g}(\boldsymbol{\mu}_i)}{\partial \boldsymbol{\mu}_i^T} \Sigma_i(\boldsymbol{\beta}) \frac{\partial \mathbf{g}^T(\boldsymbol{\mu}_i)}{\partial \boldsymbol{\mu}_i} \right\}^{-1} \tag{17}$$

  which approximately equals $(\text{cov}(\mathbf{g}(\mathbf{y}_i))^{-1}$.

- The expected Fisher information matrix is

$$F(\boldsymbol{\beta}) = \text{cov}(\mathbf{s}(\boldsymbol{\beta})) = \sum_{i=1}^{n} Z_i^T W_i(\boldsymbol{\beta}) Z_i.$$

§6.5.1 Maximum likelihood estimation in MGLM (4)

- In matrix form the above may be rewritten as

$$\begin{aligned} \mathbf{s}(\boldsymbol{\beta}) &= Z^T D(\boldsymbol{\beta})\Sigma^{-1}(\boldsymbol{\beta})[\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\beta})] \\ F(\boldsymbol{\beta}) &= Z^T W(\boldsymbol{\beta}) Z. \end{aligned}$$

  where $\mathbf{y} = (\mathbf{y}_1^T, \cdots, \mathbf{y}_n^T)^T$, $\boldsymbol{\mu}(\boldsymbol{\beta}) = (\boldsymbol{\mu}_1(\boldsymbol{\beta})^T, \cdots, \boldsymbol{\mu}_n(\boldsymbol{\beta})^T)^T$, $\Sigma(\boldsymbol{\beta}) = \text{diag}\,(\Sigma_i(\boldsymbol{\beta}))$, $W(\boldsymbol{\beta}) = \text{diag}\,(W_i(\boldsymbol{\beta}))$, $D(\boldsymbol{\beta}) = \text{diag}\,(D_i(\boldsymbol{\beta}))$, all being block diagonal, and the total design matrix $Z = \left[Z_1^T, \cdots, Z_n^T\right]^T$.

- In the situation of grouped data, $\mathbf{y}_i$ is the mean vector over n_i observations and $\Sigma_i(\boldsymbol{\beta})$ is replaced by $n_i^{-1}\Sigma_i(\boldsymbol{\beta})$.

- Under regularity conditions,

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

§6.5.1 Numerical computing in MGLM

- This is the same as in the univariate case, except using the multivariate versions.
  1. Working or pseudo observation vector
  $$\tilde{\mathbf{y}}(\beta) = (\tilde{\mathbf{y}}_1(\beta)^T, \cdots, \tilde{\mathbf{y}}_n(\beta)^T)^T, \quad \text{where}$$
  $$\tilde{\mathbf{y}}_i(\beta) = Z_i\beta + \left(D_i^{-1}(\beta)\right)^T [\mathbf{y}_i - \boldsymbol{\mu}_i(\beta)]$$
  is an approximation for $\mathbf{g}(\boldsymbol{\mu}_i(\beta))$.
  2. The IRWLS estimate is
  $$\hat{\beta}^{(k+1)} = \left(Z^T W(\hat{\beta}^{(k)})Z\right)^{-1} Z^T W(\hat{\beta}^{(k)})\tilde{\mathbf{y}}(\beta^{(k)})$$
  which is equivalent to the Fisher scoring iteration
  $$\hat{\beta}^{(k+1)} = \hat{\beta}^{(k)} + \left(Z^T W(\hat{\beta}^{(k)})Z\right)^{-1} \mathbf{s}(\beta^{(k)}).$$

§6.5.2 Testing linear hypotheses in MGLM

The linear hypotheses

$$H_0 : C\boldsymbol{\beta} = \boldsymbol{\xi} \quad \text{vs.} \quad H_1 : C\boldsymbol{\beta} \neq \boldsymbol{\xi}$$

can be tested in the same way as in the univariate GLM, except replacing the score and Fisher information by their multivariate versions.

§6.5.2 Goodness-of-fit statistics in MGLM (1)

Goodness of fit (or model adequacy) of the MGLM can be assessed based on the Pearson statistics and the deviance (the data need be grouped as much as possible).

- **General Pearson Statistics:**

$$\chi^2 = \sum_{i=1}^{g} (\mathbf{y}_i - \hat{\boldsymbol{\mu}}_i)^T \Sigma_i^{-1}(\hat{\beta})(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_i)$$

- In the case of multi-categorical response variable with multinomial distribution $n_i\mathbf{y}_i \sim M(n_i, \boldsymbol{\pi}_i)$, we have

$$\chi^2 = \sum_{i=1}^{g} \chi_p^2(\mathbf{y}_i, \hat{\boldsymbol{\pi}}_i) = \sum_{i=1}^{g} n_i \sum_{j=1}^{k} \frac{(y_{ij} - \hat{\pi}_{ij})^2}{\hat{\pi}_{ij}}$$

with $y_{ik} = 1 - y_{i1} - \cdots - y_{iq}$ and $\hat{\pi}_{ik} = 1 - \hat{\pi}_{i1} - \cdots - \hat{\pi}_{iq}$.

§6.5.2 Goodness-of-fit statistics in MGLM (2)

- **Deviance** or **likelihood ratio (LR) statistics**:

$$D = -2\sum_{i=1}^{g}\{\ell_i(\hat{\boldsymbol{\pi}}_i) - \ell_i(\mathbf{y}_i)\}$$

.

- For multinomial data,

$$D = 2\sum_{i=1}^{g}\chi_D^2(\mathbf{y}_i, \hat{\boldsymbol{\pi}}_i) = 2\sum_{i=1}^{g} n_i \sum_{j=1}^{k} y_{ij}\log\frac{y_{ij}}{\hat{\pi}_{ij}}$$

- Under regularity conditions,

$$\chi^2 \overset{a}{\sim} \chi^2(g(k-1)-p) \quad \text{and} \quad D \overset{a}{\sim} \chi^2(g(k-1)-p)$$

  where p is the number of estimated parameters.

- For sparse data with small n_i, one should use alternative asymtotics.

