7
Likelihood-based Marginal Models

In Section 5.3.1, a general overview of marginal models is presented. Specific versions, largely focused on contingency tables, were presented in Chapter 6. In this chapter, we contemplate the fully general situation. We focus on fully specified probabilistic models, in contrast to specifying a few low-order moments only, such as in generalized estimating equations (GEE). Although undoubtedly complicating both the theory and the computations, there are at least two situations in which this route is the preferred one. First, the scientific question may require careful modeling of the association structure, in addition to the univariate response function. Second, one may be interested in the joint probability of a number of events (e.g., what is the probability of side effects occurring at two subsequent measurement occasions). In such a case, the association structure is not of direct interest, but is still indirectly needed to calculate such joint, or union, probabilities. An additional reason is that, such models as the Bahadur model (Section 7.2) or the global odds ratio model (the Dale model, Section 7.7) are the underlying basis for non-likelihood methods discussed later. For example, standard GEE, such as introduced by Liang and Zeger (1986) and studied in Chapter 8, is based on Bahadur's probabilistic model, while the version proposed by Lipsitz, Laird, and Harrington (1991) can be seen as rooted in the Dale model.

We begin by presenting the Bahadur model (Section 7.2). It has a relatively simple genesis, but at the same time suffer from severe drawbacks. Section 7.3 presents a general framework, encompassing a wide class of marginal models, while details on maximum likelihood estimation are given in Section 7.4. The ideas developed in Sections 7.3 and 7.4 are exemplified, us-

ing an influenza study, in Section 7.5. Two specific families, the multivariate probit model (Section 7.6) and the multivariate Dale model, or global odds ratio model (Section 7.7) are presented next. Section 7.8 presents a hybrid model, combining marginal and conditional model specifications. Three case studies, a cross-over trial in primary dysmenorrhoea (Section 7.9), the multivariate POPS study (Section 7.10, introduced in Section 2.6), and the longitudinal fluvoxamine trial (Section 7.11) are presented.

7.1 Notation

In Chapter 4, we indicated, for each individual, subject, or experimental unit $i = 1, \ldots, N$ in a study, a series of measurements by $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{n_i})'$, along with covariate information, usually grouped into a matrix X_i. We will refer to this convention as the *regression notation*.

When data are non-Gaussian in nature, this notation can sometimes be used without too much modification, such as in the later chapters in this part (e.g., on generalized estimating equations in Chapter 8), or in Part IV on subject-specific models. On the other hand, note that in Chapter 6 we merely needed indices to indicate cells in a contingency table. For example, (i, j) in a two-way contingency table refers to row i and column j. In each such cell, *a number of* subjects are grouped. When a two-way contingency table is further split over levels of, say, a dichotomous covariate, such as in Section 6.6.1, one often merely adds an additional index. This is similar to the conventions in analysis of variance and in contrast to linear regression.

In the present chapter, we need a hybrid system. On the one hand, the focus is on (longer) sequences of measurements, together with sets of covariates that can be continuous, categorical, or a mixture of these. In later chapters, it will be sufficient to use the regression-type notation, sketched at the start of this section. However, here we will need to describe not just marginal, univariate regressions, also the association structure needs to be modeled. This brings us close to a contingency table setting. When we have, for example, one covariate with two levels and five repeated binary measures, we can view the data as consisting of two 2^5 contingency tables. But the same view can be adopted when we have covariates with more levels, and even when some or all of the covariates are continuous. For continuous covariates, measured with high accuracy, there may be one or at most a few study subjects corresponding to it. Rather than being a problem, it is merely a way of conveniently framing both genuine contingency table settings and categorical data regression settings into a single, contingency table notational convention.

Thus, in this chapter, we will let $r = 1, \ldots, N$ indicate the covariate or design levels, each containing N_r subjects. For example, when there is one covariate with two levels, $N = 2$ and the total sample size is $N_1 + N_2$.

When the covariate is continuous and such that there is only one subject per covariate level, then each $N_r = 1$ and the total sample size is N. The consequence of our choice is that, for the time being, we need an additional index i for subjects within design lvels.

The outcome for subject i at the rth level (group) is a series of measurements Y_{rij} $(j = 1, \ldots, n_r)$. In case there are subjects sharing covariates, but with a different number of repeated measurements taken, then these should be split over several design levels, implying that r defines unique combinations of covariate levels and numbers of repeated measurements. An additional notational element is that our outcome Y_{rij} can be binary (usually taking the values 0 and 1), but also categorical, ordered, or unordered. We then need additional notation and assume that in such case variable Y_{rij} can take on c_j distinct (possibly ordered) values. Without loss of generality, denote them by $1, \ldots, c_j$. In examples of a multivariate nature, the measurement sequence usually is equally long for all subjects, i.e., $n_r \equiv n$ but the number c_j of categories per outcome can be variable. In longitudinal settings, the number of measurements could also be different from subject to subject, but when the same outcome is measured repeatedly over time, one typically sees that $c_j \equiv c$. The more elaborate notation will be referred to as the *contingency table notation*.

In the specific case of categorical data with more than two, possibly ordered, categories, it is useful to make use of some additional notation. All information about the responses on the units in the rth group is contained in a cross-classification of the outcomes Y_{rij} into a $c_1 \times \ldots \times c_{n_r}$ dimensional contingency table with cell counts

$$Z_r^*(\boldsymbol{k}) \equiv Z_r^*(k_1, \ldots, k_{n_r}), \tag{7.1}$$

where cell $\boldsymbol{k} = (k_1, \ldots, k_{n_r})$ corresponds to the subjects with $Y_{rij} = k_j$, for $j = 1, \ldots, n_r$.

Along with the outcomes, a vector of explanatory variables x_{rj} is recorded. The covariate vector is allowed to change over time. It can include continuous and discrete variables. Available covariate information, along with other relevant design features, are incorporated in a design matrix X_r.

In harmony with the possibility to use cumulative measures for ordinal data, we construct the table of cumulative counts:

$$Z_r(\boldsymbol{k}) = \sum_{\boldsymbol{\ell} \leq \boldsymbol{k}} Z_r^*(\boldsymbol{\ell}). \tag{7.2}$$

Thus, $Z_r(\boldsymbol{k})$, where $\boldsymbol{k} = (k_1, \ldots, k_{n_r})$, is just the number of individuals in group r whose observed response vector is $\boldsymbol{k}$, and likewise for $Z_r(\boldsymbol{k})^*$. The corresponding probabilities are

$$\mu_r^*(\boldsymbol{k}) = P(\boldsymbol{Y}_{ri} = \boldsymbol{k} | X_r, \boldsymbol{\beta}) \tag{7.3}$$

and $\mu_r(\boldsymbol{k}) = P(\boldsymbol{Y}_{ri} \leq \boldsymbol{k}|X_r, \boldsymbol{\beta})$. Let $\boldsymbol{Z}_r$ be the vector of all cumulative cell counts with $\boldsymbol{\mu}_r$ the corresponding vector of probabilities. Note that $Z_r(c_1, \ldots, c_{n_r}) = n_r$ and $\mu_r(c_1, \ldots, c_{n_r}) = 1$. Therefore, omitting these two entries from $\boldsymbol{Z}_r$ and $\boldsymbol{\mu}_r$, respectively, yields non-redundant sets. $\boldsymbol{Z}_r^*$ and $\boldsymbol{\mu}_r^*$ are defined similarly, and simple matrix equalities

$$\boldsymbol{\mu}_r^* = B_r \boldsymbol{\mu}_r, \qquad \boldsymbol{Z}_r^* = B_r \boldsymbol{Z}_r \tag{7.4}$$

hold. As an example, consider a bivariate binary outcome vector, with probabilities $\boldsymbol{\mu}_r^* = (\mu_{11}^*, \mu_{12}^*, \mu_{21}^*, \mu_{22}^*)$ and a similar ordering for $\boldsymbol{\mu}_r$. The matrix B_r is found by

$$B_r^{-1} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix}.$$

The marginal counts are given by all counts for which all but one index are equal to their maximal value: $Z_{rjk} \equiv Z_r(c_1, \ldots, c_{j-1}, k, c_{j+1}, \ldots, c_{n_r})$. Bivariate cell counts, i.e., cell counts of a cross-classification of a pair of outcomes, follow from setting all but two indices k_s equal to c_s. Therefore, this description very naturally combines univariate, bivariate, and multivariate information. The ordering needed to stack the multi-indexed counts and probabilities into a vector will be assumed fixed. Several orderings of both $\boldsymbol{Z}_r$ and $\boldsymbol{\mu}_r$ are possible. A natural choice is the lexicographic ordering, but this has the disadvantage of dispersing the univariate marginal counts and means over the entire vector. Therefore, we will typically group the elements by dimensionality first.

7.2 The Bahadur Model

7.2.1 A General Bahadur Model Formulation

Bahadur (1961) introduced this model, with its elegant closed form, but with a number of computational problems surrounding it, stemming from the complicated and highly restrictive form of its parameter space. The model is conceived for binary data and can be introduced using the simpler regression notation, outlined in Section 7.1. Thus, let the binary response Y_{ij} indicate whether or not measurement j on subject i exhibits the event under investigation.

Assume the marginal distribution of Y_{ij} to be Bernoulli with

$$E(Y_{ij}) = P(Y_{ij} = 1) \equiv \pi_{ij}.$$

This expectation can be taken conditional upon covariates X_i. For simplicity, they are suppressed from notation. To start describing the association, the pairwise probability

$$P(Y_{ij_1} = 1, Y_{ij_2} = 1) = E(Y_{ij_1} Y_{ij_2}) \equiv \pi_{ij_1 j_2}$$

needs to be characterized. This "success probability" of two measurements taken in the same subject can be modeled in terms of the two marginal probabilities π_{ij_1} and π_{ij_2}, as well as an association parameter, this being the marginal correlation coefficient in Bahadur's model.

The marginal correlation coefficient assumes the form

$$\text{Corr}(Y_{ij_1}, Y_{ij_2}) \equiv \rho_{ij_1 j_2} = \frac{\pi_{ij_1 j_2} - \pi_{ij_1}\pi_{ij_2}}{[\pi_{ij_1}(1-\pi_{ij_2})\pi_{ij_2}(1-\pi_{ij_2})]^{1/2}}. \tag{7.5}$$

In terms of this association parameter, the joint probability $\pi_{ij_1 j_2}$ can then be written as

$$\pi_{ij_1 j_2} = \pi_{ij_1}\pi_{ij_2} + \rho_{ij_1 j_2}[\pi_{ij_1}(1-\pi_{ij_1})\pi_{ij_2}(1-\pi_{ij_2})]^{1/2}. \tag{7.6}$$

Hence, given the marginal correlation coefficient $\rho_{ij_1 j_2}$ and the univariate probabilities π_{ij_1} and π_{ij_2}, the pairwise probability $\pi_{ij_1 j_2}$ can be calculated with ease.

The first and second moments of the distribution have now been specified. However, a likelihood-based approach requires the complete representation of the joint probabilities of the vector of binary responses in each unit. The full joint distribution $f(\boldsymbol{y})$ of $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})'$ is multinomial with a 2^{n_i} probability vector. Bahadur used, apart from the conventional two-way correlation coefficient, third- and higher- order correlation coefficients to completely specify the joint distribution. To this end, let

$$\varepsilon_{ij} = \frac{Y_{ij} - \pi_{ij}}{\sqrt{\pi_{ij}(1-\pi_{ij})}} \quad \text{and} \quad e_{ij} = \frac{y_{ij} - \pi_{ij}}{\sqrt{\pi_{ij}(1-\pi_{ij})}}, \tag{7.7}$$

where y_{ij} is an actual value of the binary response variable Y_{ij}. Further, let

$$\begin{aligned} \rho_{ij_1 j_2} &= E(\varepsilon_{ij_1}\varepsilon_{ij_2}), \\ \rho_{ij_1 j_2 j_3} &= E(\varepsilon_{ij_1}\varepsilon_{ij_2}\varepsilon_{ij_3}), \\ &\vdots, \\ \rho_{i12\ldots n_i} &= E(\varepsilon_{i1}\varepsilon_{i2}\ldots\varepsilon_{in_i}). \end{aligned} \tag{7.8}$$

Then, the general Bahadur model can be represented by the expression $f(\boldsymbol{y}_i) = f_1(\boldsymbol{y}_i)c(\boldsymbol{y}_i)$, where

$$f_1(\boldsymbol{y}_i) = \prod_{j=1}^{n_i} \pi_{ij}^{y_{ij}}(1-\pi_{ij})^{1-y_{ij}}$$

and

$$c(\boldsymbol{y}_i) = 1 + \sum_{j_1<j_2} \rho_{ij_1j_2} e_{ij_1} e_{ij_2} + \sum_{j_1<j_2<j_3} \rho_{ij_1j_2j_3} e_{ij_1} e_{ij_2} e_{ij_3} + \ldots + \rho_{i12\ldots n_i} e_{i1} e_{i2} \ldots e_{in_i}.$$

Thus, the probability mass function is the product of the independence model $f_1(\boldsymbol{y}_i)$ and the correction factor $c(\boldsymbol{y}_i)$. One view-point is to consider the factor $c(\boldsymbol{y}_i)$ as a model for overdispersion.

7.2.2 The Bahadur Model for Clustered Data

To enhance understanding, let us consider the Bahadur model for the case of exchangeably clustered data. This version of the model was of use for Aerts *et al* (2002) who studied models for clustered data arising in an environmental context.

When the focus is on the special case of clustered data, this assumes on the one hand that each measurement within a unit (individual, family, litter, cluster,...) has the same response probability, i.e., $\pi_{ij} = \pi_i$. On the other hand, it usually implies that within a litter, the associations of a particular order are constant, i.e., $\rho_{ij_1j_2} = \rho_{i(2)}$ for $j_1 < j_2$, $\rho_{ij_1j_2j_3} = \rho_{i(3)}$ for $j_1 < j_2 < j_3$,..., $\rho_{i12\ldots n_i} = \rho_{i(n_i)}$, with $i = 1, \ldots, N$. Given these assumptions, we do not need to know the individual outcomes Y_{ij}, but it suffices to know

$$Z_i = \sum_{j=1}^{n_i} Y_{ij}, \tag{7.9}$$

the number of successes within a unit, with realized value z_i. Under exchangeability (or equicorrelation), the Bahadur model reduces to

$$f_1(\boldsymbol{y}_i) = \pi_i^{z_i}(1-\pi_i)^{n_i - z_i}$$

and

$$c(\boldsymbol{y}_i) = 1 + \sum_{r=2}^{n_i} \rho_{i(r)} \sum_{s=0}^{r} \left(\begin{array}{c} z_i \\ s \end{array} \right) \left(\begin{array}{c} n_i - z_i \\ r - s \end{array} \right) (-1)^{s+r} \lambda_i^{r-2s}, \tag{7.10}$$

with $\lambda_i = \sqrt{\pi_i/(1-\pi_i)}$. The probability mass function of Z_i is given by

$$f(z_i) = \binom{n_i}{z_i} f(\boldsymbol{y}_i).$$

In addition, setting all three- and higher-way correlations equal to zero, the probability mass function of Z_i simplifies further to:

$$f(z_i) \equiv f(z_i|\pi_i, \rho_{i(2)}, n_i) = \binom{n_i}{z_i} \pi_i^{z_i}(1-\pi_i)^{n_i - z_i}$$

$$\times \left[1 + \rho_{i(2)} \left\{ \begin{pmatrix} n_i - z_i \\ 2 \end{pmatrix} \frac{\pi_i}{1-\pi_i} - z_i(n_i - z_i) + \begin{pmatrix} z_i \\ 2 \end{pmatrix} \frac{1-\pi_i}{\pi_i} \right\} \right]. \tag{7.11}$$