§6.5.3 Power-divergence family (1)

For categorical data a more general single-parameter family of goodness-of-fit statistics is the **power-divergence family**, introduced by Cressie and Read (1984, JRSSB).
The power-divergence statistic with parameter $\lambda \in \mathbb{R}$ is given by

$$S_\lambda = \sum_{i=1}^{g} \mathrm{SD}_\lambda(\mathbf{y}_i, \hat{\boldsymbol{\pi}}_i)$$

where the sum of deviations over observations at $\mathbf{y}_i$ is

$$\mathrm{SD}_\lambda(\mathbf{y}_i, \hat{\boldsymbol{\pi}}_i) = \frac{2n_i}{\lambda(\lambda+1)} \sum_{j=1}^{k} y_{ij} \left[\left(\frac{y_{ij}}{\hat{\pi}_{ij}} \right)^{\lambda} - 1 \right], \quad -\infty < \lambda < \infty \qquad (18)$$

§6.5.3 Power-divergence family (2)

1. At $\lambda = 1$, $S_1 = \sum_{i=1}^{g} n_i \sum_{j=1}^{k} \frac{(y_{ij} - \hat{\pi}_{ij})^2}{\hat{\pi}_{ij}}$, the usual Pearson statistic.
2. At $\lambda \to 0$, $S_0 = 2 \sum_{i=1}^{g} n_i \sum_{j=1}^{k} y_{ij} \log \frac{y_{ij}}{\hat{\pi}_{ij}}$, the likelihood ratio statistic.
3. At $\lambda \to -1$, $S_{-1} = 2 \sum_{i=1}^{g} n_i \sum_{j=1}^{k} \hat{\pi}_{ij} \log \frac{\hat{\pi}_{ij}}{y_{ij}}$, Kullback's minimum discrimination information statistic.
4. At $\lambda = -2$, $S_{-2} = \sum_{i=1}^{g} n_i \sum_{j=1}^{k} \frac{(y_{ij} - \hat{\pi}_{ij})^2}{y_{ij}}$, Neyman's minimum modified χ^2-statistic.
5. At $\lambda = -\frac{1}{2}$, $S_{-\frac{1}{2}} = 4 \sum_{i=1}^{g} \sum_{j=1}^{k} (\sqrt{n_i y_{ij}} - \sqrt{n_i \hat{\pi}_{ij}})^2$, Freeman-Tukey statistic.

In general, $\lambda \in [-1, 2]$ is recommended.

§6.5.3 Asymptotic properties under classical “fixed cells” assumptions

Classical assumptions in asymptotic theory for grouped data imply

1. a fixed number of groups (g is fixed)
2. increasing sample sizes $n_i \to \infty, i = 1, \cdots, g$, such that $\frac{n_i}{n} \to \lambda_i$ for fixed proportions $\lambda_i > 0$, $i = 1, \cdots, g$.
3. a fixed “number of cells” k in each group
4. a fixed number of parameters.

If these assumptions hold and in addition, the model under consideration is correct, then

$$S_\lambda \overset{a}{\sim} \chi^2(g(k-1)-p), \quad p \text{ is the } \# \text{ of estimated parameters.}$$

And under local alternative hypothesis, the limit distribution of S_λ is non-central χ^2.

§6.5.3 Asymptotic properties under sparseness and "increasing-cells" assumptions

- If several explanatory variables are considered, the number of observations for a fixed explanatory variable **x** is often small and the usual asymptotic machinery will fail.
- Under such sparseness conditions, it may be assumed that with increasing sample size $n \to \infty$ the number of groups (values of the explanatory variables) is also increasing with $g \to \infty$, resulting in the "increasing-cells" setting.
- Now S_λ is no longer χ^2-distributed in the limit. Rather S_λ has an asymptotic normal distribution under H_0 that the model holds. Details are not pursued here.

§6.6 Multivariate models for correlated responses

- So far we have been assuming there is only one response variable, possibly multi-categorical, being under consideration.
- Now we consider the situations where a vector of correlated or clustered response variables is observed, together with covariates, for each unit in the sample.
- These situations often happen in longitudinal studies, repeated measurements studies, and grouped (clustered) studies, etc.
- When the response variables are approximately Gaussian, multivariate linear models are commonly used for the underlying data analysis, which has been well established.
- The situations become more difficult when the response variables are discrete, categorical or mixed discrete/categorical/continuous.
- Three main approaches will be introduced here:
  1. **conditional models**
  2. **marginal models**
  3. **random effects models**

§6.6.1 Conditional models: Asymmetric models (optional)1

Asymmetric models

- In many applications, the components of a response vector are ordered in a way that some components are prior to the other components, e.g. if they refer to events that take place earlier.
- Consider the simplest case of a response vector $\mathbf{Y} = (Y_1, Y_2)$, with categorical components $Y_1 \in \{1, \cdots, k_1\}$ and $Y_2 \in \{1, \cdots, k_2\}$. Let Y_2 refer to events that may be considered conditional on Y_1.
    - **An example**: response years in school (Y_1) and the statement about happiness (Y_2).
- Then Y_1 may be modelled to be dependent on the explanatory variables $\mathbf{x}$, using a model introduced in §6.2 to §6.4.
- And Y_2 may be modelled based on Y_1 and $\mathbf{x}$, also using a model introduced in §6.2 to §6.4.

§6.6.1 Conditional models: Asymmetric models (optional)2

Asymmetric models

- In general, consider m categorical responses $Y_1, \cdots, Y_m$, where Y_j depends on $Y_1, \cdots, Y_{j-1}$ but not on $Y_{j+1}, \cdots, Y_m$. Models that make use of this dependence structure are based on the decomposition

$$P(Y_1, \cdots, Y_m|\mathbf{x}) = P(Y_1|\mathbf{x}) \cdot P(Y_2|Y_1, \mathbf{x}) \cdots P(Y_m|Y_1, \cdots, Y_{m-1}, \mathbf{x}) \tag{19}$$

- Simple models arise if each component in (19) is specified by a GLM:

$$P(Y_j = r|Y_1, \cdots, Y_{j-1}, \mathbf{x}) = h_j(Z_j\beta) \tag{20}$$

  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 $\mathbf{x}$. Models of form (20) are sometimes called *data-driven* ones.

§6.6.1 Conditional models: Asymmetric models (optional)3

- **Markov-type transition models** follow from the additional assumption $P(Y_j = r | Y_1, \cdots, Y_{j-1}, \mathbf{x}) = P(Y_j = r | Y_{j-1}, \mathbf{x})$.
- One such example for *binary responses* is

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

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

- It is called a **regressive logistic model** when

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

- Models of the type (20) may be embedded into the GLM framework, meaning some built-in functions in `R` (e.g. `glm()`) may be used to do the computing.
- In general, however, you need to write your own program to do the analysis.

§6.6.1 Conditional models: Asymmetric models (optional)4

Example: Reported happiness. Clogg (1982) investigated the association between gender (x), years in school (Y_1), and reported happiness (Y_2). The data are given in the following. Since x and Y_1 are prior to the statement about happiness, Y_2 is modelled conditionally on Y_1 and x.

Table 8: 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

§6.6.1 Conditional models: Symmetric models (optional)1

Symmetric models

- Suppose the response vector $\mathbf{Y} = (y_1, \cdots, y_m)$.
- If there is no natural ordering among $y_1, \cdots, y_m$, or if one does not want to use this ordering, models that treat response components in a symmetric way are more sensible. An example is *Visual Impairment Study* seen in §6.1.
- Now focus on the case where all $y_1, \cdots, y_m$ are **binary**. (General cases can be handled similarly.)
- **Symmetric conditional models** can be developed by specifying a conditional distribution to be used:

$$P(y_j = 1 | y_k, k \neq j; \mathbf{x}_j), \quad j = 1, \cdots, m \tag{21}$$

  For binary responses such conditional distributions uniquely determine the joint distribution.

§6.6.1 Example: Visual impairment study

Table 9: Visual impairment data, from Liang, Zeger & Qaqish (1992)

	White				Black				
Visual	Age								
impairment	40-50	51-60	61-70	70+	40-50	51-60	61-70	70+	Total
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

- Vector binary **response** variable (y_1, y_2), where $y_1 = 1$ if left-eye impaired, 0 otherwise; $y_2 = 1$ if right-eye impaired, 0 otherwise.
- **Covariates**: `Age` (yrs.), `Race` (W or B).
- **Aim**: find the effect of race and age on visual impairment.
- **Complication**: y_1 and y_2 are correlated.
- **Methods**: multivariate models for correlated responses; conditional models; asymmetric models, marginal models, GEE, etc..

§6.6.1 Conditional models: Symmetric models (optional)2

Symmetric models

- **Symmetric logistic models** are a natural choice for (21):

$$\pi = P(y_j = 1 | y_k, k \neq j; \mathbf{x}_j) = h(\alpha(w_j; \boldsymbol{\theta}) + \mathbf{x}_j^T \boldsymbol{\beta}_j), \quad j = 1, \cdots, m \tag{22}$$

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 $m = 2$,

$$\pi = P(y_j = 1 | y_k, k \neq j; \mathbf{x}_j) = h(\theta_0 + \theta_1 y_k + \mathbf{x}_j^T \boldsymbol{\beta}_j), \quad j, k = 1, 2. \tag{23}$$

- Full likelihood estimation procedure for (22) or (23) may be computationally cumbersome, because of the normalising constant involved in the joint pmf.

§6.6.1 Conditional models: Symmetric models (optional)3

- Instead, a quasi-likelihood approach (Conolly and Liang, 1988) may be used, with an *independent working* quasi-likelihood and quasi-score function for each cluster $i(= 1, \cdots, n)$:

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

  where $\mathbf{y}_i = (y_{i1}, \cdots, y_{ij}, \cdots, y_{im})^T$ are the responses in cluster i, $\pi_{ij}(\boldsymbol{\beta}, \boldsymbol{\theta}) = P(y_{ij} = 1|\cdot)$ is defined by (22), and $\sigma_{ij}^2 = \pi_{ij}(\boldsymbol{\beta}, \boldsymbol{\theta})(1 - \pi_{ij}(\boldsymbol{\beta}, \boldsymbol{\theta}))$.

§6.6.1 Conditional models: Symmetric models (optional)4

- Denoting $M_i = \operatorname{diag}\left\{\frac{\partial \pi_{i1}}{\partial(\boldsymbol{\beta}^T, \boldsymbol{\theta}^T)^T}, \cdots, \frac{\partial \pi_{im}}{\partial(\boldsymbol{\beta}^T, \boldsymbol{\theta}^T)^T}\right\}$, $\Sigma_i = \operatorname{diag}\{\sigma_{i1}^2, \cdots, \sigma_{im}^2\}$ and $\boldsymbol{\pi} = (\pi_{i1}, \cdots, \pi_{im})^T$, we can rewrite $\mathbf{S}_i(\boldsymbol{\beta}, \boldsymbol{\theta})$ in matrix form

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

  which is a multivariate extension of the quasi-score.
- 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 $\hat{V} = \sum_{i=1}^{n} \hat{S}_i \hat{S}_i^T$.

§6.6.1 Two drawbacks of the conditional models

1. They measure the effect of $\mathbf{x}$ on a binary component y_j conditional on incorporating into the effects of other y_k, $k \neq j$. Thus the model is not able to provide prediction based on $\mathbf{x}$ alone.
2. Interpretation of the effects depends on the dimension of $\mathbf{y}$.

Both drawbacks can be avoided by using a marginal modelling approach.

§6.6.2 Marginal models

- In many situations the primary interest is to analyse the **marginal mean** of the response given the covariates. The association between the responses is often of secondary interest. An example is the *visual impairment study*.
- Marginal models were first proposed by Liang & Zeger (1986) and Zeger & Liang (1986) in the context of longitudinal data with many short time series. Their modelling and estimation approach is based on GEE and has subsequently been modified and further developed.
- Focus in this subsection is on modelling and estimating marginal means of correlated and categorical response data by the first-order generalised estimating equations (GEE1).

§6.6.2 Marginal models for correlated univariate responses (1)

- In marginal models, the effects of covariates on responses and the association between responses is modelled separately.
- Let $\mathbf{y}_i = (y_{i1}, \cdots, y_{im_i})^T$ be the vector of responses, and

$$\mathbf{x}_i^T = (\mathbf{x}_{i1}^T, \cdots, \mathbf{x}_{im_i}^T)$$

  be the vector of covariates from cluster i, $i = 1, \cdots, n$. Here the cluster size m_i may vary with the cluster index i.
- Response observations within the same cluster are **correlated**. Response observations between different clusters are assumed **independent** of each other.

§6.6.2 Marginal models for correlated univariate responses (2)

Specification of a **marginal model** is as following

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

$$\mu_{ij}(\boldsymbol{\beta}) = E(y_{ij}|\mathbf{x}_{ij}) = h(\mathbf{z}_{ij}^T\boldsymbol{\beta}) \tag{24}$$

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

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

$$\sigma_{ij}^2 = \text{var}(y_{ij}|\mathbf{x}_{ij}) = v(\mu_{ij})\phi \tag{25}$$

   where $v(\cdot)$ is the **variance function** of known form.