This very tractable expression of the Bahadur probability mass function is advantageous over other representations, such as the multivariate probit (Section 7.6) and Dale (Section 7.7) models, for which no closed form solutions, free of integrals, exist. However, a drawback is the fact that the correlation between two responses is highly constrained when the higher order correlations are removed. Even when higher order parameters are included, the parameter space of marginal parameters and correlations is known to be peculiar. Bahadur (1961) discusses restrictions on the correlation parameters. The second-order approximation in (7.11) is only useful if it is a probability mass function. Bahadur indicates that the sum of the probabilities of all possible outcomes is one. However, depending on the values of π_i and $\rho_{i(2)}$, expression (7.11) may fail to be non-negative for some outcomes. The latter results in restrictions on the parameter space, which, in case of the second-order approximation, are described by Bahadur (1961). From these, it can be deduced that the lower bound for $\rho_{i(2)}$ approaches zero as the cluster size increases. However, it is important to note that also the upper bound for this correlation parameter is constrained. Indeed, even though it is one for clusters of size two, the upper bound varies between $1/(n_i - 1)$ and $2/(n_i - 1)$ for larger clusters. Taking a cluster size of, for example, 12, the upper bound is in the range $(0.09; 0.18)$. Kupper and Haseman (1978) present numerical values for the constraints on $\rho_{i(2)}$ for choices of π_i and n_i. Restrictions for a specific version where a third-order association parameter is included as well are studied by Prentice (1988), while a more general situation is studied by Declerck, Aerts, and Molenberghs (1998). See also Aerts *et al* (2002).

The marginal parameters π_i and $\rho_{i(2)}$ can be modeled using a composite link function. Because Y_{ij} is binary, the logistic link function for π_i is a natural choice. In principle, any link function, such as the probit link, the log-log link or the complementary log-log link, could be chosen. A convenient transformation of $\rho_{i(2)}$ is Fisher's z-transform. This leads to the following generalized linear regression relations

$$\begin{pmatrix} \ln\left(\frac{\pi_i}{1-\pi_i}\right) \\ \ln\left(\frac{1+\rho_{i(2)}}{1-\rho_{i(2)}}\right) \end{pmatrix} \equiv \boldsymbol{\eta}_i = X_i \boldsymbol{\beta}, \tag{7.12}$$

where X_i is a design matrix and $\boldsymbol{\beta}$ is a vector of unknown parameters. Note that (7.12) is not encompassed by (6.2).

Denote the log-likelihood contribution of the ith unit by

$$\ell_i = \ln f(z_i|\pi_i, \rho_{i(2)}, n_i).$$

The maximum likelihood estimator $\widehat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ is defined as the solution to the score equations $\boldsymbol{U}(\boldsymbol{\beta}) = \mathbf{0}$. The score function $\boldsymbol{U}(\boldsymbol{\beta})$ can be written as

$$\boldsymbol{U}(\boldsymbol{\beta}) = \sum_{i=1}^{N} X_i'(T_i')^{-1} L_i \tag{7.13}$$

where

$$T_i = \frac{\partial \boldsymbol{\eta}_i}{\partial \boldsymbol{\Theta}_i} = \begin{pmatrix} \frac{\partial \eta_{i1}}{\partial \pi_i} & \frac{\partial \eta_{i2}}{\partial \pi_i} \\ \frac{\partial \eta_{i1}}{\partial \rho_{(2)}} & \frac{\partial \eta_{i2}}{\partial \rho_{i(2)}} \end{pmatrix} = \begin{pmatrix} \frac{1}{\pi_i(1-\pi_i)} & 0 \\ 0 & \frac{2}{(1-\rho_{i(2)})(1+\rho_{i(2)})} \end{pmatrix},$$

$$L_i = \frac{\partial \ell_i}{\partial \boldsymbol{\Theta}_i} = \begin{pmatrix} \frac{\partial \ell_i}{\partial \pi_i} \\ \frac{\partial \ell_i}{\partial \rho_{i(2)}} \end{pmatrix}$$

and $\boldsymbol{\Theta}_i = (\pi_i, \rho_{i(2)})'$, the set of natural parameters. A Newton-Raphson algorithm can be used to obtain the maximum likelihood estimates $\widehat{\boldsymbol{\beta}}$ and an estimate of the asymptotic covariance matrix of $\widehat{\boldsymbol{\beta}}$ can be obtained from the observed information matrix at maximum.

When including higher order correlations, implementing the score equations and the observed information matrices becomes increasingly cumbersome. Although the functional form (7.13) does not change, the components T_i and L_i become fairly complicated. Fisher's z transform can be applied to all correlation parameters $\rho_{i(r)}$. The design matrix X_i would then extend in a straightforward fashion as well. Unfortunately, fitting a higher order Bahadur model, is not straightforward, due to increasingly complex restrictions on the parameter space.

Observing that interest is often restricted to the marginal mean function and the pairwise association parameter, one can replace a full likelihood approach by estimating equations where only the first two moments are modeled and working assumptions are adopted about third- and fourth-order moments. This is treated as one of the extensions to standard generalized estimating equations in Section 8.5. See also Liang, Zeger, and Qaqish (1992).

7.2.3 Analysis of the NTP Data

Table 7.1 presents parameter estimates and standard errors for the Bahadur model, in the specific context of clustered outcomes as in Section 7.2.2, fitted to several outcomes in three of the NTP datasets, described in Section 2.7. Apart from the external, visceral, and skeletal malformation outcomes, we also consider the so-called collapsed outcome, which is 1 if at least one of the three malformations occur and 0 otherwise.

TABLE 7.1. *NTP Data. Parameter estimates (standard errors) for the Bahadur model, fitted to various outcomes in three studies.* β_0 *and* β_d *are the marginal intercept and dose effect, respectively;* β_a *is the Fisher z transformed correlation;* ρ *is the correlation.*

Outcome	Parameter	DEHP	EG	DYME
External	β_0	-4.93(0.39)	-5.25(0.66)	-7.25(0.71)
	β_d	5.15(0.56)	2.63(0.76)	7.94(0.77)
	β_a	0.11(0.03)	0.12(0.03)	0.11(0.04)
	ρ	0.05(0.01)	0.06(0.01)	0.05(0.02)
Visceral	β_0	-4.42(0.33)	-7.38(1.30)	-6.89(0.81)
	β_d	4.38(0.49)	4.25(1.39)	5.49(0.87)
	β_a	0.11(0.02)	0.05(0.08)	0.08(0.04)
	ρ	0.05(0.01)	0.02(0.04)	0.04(0.02)
Skeletal	β_0	-4.67(0.39)	-2.49(0.11)	-4.27(0.61)
	β_d	4.68(0.56)	2.96(0.18)	5.79(0.80)
	β_a	0.13(0.03)	0.27(0.02)	0.22(0.05)
	ρ	0.06(0.01)	0.13(0.01)	0.11(0.02)
Collapsed	β_0	-3.83(0.27)	-2.51(0.09)	-5.31(0.40)
	β_d	5.38(0.47)	3.05(0.17)	8.18(0.69)
	β_a	0.12(0.03)	0.28(0.02)	0.12(0.03)
	ρ	0.06(0.01)	0.14(0.01)	0.06(0.01)

Specifically, a marginal logit model linear in dose and a constant association $\rho_{i(2)} = \rho_{(2)}$ are chosen, implying that X_i in (7.12) takes the form:

$$X_i = \begin{pmatrix} 1 & d_i & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{7.14}$$

and

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_d \\ \beta_a \end{pmatrix}, \tag{7.15}$$

where β_0 is an intercept, β_d the dose effect, and β_a the Fisher z transformed correlation.