3. The correlation between y_{ij} and y_{ik} is a function of $\mu_{ij} = \mu_{ij}(\boldsymbol{\beta})$, $\mu_{ik} = \mu_{ik}(\boldsymbol{\beta})$ and perhaps of additional association parameters $\boldsymbol{\alpha}$:

$$\text{corr}(y_{ij}, y_{ik}) = c(\mu_{ij}, \mu_{ik}, \boldsymbol{\alpha}) \tag{26}$$

   with the function $c(\cdot, \cdot, \cdot)$ being of known form.

§6.6.2 Marginal models for correlated univariate responses (3)

- **Remarks:**
  1. Parameters $(\boldsymbol{\beta}, \boldsymbol{\alpha})$ are the same for each cluster. Hence marginal models are appropriate for analysing **population-averaged** effects.
  2. Marginal effects $\boldsymbol{\beta}$ can be consistently estimated even if the correlation function is incorrectly specified. However, the efficiency of the estimator $\hat{\boldsymbol{\beta}}$ can be compromised.
- Since the primary scientific objective is often the regression relationship, it is important to **correctly specify** the marginal mean structure (24), while $c(\mu_{ij}, \mu_{ik}, \boldsymbol{\alpha})$ only needs to be a **working correlation** for the association between y_{ij} and y_{ik}. Together with (25) a **working covariance matrix** $\mathrm{cov}(\mathbf{y}_i) = \Sigma_i(\boldsymbol{\beta}, \boldsymbol{\alpha})$ can be obtained, where an additional dispersion parameter ϕ is involved but suppressed for presentation simplicity.
- Two main approaches for specifying the working correlations or covariances have been considered in the literature.

§6.6.2 Marginal models for correlated univariate responses (4)

First approach: The variance structure (25) is supplemented by a **working correlation matrix** $R_i(\boldsymbol{\alpha})$, so that the **working covariance matrix** is of the form

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

where $C_i(\boldsymbol{\beta}) = \text{diag}\,[\text{var}(y_{ij}|x_{ij})] = \text{diag}\{\sigma_{i1}^2,\cdots,\sigma_{im_i}^2\}$. 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. $R_i(\boldsymbol{\alpha})$ completely unspecified, except being positive definite, i.e. $\alpha_{jk} = \text{corr}(y_{ij}, y_{ik})$, $j < k$.

§6.6.2 Marginal models for correlated univariate responses (5)

Second approach: For binary responses the odds ratio is an alternative measure of association that is easier to interpret and has some desirable properties.

- The odds ratio for y_{ij}, y_{ik}, $j, k = 1, \cdots, m$, $j \neq k$, is defined as

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

- From this definition, the probability $P(y_{ij} = y_{ik} = 1)$, denoted as π_{i11}, can be shown to satisfy

$$\pi_{i11} = E(y_{ij}y_{ik}) = \begin{cases} \frac{1-(\pi_{ij}+\pi_{ik})(1-\gamma_{ijk})-s(\pi_{ij},\pi_{ik},\gamma_{ijk})}{2(\gamma_{ijk}-1)} & \text{if } \gamma_{ijk} \neq 1 \\ \pi_{ij}\pi_{ik} & \text{if } \gamma_{ijk} = 1 \end{cases} \tag{27}$$

with

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

- It can be seen that the covariance matrix of $\mathbf{y}_i = (y_{i1} \cdots, y_{im_i})^T$ can be expressed as a function of $(\pi_{ij}, \pi_{ik}, \gamma_{ijk})$.

§6.6.2 Marginal models for correlated univariate responses (6)

- One advantage of the odds ratio is that it is unconstrained, being able to take any positive values. In contrast,

$$\mathrm{corr}(y_{ij}, y_{ik}) = \frac{\pi_{i11} - \pi_{ij}\pi_{ik}}{\sqrt{\pi_{ij}(1-\pi_{ij})\pi_{ik}(1-\pi_{ik})}},$$

  the correlation between y_{ij} and y_{ik} is constrained (i.e. its possible values may all fall into a subset of $[-1, 1]$), because the involved π_{i11} is constrained by

$$\max(0, \pi_{ij} + \pi_{ik} - 1) \leq \pi_{i11} \leq \min(\pi_{ij}, \pi_{ik}).$$

  Note
  $\pi_{i11} = P(y_{ij} = 1) + P(y_{ik} = 1) - P(y_{ij} = 1 \text{ or } y_{ik} = 1) \geq \pi_{ij} + \pi_{ik} - 1.$

§6.6.2 Marginal models for correlated univariate responses (7)

- To reduce the number of unknown parameters involved in the odds ratio, we assume $\gamma_{ijk} = \gamma_{ijk}(\boldsymbol{\alpha})$, a function of a vector parameter $\boldsymbol{\alpha}$.
- Common choices of $\gamma_{ijk}(\boldsymbol{\alpha})$:
  1. $\gamma_{ijk} = \gamma$, for all i, j, k.
  2. $\log \gamma_{ijk} = \boldsymbol{\alpha}^T \omega_{ijk}$.

  Using the relation (27), $\text{cov}(\mathbf{y}_i) = \Sigma_i(\boldsymbol{\beta}, \boldsymbol{\alpha})$ is also a function of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$.

§6.6.2 Marginal models for correlated univariate responses (8)

Some examples of marginal models:

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:

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

§6.6.2 GEE approach in statistical inference (1)

For cluster $i = 1, \cdots, n$, let $\mathbf{y}_i = (y_{i1}, \cdots, y_{im_i})^T$, $\mathbf{x}_i = (\mathbf{x}_{i1}^T, \cdots, \mathbf{x}_{im_i}^T)^T$, $\boldsymbol{\mu}_i(\boldsymbol{\beta}) = (\mu_{i1}(\boldsymbol{\beta}_1), \cdots, \mu_{im_i}(\boldsymbol{\beta}_{m_i}))^T$ and $\Sigma_i(\boldsymbol{\beta}, \boldsymbol{\alpha})$ be defined as before.

- In linear models for Gaussian responses, there is no problem to perform maximum likelihood estimation, since specification of means and covariances determines the likelihood function.
- This is not the case with non-Gaussian data. Therefore a generalised estimating approach is proposed.
- Keeping the association parameter $\boldsymbol{\alpha}$ and ϕ, if present, fixed for the moment, the **generalised estimating equation** (GEE) for effect $\boldsymbol{\beta}$ is

$$\mathbf{S}_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{n} Z_i^T D_i(\boldsymbol{\beta}) \Sigma_i^{-1}(\boldsymbol{\beta}, \boldsymbol{\alpha})(\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})) = \mathbf{0}, \qquad (28)$$

  with m_i-row design matrix $Z_i = (\mathbf{z}_{i1}, \cdots, \mathbf{z}_{im_i})^T$ 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 $\eta_{ij} = \mathbf{z}_{ij}^T \boldsymbol{\beta}$.
- The middle-part of (28) is a multivariate quasi-score function.

§6.6.2 GEE approach in statistical inference (2)

A GEE estimate $\hat{\boldsymbol{\beta}}$ is computed by iterating a modified Fisher scoring algorithm w.r.t. $\boldsymbol{\beta}$ and estimation of $\boldsymbol{\alpha}$ (and ϕ):

(i) Given current estimates $\hat{\boldsymbol{\alpha}}$ (and $\hat{\phi}$), the GEE (28) for $\hat{\boldsymbol{\beta}}$ is solved by the iterations

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

with

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

being the observed quasi-information matrix.

§6.6.2 GEE approach in statistical inference (3)

(ii) If $\boldsymbol{\alpha}$ is unknown, it can be consistently estimated by a method of moments based on Pearson residuals $\hat{r}_{ij} = \dfrac{y_{ij} - \hat{\mu}_{ij}}{\sqrt{v(\hat{\mu}_{ij})}}$.

- The dispersion parameter ϕ is estimated consistently by $\hat{\phi} = \dfrac{1}{N-p} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \hat{r}_{ij}^2$, with $N = \sum_{i=1}^n m_i$ and $p = \dim(\boldsymbol{\beta})$.
- Estimation of $\boldsymbol{\alpha}$ depends on the choice of $R_i(\boldsymbol{\alpha})$. For exchangeable 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

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

if all cluster sizes $m_i = m$ and $m << n$.

§6.6.2 GEE approach in statistical inference (4)

- Cycling between Fisher scoring steps for $\boldsymbol{\beta}$ and consistent estimation of $\boldsymbol{\alpha}$ and ϕ leads to a consistent estimation of $\boldsymbol{\beta}$.
- Alternatively, $\boldsymbol{\alpha}$ (and possible ϕ) can be estimated by simultaneously solving an additional estimating equation. Details are not pursued here.
- Under regularity conditions, the GEE estimator $\hat{\boldsymbol{\beta}}$ satisfies

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

  with $F = \sum_{i=1}^{n} Z_i^T D_i \Sigma_i^{-1} D_i Z_i$ and $V = \sum_{i=1}^{n} Z_i^T D_i \Sigma_i^{-1} S_i \Sigma_i^{-1} D_i Z_i$, where S_i is the *true* covariance matrix of $\mathbf{y}_i$.
- cov($\hat{\boldsymbol{\beta}}$) is approximated by the "sandwich matrix":

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

- $\text{cov}(\mathbf{S}_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\alpha})) = V$ and $\mathbf{S}_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\alpha}) \overset{a}{\sim} N(0, V)$.

§6.6.2 Visual impairment example (1)

We use GEE approach to fit a marginal model to the visual impairment data. R packages gee with command gee() and geepack with command geeglm() can be used. (There are a few problems with both packages).

```
library(gee); library(geepack)
VI.dat <- read.csv("D:/MAST90139/Visual-impairment.csv"); VI.dat
```

```
   ID Yes  No    Age  Race
1   1  15 617 40to50 White
2   1  19 613 40to50 White
3   2  24 557 51to60 White
4   2  25 556 51to60 White
5   3  42 789 61to70 White
6   3  48 783 61to70 White
7   4 139 673    70+ White
8   4 146 666    70+ White
9   5  29 750 40to50 Black
10  5  31 748 40to50 Black
11  6  38 574 51to60 Black
12  6  37 575 51to60 Black
13  7  50 473 61to70 Black
14  7  49 474 61to70 Black
15  8  85 344    70+ Black
16  8  93 336    70+ Black
```

§6.6.2 Visual impairment example (2)

```
VI1e<- gee(cbind(Yes, No) ~ Age + Race, id = ID,
           data = VI.dat, family = binomial, corstr = "exchangeable")
summary(VI1e)
```

```
 GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
 gee S-function, version 4.13 modified 98/01/27 (1998)

Model:
 Link:                      Logit
 Variance to Mean Relation: Binomial
 Correlation Structure:     Exchangeable

Call:
gee(formula = cbind(Yes, No) ~ Age + Race, id = ID, data = VI.dat,
    family = binomial, corstr = "exchangeable")

Summary of Residuals:
   Min     1Q Median     3Q    Max
  15.0   28.0   39.9   58.6  145.8


Coefficients:
            Estimate Naive S.E. Naive z Robust S.E. Robust z
(Intercept)   -3.225     0.1125  -28.66      0.0254  -127.22
Age51to60      0.478     0.1451    3.30      0.0167    28.65
```

§6.6.2 Visual impairment example (3)

```
Call:
gee(formula = cbind(Yes, No) ~ Age + Race, id = ID, data = VI.dat,
    family = binomial, corstr = "exchangeable")

Summary of Residuals:  ###Residuals calculation is not correct in gee().
   Min      1Q Median     3Q    Max
  15.0   28.0   39.9   58.6  145.8

Coefficients:
            Estimate Naive S.E. Naive z Robust S.E. Robust z
(Intercept)   -3.225     0.1125  -28.66      0.0254  -127.22
Age51to60      0.478     0.1451    3.30      0.0167    28.65
Age61to70      0.837     0.1348    6.21      0.0905     9.26
Age70+         1.973     0.1227   16.08      0.0531    37.13
RaceWhite     -0.350     0.0768   -4.56      0.0662    -5.28

Estimated Scale Parameter:  0.571
Number of Iterations:  1

Working Correlation
      [,1]  [,2]
[1,] 1.000 0.886
[2,] 0.886 1.000
```

§6.6.2 Visual impairment example (4)

```
total=VI.dat$Yes+VI.dat$No

pearson.res1e=
    (VI.dat$Yes-total*VI1e$fitted)/sqrt(total*(VI1e$fitted)*(1-VI1e$fitted))

sum(pearson.res1e^2)/11
[1] 0.5708722 #estimate of \phi based on Pearson residuals

> VI1e$scale
[1] 0.5708722

(VI1e$scale*3)^(-1)*sum(pearson.res1e[2*(1:8)-1]*pearson.res1e[2*(1:8)])
[1] 0.8861748

> VI1e$working.correlation
          [,1]      [,2]
[1,] 1.0000000 0.8861748
[2,] 0.8861748 1.0000000
```