We conclude that the background risk for malformation in all cases is very small, but that it increases with dose. For the external malformation outcome in the DEHP study, for example, the background risk is estimated to be small:

$$\frac{e^{-4.93}}{1 + e^{-4.93}} = 0.0071.$$

When the dose level equals its highest value ($d = 1.0$), the risk becomes

$$\frac{e^{-4.93+5.15}}{1+e^{-4.93+5.15}} = 0.55,$$

implying that more than one out of two foetuses would be malformed.

The dose-response curve that follows from the marginal logistic regression:

$$P(Y_{ij} = 1|d_i) = \frac{e^{-4.93+5.15d_i}}{1+e^{-4.93+5.15d_i}}$$

is supplemented with information on the association. In addition, one obtains a correlation of

$$\widehat{\rho} = \frac{e^{\widehat{\beta}_a} - 1}{e^{\widehat{\beta}_a} + 1} = 0.05.$$

Although small, the within-cluster association is significant, as it is for most but not all outcomes.

7.2.4 Analysis of the Fluvoxamine Trial

The fluvoxamine trial, introduced in Section 2.4, were analyzed in some detail in Chapter 6. In Section 6.6, several two-way contingency tables, either based on a single outcome at two measurement occasions, or side effects and therapeutic effect at the same time, were analyzed. This initial setting was extended to categorical covariates and three-way tables in Sections 6.6.1 and 6.6.2, respectively. Using the Bahadur model, we are able to extend this further to sequences of arbitrary length, and a combination of continuous and categorical covariates. This is true in principle, as the Bahadur model is restrictive due to constraints on the parameter space, as stated before. In Section 7.2.3, we were able to analyze the NTP data, with dose treated as a continuous covariate, in spite of the fact that some litters consist of around 15 littermates, but we could do so only by carefully exploiting the exchangeable nature of the data, with only three regression parameters as a result.

Here, we would like to study three side-effects measures simultaneously, regressed on age and sex of the patient, prior duration of the mental illness, and initial severity of the disease. We are confronted with two stumbling blocks. First, because the Bahadur model is formulated for binary outcomes, we need to collapse the original four-category side effects outcome into a dichotomous variable. This is done by transforming the lower two levels of the side effects variable into 0 and the upper two into 1. Second, due to the parameter restrictions, it was not possible to consider all four covariates simultaneously. Thus, we resrict attention to the sex and prior duration variables. Parameter estimates are given in Table 7.2. The three-way correlation coefficient is set to zero. The effect of the covariates is not significant, but the correlation parameters are. For ease of interpretation,

TABLE 7.2. *Fluvoxamine Study. Longitudinal analysis using Bahadur's model. The side effects at three successive times are regressed on sex and duration. The entries represent the parameter estimates (standard errors).*

	Side effects at time		
Parameter	1	2	3
Intercept	0.81(0.47)	0.15(0.37)	0.57(0.44)
Sex	-0.56(0.26)	0.02(0.20)	0.14(0.24)
Duration	0.008(0.009)	0.01(0.01)	-0.01(0.01)
Association Parameters			
12	13	23	123
Fisher z transformed correlations			
1.42(0.16)	0.84(0.13)	1.37(0.15)	—
Correlations			
0.61(0.05)	0.39(0.05)	0.59(0.05)	—

the Fisher z transformed correlation, as they figure in the model and fitting program, are transformed again to their original scale, supplemented with standard errors obtained by means of the delta method.

Thus, while the Bahadur model can be of some use in a restricted number of situations, including exchangeably clustered outcomes, there are practical limitations when used in multivariate and longitudinal settings. Therefore, in spite of the relatively simple model formulation, there is a need for alternative models, when a full likelihood based analysis of a marginally formulated model is envisaged. In the next section, we will sketch a general framework to achieve this, then consider the probit (Section 7.6) and Dale model (Section 7.7) cases, whereafter we analyze several sets of data. In particular, we return to the fluvoxamine study in Section 7.11.

7.3 A General Framework for Fully Specified Marginal Models

We will now use the contingency table notation laid out in Section 7.1. A marginal model can be built in several ways. In a few cases it is possible to write down the multivariate probability mass function immediately, such as in the Bahadur model of Section 7.2. In most cases, one starts from the univariate margins, on top of which an association structure is assumed, of the second and higher orders, to complete model specification. We will proceed here in this at first sight laborious way.

By means of (7.3), the set of cell probabilities at design level r has been defined. To proceed with modeling, we typically map these onto a set of link functions $\boldsymbol{\eta}_r$, which can then be expressed in terms of parameters of scientific interest. In the Bahadur model for clustered data, this was done by means of (7.12). In general, we map the C_r-vector $\boldsymbol{\mu}_r$ ($C_r = c_1 \cdot c_2 \cdot \ldots \cdot c_{n_r}$) to

$$\boldsymbol{\eta}_r = \boldsymbol{\eta}_r(\boldsymbol{\mu}_r), \tag{7.16}$$

a $C_r^{'}$-vector. In many models, $C_r = C_r^{'}$, and $\boldsymbol{\eta}_r$ and $\boldsymbol{\mu}_r$ have the same ordering. A counterexample is provided by the probit model, where the number of link functions is smaller than the number of mean components, as soon as $n_r > 2$, i.e., there are three of more repeated measures [see (7.25)–(7.27)]. As already indicated in Section 6.2.1, an important class of link functions is discussed by McCullagh and Nelder (1989):

$$\boldsymbol{\eta}_r(\boldsymbol{\mu}_r) = C \ln(A\boldsymbol{\mu}_r), \tag{7.17}$$

a definition in terms of contrasts of log probabilities, where the probabilities involved are linear combinations $A\boldsymbol{\mu}_r$. The same class was presented in (6.2) for the specific case of marginal models for a contingency table.

7.3.1 Univariate Link Functions

We consider particular choices of link functions. To this end, let us abbreviate the univariate marginal probabilities by

$$\mu_{rjk} = \mu_r(c_1, \ldots, c_{j-1}, k, c_{j+1}, \ldots, c_{n_r}),$$

then the logit link becomes

$$\eta_{rjk} = \ln(\mu_{rjk}) - \ln(1 - \mu_{rjk}) = \text{logit}(\mu_{rjk}). \tag{7.18}$$

Some link functions that are occasionally of interest, such as the probit or complementary log-log link are not supported by (7.17) but they can easily be included in (7.16). The probit link is

$$\eta_{rjk} = \Phi_1^{-1}(\mu_{rjk}),$$

with Φ_1 the univariate standard normal distribution.

7.3.2 Higher-order Link Functions

However, univariate links alone do not fully specify $\boldsymbol{\eta}_r$ and hence leave the joint distribution partly undetermined. Full specification of the association requires addressing the form of pairwise and higher-order probabilities. First, we will consider the pairwise associations. Let us denote the bivariate probabilities, pertaining to the j_1th and j_2th outcomes, by

$$\mu_{r,jh,k\ell} = \mu_r(c_1, \ldots, c_{j-1}, k, c_{j+1}, \ldots, c_{h-1}, \ell, c_{h+1}, \ldots, c_{n_r}).$$

TABLE 7.3. *Association structure of selected marginal models.*

Name	Association structure	Equation
Success probability	Logit of joint probability	(7.19)
Bahadur model	Marginal correlation coefficients	(7.5)
Dale model	Global marginal odds ratio	(7.21)–(7.23)
	Local marginal odds ratio	(7.24)
Probit model	Polychoric correlation	(7.25)–(7.27)

Some association parameterizations are summarized in Table 7.3.

The success probability parameterization of Ekholm (1991) consists of choosing a link function for the univariate marginal means (e.g., a logit link) and then applying the same link function to the two- and higher order success probabilities (i.e., the probabilities for observing a single success when looking at one outcome at a time, a pair of successes when looking at pairs of outcomes,...). For categorical data, a logit link for two-way probabilities is given by

$$\eta_{r,jh,k\ell} = \ln(\mu_{r,jh,k\ell}) - \ln(1 - \mu_{r,jh,k\ell}) = \text{logit}(\mu_{i,jh,k\ell}), \tag{7.19}$$

for $k = 1, \ldots, c_j - 1$ and $\ell = 1, \ldots, c_h - 1$. Ekholm, Smith, and McDonald (1995) and Ekholm, McDonald, and Smith (2000) used these to define dependence ratios, in the specific case of binary data. The marginal correlation coefficient (Bahadur 1961) is defined as

$$\rho_{r,jh,k\ell} = \frac{\mu_{r,jh,k\ell} - \mu_{rjk}\mu_{rh\ell}}{\sqrt{\mu_{rjk}(1 - \mu_{rjk})\mu_{rh\ell}(1 - \mu_{rh\ell})}}. \tag{7.20}$$

This model has been developed, for binary data, including the higher order correlations, in Section 7.2.

We will put strong emphasis on the marginal global odds ratio, defined by

$$\psi_{r,jh,k\ell} = \frac{(\mu_{r,jh,k\ell})(1 - \mu_{rjk} - \mu_{rh\ell} + \mu_{r,jh,k\ell})}{(\mu_{rh\ell} - \mu_{r,jh,k\ell})(\mu_{rjk} - \mu_{r,jh,k\ell})} \tag{7.21}$$

and usefully modeled on the log scale as

$$\begin{aligned} \eta_{r,jh,k\ell} &= \ln \psi_{r,jh,k\ell} \\ &= \ln(\mu_{r,jh,k\ell}) - \ln(\mu_{rjk} - \mu_{r,jh,k\ell}) \\ &\quad - \ln(\mu_{rh\ell} - \mu_{r,jh,k\ell}) + \ln(1 - \mu_{rjk} - \mu_{rh\ell} + \mu_{r,jh,k\ell}). \end{aligned}$$

Higher order global odds ratios are easily introduced, for example, using ratios of conditional odds (ratios). Let

$$\mu_{rj|h}(z_h) = P(Z_{rijk_j} = 1 | Z_{rihk_h} = z_h, X_r, \boldsymbol{\beta}) \tag{7.22}$$

be the conditional probability of observing a success at occasion j, given the value z_h is observed at occasion h, and write the corresponding conditional odds as

$$\psi_{rj|h}(z_h) = \frac{\mu_{rj|h}(z_h)}{1 - \mu_{rj|h}(z_h)}.$$

The pairwise marginal odds ratio, for occasions j and h, is defined as

$$\begin{aligned} \psi_{rjh} &= \frac{\{\mathrm{pr}(Z_{rijk_j} = 1, Z_{rihk_h} = 1)\}\{\mathrm{pr}(Z_{rijk_j} = 0, Z_{rihk_h} = 0)\}}{\{\mathrm{pr}(Z_{rijk_j} = 0, Z_{rihk_h} = 1)\}\{\mathrm{pr}(Z_{rijk_j} = 1, Z_{rihk_h} = 0)\}} \\ &= \frac{\psi_{rj|h}(1)}{\psi_{rj|h}(0)}, \end{aligned}$$

in accordance with (7.21). This formulation can be exploited to define the higher order marginal odds ratios in a recursive fashion:

$$\psi_{rj_1\ldots j_m j_{m+1}} = \frac{\psi_{rj_1\ldots j_m|j_{m+1}}(1)}{\psi_{rj_1\ldots j_m|j_{m+1}}(0)}, \tag{7.23}$$

where $\psi_{rj_1\ldots j_m|j_{m+1}}(z_{m+1})$ is defined by conditioning all probabilities occurring in the expression for $\psi_{rj_1\ldots j_m}$ on $Z_{rij_{m+1}} = z_{j_{m+1}}$. The choice of the variable to condition on is immaterial. Observe that multi-way marginal global odds ratios are defined solely in terms of conditional probabilities. We will return to these in Section 7.7.4, when more detail is given about the multivariate Dale model.

Another type of marginal odds ratios is given by the marginal *local* odds ratios. These were used in Section 6.2.2. This type of odds ratio changes (7.21) to

$$\psi^*_{r,jh,k\ell} = \frac{\mu^*_{r,jh,k\ell}\mu^*_{r,jh,k+1,\ell+1}}{\mu^*_{r,jh,k+1,\ell}\mu^*_{r,jh,k,\ell+1}}, \tag{7.24}$$

with the cell probabilities as in (7.3). Higher order marginal local odds ratios are constructed in the same way as their global counterparts. The global odds ratio model will be studied further in Section 7.7.

Observe that the multivariate probit model (Ashford and Sowden 1970, Lesaffre and Molenberghs 1991) also fits within the class defined by (7.16). To see this, let $g = h^{-1}$. For three categorical outcome variables, the inverse link is specified by

$$\mu_{rjk} = \Phi_1(\eta_{rjk}), \tag{7.25}$$

$$\mu_{r,jh,k\ell} = \Phi_2(\eta_{rjk}, \eta_{rh\ell}, \eta_{r,jh,k\ell}), \tag{7.26}$$

$$\mu_{r,123,k\ell m} = \Phi_3(\eta_{r1k}, \eta_{r2\ell}, \eta_{r3m}, \eta_{r,12,k\ell}, \eta_{r,13,km}, \eta_{r,23,\ell m}), \tag{7.27}$$

where the notation for the three-way probabilities is obvious. The association links $\eta_{r,jh,k\ell}$ represent any transform (e.g., Fisher's z-transform such

as in the Bahadur model of Section 7.2) of the correlation coefficient. It is common practice to keep each correlation constant throughout a table, rather than having it depend on the categories: $\eta_{r,jh,k\ell} \equiv \eta_{r,jh}$. Relaxing this requirement may still give a valid set of probabilities, but the correspondence between the categorical variables and a latent multivariate normal variable is lost. Finally, observe that univariate links and bivariate links (representing correlations) fully determine the joint distribution. This implies that the mean vector and the link vector will have a different length, except in the univariate and bivariate cases.

In summary, marginal models are characterized by jointly specifying marginal response functions and marginal association measures. Models can be classified by the association measures, as exemplified in Table 7.3.

Finally, model formulation is completed by constructing appropriate design matrices. Let us give an example to indicate how model assumptions are reflected by choosing particular types of design. We deliberately restrict ourselves to linear predictors, while, in principle, there is no obstacle to include non-linear effects (Chapter 20). The design matrix X_r for the rth design level includes all information which is needed to model both the marginal mean functions and associations. Each row corresponds to an element in the vector of link functions $\boldsymbol{\eta}_r$. Its generality is best illustrated using an example.