§6.6.2 Visual impairment example (5)

```
VI3 <- geeglm(cbind(Yes, No)~Age + Race, family=binomial(link="logit"),
                     data=VI.dat, id=ID, corstr = "exchangeable", std.err="san.se")
summary(VI3)
```

```
Call:
geeglm(formula = cbind(Yes, No) ~ Age + Race, family = binomial(link = "logit"),
    data = VI.dat, id = ID, corstr = "exchangeable", std.err = "san.se")
Coefficients:
            Estimate Std.err    Wald Pr(>|W|)
(Intercept)  -3.2253  0.0254 16184.7  < 2e-16 ***
Age51to60     0.4783  0.0167   820.7  < 2e-16 ***
Age61to70     0.8374  0.0905    85.7  < 2e-16 ***
Age70+        1.9728  0.0531  1378.3  < 2e-16 ***
RaceWhite    -0.3498  0.0662    27.9  1.3e-07 ***
---
Estimated Scale Parameters: #IScale seems defined differently in gee() and geeglm()
            Estimate  Std.err
(Intercept) 0.000604 0.000323

Correlation: Structure = exchangeable  Link = identity
Estimated Correlation Parameters:
      Estimate Std.err     #The estimation here is about half of that in gee().
alpha    0.441   0.274
Number of clusters:   8   Maximum cluster size: 2
```

§6.6.2 Visual impairment example (6)

```
pearson.res3=(VI.dat$Yes/total-VI3$fitted)/sqrt((VI3$fitted)*(1-VI3$fitted)/total)
phi.est=sum(pearson.res3^2)/11   #=0.5709

sum((summary(VI3)$deviance.resid)^2)/11

[1] 0.577  #very close to 0.5709.
               #Using deviance residuals or Pearson residuals give similar results.

VI3$geese$gamma     #How is gamma defined?

(Intercept)
   0.000604

(phi.est*3)^(-1)*sum(pearson.res3[2*(1:8)-1]*pearson.res3[2*(1:8)])

[1] 0.8862  #estimated alpha

 (0.577*3)^(-1)*sum(summary(VI3)$devi[2*(1:8)-1]*summary(VI3)$devi[2*(1:8)])

[1] 0.872

VI3$geese$alpha

 alpha
0.4415 #approximately half of 0.8862. Then How is alpha defined here?
```

§6.6.2 Marginal models for correlated responses having k categories (1)

- 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 same k categories and is coded by a dummy vector

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

- 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 to the responses with each other with in each cluster.

§6.6.2 Marginal models for correlated responses having *k* categories (2)

Here we focus on ordinal responses. Other types of categorical responses can be handled similarly.

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

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

with $\pi_{ijr} = P(Y_{ij} \leq r|\mathbf{x}_{ij}) = \sum_{\ell=1}^{r} P(y_{ij\ell} = 1|\mathbf{x}_{ij})$, and the response function $\mathbf{h}(\cdot)$ and design matrix Z_{ij}.

(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.

§6.6.2 Marginal models for correlated responses having k categories (3)

(iii) Association between Y_{ij} and Y_{ik} is modelled by a working correlation matrix R_i or by global cross-ratios. For example, 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 $Q_{q\times q}$ contains $\boldsymbol{\alpha}$ to be estimated by a method of moments. **Global cross-ratios** (GCR), for each pair of categories ℓ and m of Y_{ij} and Y_{ik}, are defined as

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

GCR can be modelled log-linearly, i.e. $\log(\gamma_{ijk}(\ell, m)) = \alpha_{\ell m}$ or by a regression model including covariate effects.

- For the model specified by (i), (ii) and (iii), the involved regression and association parameters can be estimated by a multivariate extension of the GEE approach discussed before. Details not pursued here.
- R packages `multgee`, `geepack` and `repolr` may be used to fit the above model.

§6.6.2 Likelihood-based inference for marginal models

- The GEE approach is not likelihood-based as it does not require specification of the joint distribution of multivariate response vector $\mathbf{y}_i$, $i = 1, \cdots, n$.
- Difficulty with the likelihood-based inference is due to the difficulty in formulating this joint distribution.
- So in general it is difficult to perform likelihood-based inference for marginal models.
- But there are new developments:
  - Constructing joint multivariate distribution by copulas
  - Vector GLM
  - Bayesian inference.

§6.6.3 Marginal models for longitudinal data (1)

- In many longitudinal studies, data consist of a small or moderate number of repeated observations for many subjects; and the main objective is to analyse the effects of covariates on a response variable, without conditioning on previous responses.
- In this setting it is more natural to view the data as a cross section of individual time series $\mathbf{y}_i = (\mathbf{y}_{i1}^T, \cdots, \mathbf{y}_{iT_i}^T)^T$, $\mathbf{x}_i = (\mathbf{x}_{i1}^T, \cdots, \mathbf{x}_{iT_i}^T)^T$, $i = 1, \cdots, n$. It is also often assume that, given the covariates, individual time series are mutually independent.
- Marginal models introduced in §6.5.2 can be used for these cases.
- Suppose $\mathbf{y}_{it} = (y_{it1}, \cdots, y_{itq})^T$ are $q = k - 1$ dummy variables describing a multi-categorical response Y_{it} with k categories.
  For simplicity, assume the number of categories is the same for all Y_{it} responses.

§6.6.3 Marginal models for longitudinal data (2)

Marginal GEE models for longitudinal data are then defined as

(i) The marginal mean is correctly specified by a response model

$$\boldsymbol{\mu}_{it}(\boldsymbol{\beta}) = E(\mathbf{y}_{it}|\mathbf{x}_{it}) = \mathbf{h}(\boldsymbol{\eta}_{it}), \quad \boldsymbol{\eta}_{it} = Z_{it}\boldsymbol{\beta}$$

where $\boldsymbol{\eta}_{it}$ is q-dimensional with components $\eta_{itr} = \mathbf{z}_{itr}^T\boldsymbol{\beta}$, $\mathbf{z}_{itr}^T$ is the rth row of Z_{it}.