Consider the case of three outcomes, recorded on a three-point scale. Let the measurement times be $t_1 \equiv 0$, t_2, and t_3. Assume the recording of four explanatory variables, $x_1, \ldots, x_4$, with only x_3 and x_4 time-varying. We first turn our attention to the marginal distributions. Let x_1 have a constant effect on each outcome, i.e., a single parameter describes the effect of x_1 on the cumulative logits of the three outcome probabilities. On the other hand, the effect of x_2 is allowed to change over time. We also introduce a single parameter to describe the effect of x_3 and three separate parameters to account for the influence of x_4. These assumptions call for the following parameter vector

$$\boldsymbol{\beta}_1 = (\beta_{01}, \beta_{02}, \tau_2, \tau_3, \beta_1, \beta_{21}, \beta_{22}, \beta_{23}, \beta_3, \beta_{41}, \beta_{42}, \beta_{43})',$$

with β_{0k} intercepts, τ_j the effect of measurement time j, β_1 and β_3 the parameters, needed to describe the effect of x_1 and x_3 respectively, and β_{tj} the parameter describing the effect of $x_t^{(j)}$ at time t ($t = 2, 4; j = 1, 2, 3$). Next, assume that the two-way associations depend on the pair of variables they refer to, as well as on the cumulative category within that variable. Finally, assume dependence on the covariate x_{1r}. This introduces extra parameters

$$\boldsymbol{\alpha}_2 = (\gamma, \gamma_{11}, \gamma_{12}, \gamma_{21}, \gamma_{22}, \gamma_{31}, \gamma_{32}, \phi_1, \phi_2, \phi_3)',$$

with γ the intercept, γ_{jk} the dependence on category k of outcome j ($j = 1, 2; k = 1, 2$), and ϕ_j the dependence on x_1. Finally, assume a constant

	β_{01}	β_{02}	τ_2	τ_3	β_1	β_{21}	β_{22}	β_{23}	β_3	β_{41}	β_{42}	β_{43}
$\eta_r(1,3,3)$	1	0	0	0	x_{1r}	x_{2r}	0	0	$x_{3r}^{(1)}$	$x_{4r}^{(1)}$	0	0
$\eta_r(2,3,3)$	0	1	0	0	x_{1r}	x_{2r}	0	0	$x_{3r}^{(1)}$	$x_{4r}^{(1)}$	0	0
$\eta_r(3,1,3)$	1	0	1	0	x_{1r}	0	x_{2r}	0	$x_{3r}^{(2)}$	0	$x_{4r}^{(2)}$	0
$\eta_r(3,2,3)$	0	1	1	0	x_{1r}	0	x_{2r}	0	$x_{3r}^{(2)}$	0	$x_{4r}^{(2)}$	0
$\eta_r(3,3,1)$	1	0	0	1	x_{1r}	0	0	x_{2r}	$x_{3r}^{(3)}$	0	0	$x_{4r}^{(3)}$
$\eta_r(3,3,2)$	0	1	0	1	x_{1r}	0	0	x_{2r}	$x_{3r}^{(3)}$	0	0	$x_{4r}^{(3)}$

FIGURE 7.1. *Design matrix for marginal means and pairwise associations. Marginal means.*

value for the three-way associations, α_3 say. The entire parameter vector is denoted as

$$\boldsymbol{\beta} = (\boldsymbol{\beta}^{'}, \boldsymbol{\alpha}_2^{'}, \alpha_3)'.$$

The design matrix for design level r, X_r, is block diagonal with blocks X_{r1} (mean functions, shown in Figure 7.1), X_{r2} (pairwise association, shown in Figure 7.2), and X_{r3} (three-way association).

Observe that, apart from the intercepts β_{0k}, the design is identical for each cumulative logit in Figures 7.1 and 7.2. This reflects the proportional odds assumption when marginal logits are used. If this assumption is considered unrealistic, the design can be generalized without any difficulty. Nominal covariates and interactions between covariates are also easily included.

The second block of the design matrix, X_{2r}, corresponds to the pairwise associations and is given by Figure 7.2. Finally, the design for the three-way associations in our example is a 8×1 column vector of ones, corresponding to the 8 link functions $\eta_r(k_1, k_2, k_3)$ ($k_j = 1, 2; j = 1, 2, 3$). Replacing the elements of this vector by zeros has the effect of setting higher order association components equal to one (zero on the log scale).

Generalizations include non-block diagonal designs, and structured association such as exchangeable association, temporal association (as introduced by Fitzmaurice and Lipsitz 1995), and banded association. In many circumstances, the association structure of a given table, representing a two- or multi-way classification of several variables is of direct interest, rather than the dependence of the outcomes on covariates. Association measures are extensively studied in Goodman (1981b). We will discuss these further in Chapter 11. With the current approach, we are also able to explore the association structure of contingency tables. A typical form for the linear predictor, pertaining to a two-way association, is given by

$$\eta_{r,jh,k\ell} = \gamma + \gamma_{jh} + \gamma_{jk} + \gamma_{h\ell} + \delta_{jk}\delta_{h\ell},$$

	γ	γ_{11}	γ_{12}	γ_{21}	γ_{22}	γ_{31}	γ_{32}	ϕ_1	ϕ_2	ϕ_3
$\eta_r(1,1,3)$	1	1	0	1	0	0	0	x_{1r}	0	0
$\eta_r(1,2,3)$	1	1	0	0	1	0	0	x_{1r}	0	0
$\eta_r(2,1,3)$	1	0	1	1	0	0	0	x_{1r}	0	0
$\eta_r(2,2,3)$	1	0	1	0	1	0	0	x_{1r}	0	0
$\eta_r(1,3,1)$	1	1	0	0	0	1	0	0	x_{1r}	0
$\eta_r(1,3,2)$	1	1	0	0	0	0	1	0	x_{1r}	0
$\eta_r(2,3,1)$	1	0	1	0	0	1	0	0	x_{1r}	0
$\eta_r(2,3,2)$	1	0	1	0	0	0	1	0	x_{1r}	0
$\eta_r(3,1,1)$	1	0	0	1	0	1	0	0	0	x_{1r}
$\eta_r(3,1,2)$	1	0	0	1	0	0	1	0	0	x_{1r}
$\eta_r(3,2,1)$	1	0	0	0	1	1	0	0	0	x_{1r}
$\eta_r(3,2,2)$	1	0	0	0	1	0	1	0	0	x_{1r}

FIGURE 7.2. *Design matrix for marginal means and pairwise associations. Pairwise associations.*

including an overall intercept, effects specific to times j and h: γ_{ts}, 'row' and 'column' effects γ_{jk} and $\gamma_{h\ell}$ and multiplicative interactions. Obviously, this model is overparameterizing the association, calling for the usual restrictions.

7.4 Maximum Likelihood Estimation

In the previous section, a general framework for formulating marginal models has been sketched. We will zoom in on specific instances, the multivariate probit and Dale models, in Sections 7.6 and 7.7, respectively. But before doing so, we will discuss a general form for the likelihood equations and discuss algorithms to obtain the maximum likelihood estimator, as well as estimates of precision. When performing maximum likelihood estimation for marginal models, a crucial element is the determination of the joint probabilities. Details on these important but technical aspects are provided in Appendix 7.12.

7.5 An Influenza Study

Consider the following clinical trial. A group of 498 medical students, between 17 and 29 years of age (median 21 years), are randomized to two

treatment groups. Those in the HI group receive hepatitis B vaccination (H), followed by influenza vaccination (I), whereas the reverse order is applied in the IH group. For each type of vaccination, vaccines from a company A and a company B are used. In each treatment period, the vaccines are evaluated with respect to the side effects they caused. We are interested in the outcomes *headache* and *respiratory problems*. Because both outcomes are measured in each of the two periods, we obtain a four-dimensional response variable. It is of interest to assess the strength of the association between both headache outcomes, between both respiratory outcomes, as well as to determine whether both complaints are dependent. In addition, a three-point ordinal variable, level of pain, is recorded for six days in row during the first period, supplementing the cross-over study with a longitudinal one. The first three days will be evaluated here. In order to analyze these data, we need tools for longitudinal categorical data, as well as tools for more complex designs, such as cross-over trials with several outcomes in each period. Whereas the association between outcomes is often considered a nuisance characteristic in longitudinal studies, it is usually of direct interest in multivariate settings, such as the bivariate cross-over study considered here.

We analyze the cross-over and longitudinal parts of the influenza study in turn.

7.5.1 The Cross-over Study

Let us now analyze presence/absence of headache (H) and presence/absence of respiratory problems (R), measured in both trial periods. Explicitly, the probability of absence of symptoms will be modeled. We combine marginal logits with marginal log odds ratios. The modeling is in stages. First, period effect is included. Then, a contrast between the two companies, a contrast between the two vaccinations, and an interaction term between companies and treatments is added. Further, the baseline covariates 'age' (in years) and 'sex' (0 =male, 1 =female) are included. There are three types of two-way association: between the two headache outcomes, between the two respiratory problems outcomes, and between a headache and a respiratory outcome. The two-way associations are graphically depicted in Figure 7.3. Three-way and four-way associations are assumed to be constant throughout. The results are presented in Table 7.4.

Respiratory problems are on average very infrequent, as can be seen from the high value of the intercept. For both outcomes, there is a significant period effect: there are less headaches and respiratory problems in the second period. Also, the influenza vaccination causes less headaches, but more respiratory problems. Headaches are more frequently seen in younger people, whereas the opposite holds for respiratory problems. Men suffer more from headaches after vaccination than women. The odds ratio between two respiratory problems is high (7.9), while a somewhat smaller association is

TABLE 7.4. *Influenza Study. Parameter estimates (standard errors) for the cross-over trial.*

Effect	Estimate (s.e.)
Headache	
intercept	0.055 (1.092)
period effect	0.434 (0.140)
company A effect	-0.341 (0.221)
influenza effect	0.132 (0.212)
company A-influenza interaction	-0.053 (0.281)
age	0.052 (0.054)
sex	0.875 (0.217)
Respiratory problems	
intercept	5.217 (1.297)
period effect	0.167 (0.156)
company A effect	-0.229 (0.267)
influenza effect	-0.119 (0.226)
company-influenza interaction	0.257 (0.312)
age	-0.159 (0.063)
sex	0.133 (0.243)
Associations (log odds ratios)	
headache-headache (ψ_{HH})	1.130 (0.251)
respiratory-respiratory (ψ_{RR})	2.061 (0.309)
headache-respiratory (ψ_{HR})	1.090 (0.191)
three-way interaction	0.219 (0.395)
four-way interaction	2.822 (1.462)

seen between the pair of headache measures (3.1) and between the mixed pair (3.0). This is due to the fact that respiratory problems are more severe and probably more strongly related with vaccination than headache, which can have various causes. Extending the two-way association structure to include a company effect was not significant. We found no higher-order association, although the four-way association was close to significance.

7.5.2 The Longitudinal Study

Pain was measured on six consecutive days after vaccination. Changes in response are mainly observed during the first three days. Significant predictors for the evolution of pain level are 'sex,' 'age,' the use of medication ('med'), and the actual vaccination. The effect of all covariates is allowed to change over time. As there are four vaccinations, we decompose them into two factors (company, influenza, and the interaction). At each measurement time, there are two intercepts, corresponding to two cumulative logits [no pain (0) versus pain (1 and 2); no or mild pain (0 and 1) versus moderate

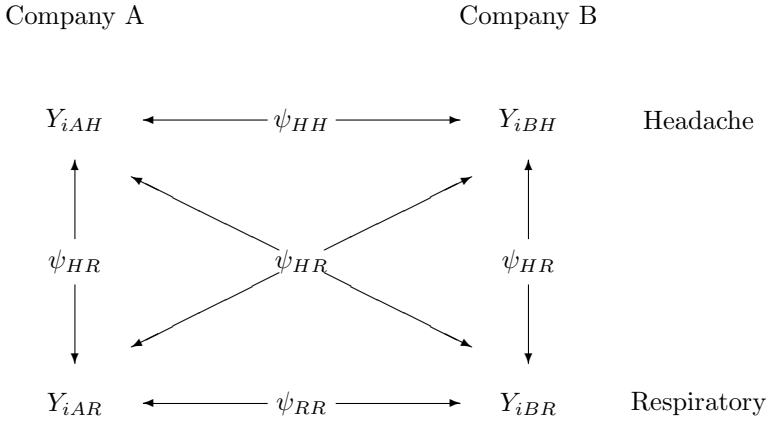

FIGURE 7.3. *Influenza Study. Association structure for the cross-over study.*

pain (2)]. All covariates are allowed to have a different effect at each measurement, presented as 'sex' (overall), 'sex' (linear), and 'sex' (quadratic). The results are presented in Table 7.5. We observe strong quadratic time effects for company A and for the interaction between company A and influenza. Considering the hepatitis vaccine for company B as the baseline, the differences (for each measurement time) on the logit scale between each vaccine and the baseline are: for the influenza vaccine of company A: −5.33, 0.85, and −1.10; for the influenza vaccine of company B: −1.36, 1.95, and 0.43; for the hepatitis vaccine of company A: −4.18, −1.15, and −1.71. The combination of a strong interaction between company and type of vaccine and of the change of the effects over time, yields a complex picture. As the outcomes are modeled via marginal logits, they are interpreted using standard logistic regression methodology. Making comparisons for the three measurement times, we are able to study the evolution of differences over time.

7.6 The Multivariate Probit Model

Section 7.3 presented a general framework to formulate marginal models for categorical data. One of the models mentioned in particular was the multivariate probit model. In this section, we will study this model in more detail. We will refer to the bivariate version as the BPM (bivariate probit

TABLE 7.5. *Influenza Study. Parameter estimates (standard errors) for the longitudinal data.*

	Estimate (s.e.)		
Marginal parameters			
Effect	Average	Linear	Quadratic
intercept 1	-2.34(1.00)	1.24(0.93)	-0.60(0.36)
intercept 2	0.34(1.00)	0.89(0.93)	-0.80(0.37)
age	0.15(0.05)	-0.01(0.05)	0.05(0.02)
sex	0.43(0.19)	-0.31(0.17)	-0.01(0.07)
medication	-0.47(0.22)	-0.26(0.19)	0.07(0.08)
company A effect	1.23(0.27)	0.37(0.27)	-0.39(0.11)
influenza effect	-0.74(0.21)	0.08(0.19)	-0.11(0.07)
company-influenza interaction	-1.06(0.34)	-0.26(0.32)	0.34(0.12)
Associations (log odds ratios)			
time 1–time 2	1.81(0.21)		
time 1–time 3	0.98(0.26)		
time 2–time 3	3.40(0.41)		
three-way interaction	0.88(0.63)		

model), TPM (trivariate probit model) for the trivariate version, and MPM (multivariate probit model) for the general case.

7.6.1 Probit Models

The bivariate probit methodology will be introduced with the data from the BIRNH study, where smoking and drinking behavior in a general population is studied (Kesteloot, Geboers, and Joossens 1989). Risk factors for these two endpoints are determined but the main interest lies in the association between smoking and drinking. The main question is whether this association changes over demographic variables such as age, sex, and social status. The same data will be analyzed with the bivariate Dale model (BDM).

The BIRNH (Belgian Interuniversity Research on Nutrition and Health) study was conducted in the period 1980–1984 (Kesteloot, Geboers, and Joossens 1989). A stratified random sample from 42 counties of Belgium was taken to study the effect of nutrition on health. We are interested in modeling the relationship between alcohol drinking and smoking habits on the one hand and certain demographic variables on the other hand. Complete data were obtained from 5485 men and 4856 women.