(ii) The association structure for $\mathbf{y}_i = (\mathbf{y}_{i1}^T, \cdots, \mathbf{y}_{iT_i}^T)^T$ is specified by a "working" covariance matrix $\Sigma_i(\boldsymbol{\beta}, \boldsymbol{\alpha})$, depending on regression parameters $\boldsymbol{\beta}$, association parameters $\boldsymbol{\alpha}$, and possibly an additional nuisance parameter ϕ.

§6.6.3 Marginal models for longitudinal data (3)

- $\Sigma_i(\boldsymbol{\beta}, \boldsymbol{\alpha})$ can be determined by a variance function for $\mathbf{y}_{it}$ together with a "working" correlation matrix $R_i(\boldsymbol{\alpha})$ or some odds ratio model $\boldsymbol{\gamma}(\boldsymbol{\alpha})$.
- Common choices for $R_i(\boldsymbol{\alpha})$:
  1. stationary correlations between $\mathbf{y}_{is}$ and $\mathbf{y}_{it}$

  $$(R_i(\boldsymbol{\alpha}))_{st} = \alpha(|t-s|) \tag{29}$$

  2. its special form $\alpha(|t-s|) = \alpha^{|t-s|}$. AR(1) correlation.

Statistical inference in marginal GEE models (1)

- Estimation of $\boldsymbol{\beta}$ using GEEs is carried out in complete analogy to §6.5.2.
- Denote $E(\mathbf{y}_i|\mathbf{x}_i) = \boldsymbol{\mu}_i(\boldsymbol{\beta}) = (\boldsymbol{\mu}_{i1}^T, \cdots, \boldsymbol{\mu}_{iT_i}^T)^T$, $Z_i^T = (Z_{i1}, \cdots, Z_{iT_i})$, and (block-) diagnol matrices $D_i(\boldsymbol{\beta}) = \text{diag}\{D_{it}(\boldsymbol{\beta})\}$, with $D_{it}(\boldsymbol{\beta}) = \dfrac{\partial \mathbf{h}^T}{\partial \boldsymbol{\eta}_{it}} |_{\boldsymbol{\eta}_{it}} = Z_{it}\boldsymbol{\beta}$.
- The GEE for $\boldsymbol{\beta}$ is

$$\mathbf{s}_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{n} Z_i^T D_i(\boldsymbol{\beta}) \Sigma_i^{-1}(\boldsymbol{\beta}, \boldsymbol{\alpha})(\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})) = \mathbf{0} \qquad (30)$$

- Given current estimates $\hat{\boldsymbol{\alpha}}$ and $\hat{\phi}$, the GEE (30) for $\hat{\boldsymbol{\beta}}$ is solved by

$$\hat{\boldsymbol{\beta}}^{(k+1)} = \hat{\boldsymbol{\beta}}^{(k)} + \left(\hat{F}^{(k)}\right)^{-1} \hat{\mathbf{s}}^{(k)}, \quad k = 1, 2, \cdots$$

with the "working" Fisher information matrix $\hat{F}^{(k)} = \sum_{i=1}^{n} Z_i^T D_i(\hat{\boldsymbol{\beta}}^{(k)}) \Sigma_i^{-1}(\hat{\boldsymbol{\beta}}^{(k)}, \hat{\boldsymbol{\alpha}}) D_i^T(\hat{\boldsymbol{\beta}}^{(k)}) Z_i$ and $\hat{\mathbf{s}}^{(k)} = \mathbf{s}_{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}^{(k)}, \hat{\boldsymbol{\alpha}})$.

Statistical inference in marginal GEE models (2)

- Given a current estimate of $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$ and ϕ can be estimated from Pearson residuals
$$\hat{r}_{it} = \frac{y_{it} - \hat{\mu}_{it}}{(v(\hat{\mu}_{it}))^{1/2}}$$
by the method of moments.

1. $\hat{\phi} = \frac{1}{N-p} \sum_{i=1}^{n} \sum_{t=1}^{T_i} \hat{r}_{it}^2, \quad N = \sum_{i=1}^{n} T_i$
2. Estimation of $\boldsymbol{\alpha}$ depends on the choice of $R_i(\boldsymbol{\alpha})$. For the exchangeable correlation model
$$\hat{\alpha} = \left\{ \hat{\phi} \left[\sum_{i=1}^{n} \frac{1}{2} T_i(T_i - 1) - p \right] \right\}^{-1} \sum_{i=1}^{n} \sum_{t>s} \hat{r}_{it} \hat{r}_{is}$$
3. A totally unspecified $R = R(\boldsymbol{\alpha})$ can be estimated if $T << n$.
4. In particular for binary and categorical observations, estimation of $\boldsymbol{\alpha}$ via GEE2 is often preferrable.

Statistical inference in marginal GEE models (3)

- Consistency and asymptotic normality results for $\hat{\boldsymbol{\beta}}$ can be obtained along the lines of asymptotic theory of quasi-MLEs for independent observations if T_i, $i = 1, \cdots, n$ are fixed and $n \to \infty$.
- If $\hat{\boldsymbol{\alpha}}$ is $\sqrt{n}$-consistent given $\boldsymbol{\beta}$ and ϕ; and $\hat{\phi}$ is $\sqrt{n}$-consistent given $\boldsymbol{\beta}$, then $\hat{\boldsymbol{\beta}}$ is $\sqrt{n}$-consistent and asymptotically normal, i.e.

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

with $F = \sum_{i=1}^{n} Z_i^T D_i \Sigma_i^{-1} D_i^T Z_i$, $V = \sum_{i=1}^{n} Z_i^T D_i \Sigma_i^{-1} \mathrm{Cov}(\mathbf{y}_i) \Sigma_i^{-1} D_i^T Z_i$.

- Asymptotic covariance matrix $A = F^{-1}VF^{-1}$ can be consistently estimated by replacing $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$ and ϕ by their respective consistent estimates and $\mathrm{Cov}(\mathbf{y}_i)$ by $(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_i)(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_i)^T$, i.e. by sandwich matrix

$$\hat{A} = \widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}) \overset{a}{\approx} \hat{F}^{-1} \left\{ \sum_{i=1}^{n} Z_i^T \hat{D}_i \hat{\Sigma}_i^{-1} (\mathbf{y}_i - \hat{\boldsymbol{\mu}}_i)(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_i)^T \hat{\Sigma}_i^{-1} \hat{D}_i^T Z_i \right\} \hat{F}^{-1} \quad (31)$$

- Based on the preceeding asymptotic result, confidence intervals for $\boldsymbol{\beta}$ and Wald and score tests may be constucted; cf. §2.3.1.

Example 6.4 Ohio children (1)

- The `ohio` data concern 537 children from Steubenville, Ohio and were taken as part of a study on the effects of air pollution.
- Children were in the study for four years from age seven to ten.
- The response is whether they wheezed or not.
- The variables are

resp: an indicator of wheeze status (1=yes, 0=no)

id: an identifier for the child, taking values from 0 to 536

age: 7 yrs $= -2$; 8 yrs $= -1$; 9 yrs $=0$; 10 yrs $=1$

smoke: mother's smoking status at start of the study (1=smoker, 0=nonsmoker)

Example 6.4 Ohio children (2)

Some analysis has been done to the data in R, producing the following output.

```
> head(ohio)

  resp id age smoke
1    0  0  -2     0
2    0  0  -1     0
3    0  0   0     0
4    0  0   1     0
5    0  1  -2     0
6    0  1  -1     0
```

```
> tail(ohio)

     resp  id age smoke
2143    1 535   0     1
2144    1 535   1     1
2145    1 536  -2     1
2146    1 536  -1     1
2147    1 536   0     1
2148    1 536   1     1
```

Example 6.4 Ohio children (3)

```
> str(ohio)

'data.frame':   2148 obs. of  4 variables:
 $ resp : int  0 0 0 0 0 0 0 0 0 0 ...
 $ id   : int  0 0 0 0 1 1 1 1 2 2 ...
 $ age  : int  -2 -1 0 1 -2 -1 0 1 -2 -1 ...
 $ smoke: int  0 0 0 0 0 0 0 0 0 0 ...

> library(geepack}

> fit.exch <- geeglm(resp~age+smoke, family=binomial(link="logit"),
data=ohio, id=id, corstr = "exchangeable", std.err="san.se")

> summary(fit.exch)
```

Example 6.4 Ohio children (4)

```
> summary(fit.exch)

Call:
geeglm(formula = resp ~ age + smoke, family = binomial(link = "logit"),
    data = ohio, id = id, corstr = "exchangeable", std.err = "san.se")

 Coefficients:
            Estimate  Std.err    Wald Pr(>|W|)
(Intercept) -1.880    0.114   272.597  < 2e-16 ***
age         -0.113    0.044     6.684  0.00973 **
smoke        0.265    0.178     2.224  0.13588
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Estimated Scale Parameters:
            Estimate Std.err
(Intercept)   0.9985  0.1116

Correlation: Structure = exchangeable  Link = identity

Estimated Correlation Parameters:
      Estimate Std.err
alpha   0.3543 0.06244
Number of clusters:   537   Maximum cluster size: 4
```

Example 6.4 Ohio children (5)

Use the above output to answer the following questions.

1. Let y_{it} be the response value `resp` of child i at age t. Write down the model involved in the analysis, including the mean, variance and correlation coefficient of y_{it}'s. Give the estimates of the parameters appearing in the model.
2. Write down the model's design matrix for data where `id`=536.
3. Estimate the odds ratio of wheezing for a child for every one year older in age. Also calculate an approximate standard error of your odds ratio estimate.
4. Estimate the odds ratio of wheezing for a 10-year old child with a smoking mother versus a 9-year old child with nonsmoking mother.

Example 6.4 Ohio children (6)

1 Let y_{it} be the response value `resp` of child i at age t. Write down the model involved in the analysis, including the mean, variance and correlation coefficient of y_{it}'s. Give the estimates of the parameters appearing in the model.

- Let $\pi_{it} = P(y_{it} = 1|\text{age, smoke})$, $t = 1, \cdots, 4$; $i = 1, \cdots, 537$. Then the model is specified by the following

$$
\begin{aligned}
\log \frac{\hat{\pi}_{it}}{1 - \hat{\pi}_{it}} &= \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{age} + \hat{\beta}_2 \cdot \text{smoke} \\
&= -1.880 - 0.113 \times \text{age} + 0.265 \times \text{smoke} \\
\widehat{\text{Var}}(y_{it}) &= \hat{\phi}\hat{\pi}_{it}(1 - \hat{\pi}_{it}) = \frac{0.9985 e^{-1.880 - 0.113 \cdot \text{age} + 0.265 \cdot \text{smoke}}}{\left(1 + e^{-1.880 - 0.113 \cdot \text{age} + 0.265 \cdot \text{smoke}}\right)^2} \\
\widehat{\text{Corr}}(y_{it}, y_{is}) &= \hat{\alpha} = 0.3543 \quad \text{for } t \neq s
\end{aligned}
$$

Example 6.4 Ohio children (7)

2 Write down the model's design matrix for data where `id`=536.

- $Z(\text{id} = 536) = \begin{bmatrix} 1 & -2 & 1 \\ 1 & -1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix}$

Example 6.4 Ohio children (8)

3 Estimate the odds ratio of wheezing for a child for every one year older in age. Also calculate an approximate standard error of your odds ratio estimate.

- $\widehat{\mathrm{OR}} = e^{\hat{\beta}_1} = e^{-0.113} = 0.893$
- $\text{s.e.}(\widehat{\mathrm{OR}}) \approx e^{\hat{\beta}_1} \cdot \text{s.e.}(\hat{\beta}_1) = e^{-0.113} \times 0.044 = 0.039.$

Example 6.4 Ohio children (9)

4 Estimate the odds ratio of wheezing for a 10-year old child with a smoking mother versus a 9-year old child with nonsmoking mother.

- For a 10-year old child with a smoking mother

$$\log \frac{\hat{\pi}_{it}}{1-\hat{\pi}_{it}} = \hat{\beta}_0 + \hat{\beta}_1 \cdot 1 + \hat{\beta}_2 \cdot 1 = -1.880 - 0.113 + 0.265$$

  For a 9-year old child with a non-smoking mother

$$\log \frac{\hat{\pi}_{it}}{1-\hat{\pi}_{it}} = \hat{\beta}_0 + \hat{\beta}_1 \cdot 0 + \hat{\beta}_2 \cdot 0 = -1.880$$

- Therefore

$$\widehat{\text{OR}} = \frac{e^{\hat{\beta}_0+\hat{\beta}_1+\hat{\beta}_2}}{e^{\hat{\beta}_0}} = e^{\hat{\beta}_1+\hat{\beta}_2} = e^{-0.113+0.265} = e^{0.152} = 1.164.$$