Alcohol is divided into 4 classes according to daily intake: (0, 0–10, 10–30, >30). Smoking is divided into 3 classes: (never smoked, ex-smoker, smoker). Predictors variables are: 'sex' (coded as 1 for males and 2 for females), 'age,' 'weight,' 'height,' body mass index ('BMI'), 'site' within

Belgium (1: Flanders, 2: elsewhere), and social status. Age, weight, and height are categorized using the midpoints of their 10 unit classes, for BMI we chose classes of 5 units. Two variables describe social status: 'social 1' [employment (1) versus unemployment or housework (0)] and 'social 2' [working at home (1) *versus* working outside (0)]. Four questions were of interest:

1. Is there a relationship between drinking and smoking behavior?
2. Is alcohol consumption related to the demographic variables?
3. Is smoking behavior related to the demographic variables?
4. Is the association between smoking and drinking dependent on certain demographic variables, i.e., does the relationship change in certain subgroups?

It will be shown below that the BPM is adequate to answer all those questions.

7.6.2 Tetrachoric and Polychoric Correlation

Assume first that we have divided the study population into drinkers/non-drinkers and smokers/non-smokers. With a homogeneous group, a fourfold table will show whether there is an association between drinking and smoking. For measures of association we can take the cross-product (odds) ratio or the tetrachoric correlation ρ, i.e., the correlation of the underlying, doubly dichotomized bivariate normal, which was introduced almost a century ago (Pearson 1900). For the latter case, we assume that the dichotomous variables smoking (1 =no, 2 =yes) and alcohol drinking (1 =no, 2 =yes) are each discrete categorizations of continuous unobservable random variables. These latent variables follow from a bivariate standard normal (Φ_2) distribution with correlation ρ, and for each variable there is a single threshold that partitions the distribution. The two cutpoints ϕ_1 and ϕ_2 give rise to four quadrants (Figure 7.4). The percentages in the four cells then correspond to the probabilities of the four quadrants under Φ_2, and an estimate of the thresholds is obtained by equating the observed proportions to the theoretical probabilities. The correlation coefficient, called the tetrachoric correlation, is estimated from the thresholds and a series expansion.

If the original classification of drinking and smoking is used, then a 4×3-contingency table arises. Similar to the above, we can assume a bivariate normal distribution with correlation ρ. However, there are now three cutpoints in the 'drinking' latent variable and two cutpoints in the 'smoking' variable (see Figure 7.5). The underlying correlation is now called the polychoric correlation. For computational reasons considering a 10% random sample, we obtained a polychoric correlation of 0.25, which is highly

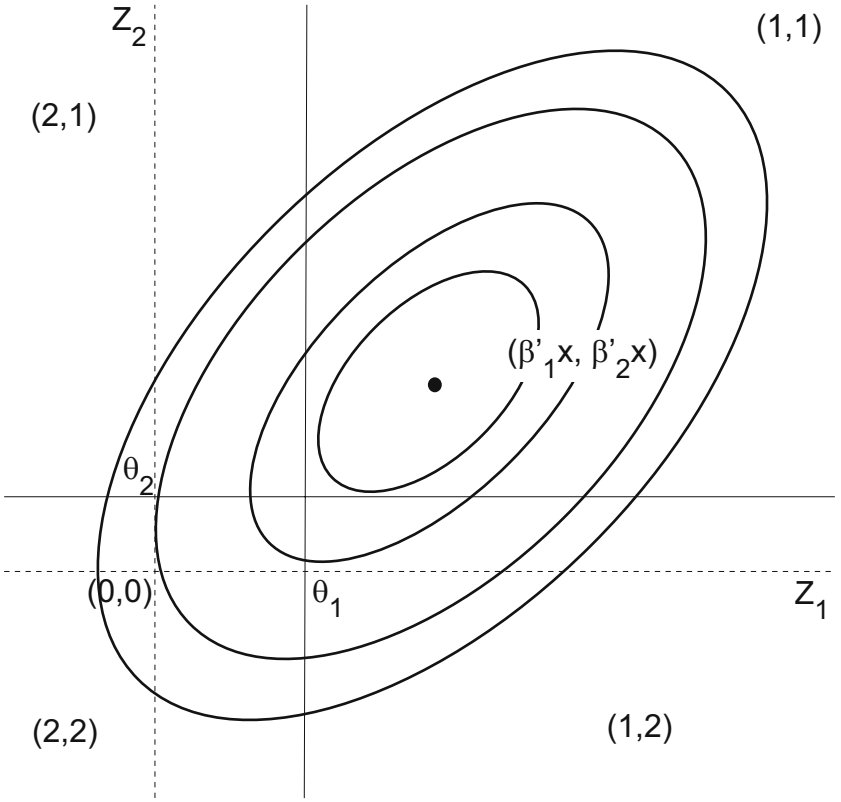

FIGURE 7.4. *Two-dimensional latent space, thresholds θ_1 and θ_2. The bivariate normal density with mean $(\boldsymbol{\beta}_1'\boldsymbol{x}, \boldsymbol{\beta}_2'\boldsymbol{x})'$ and correlation ρ is indicated by the elliptical contours.*

significant ($p < 0.0001$); the p-value is obtained from a Wald test for no correlation.

7.6.3 The Univariate Probit Model

To investigate the relationship between alcohol drinking (yes/no) and the explanatory variables, we would normally use the logistic model. Alternatively, the univariate probit model can be employed. Specifically this model states that the probability of alcohol drinking equals $\Phi(\boldsymbol{\beta}_1^{'}x)$, where $\boldsymbol{x} = (1, x_1, \ldots, x_{p-1})^{'}$ is the vector of covariates and $\boldsymbol{\beta}_1$ the vector of unknown regression parameters. Although not necessary, this model can be justified by the existence of an unobservable latent variable that has a normal distribution with a mean dependent on the covariates. A similar model can be proposed to model smoking behavior. Furthermore, the univariate

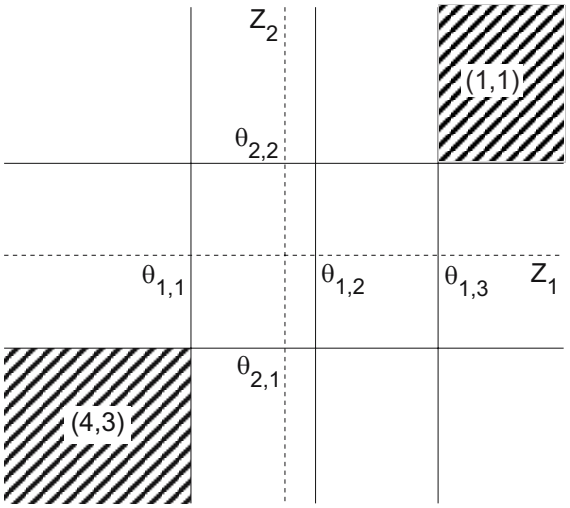

FIGURE 7.5. *Integration areas for a* 4×3 *BPM. The areas for the response combinations* $(1, 1)$ *and* $(4, 3)$ *are shaded.*

probit model (and the logistic model) can be extended to handle discrete, ordinal response variables.

However, the weakness of this approach lies in the fact that the two response variables are modeled separately, thereby neglecting their association. This will result in less efficient estimates of the parameters even though they are consistent. More importantly, we would obtain severely distorted estimates of the probabilities of combined responses, the so-called joint or union probabilities. This will be illustrated further in Section 7.10.

For the BIRNH study, parameter estimates (standard errors) are presented in Table 7.7. We selected the important risk factors using forward selection based on the score statistic (but the selected models based on the log-likelihood ratio criterion were identical). In Table 7.6, we show the two estimated univariate probit models, next to the bivariate probit model, based on the 10% random sample. The intercepts are the threshold values that determine the classes of the ordinal response variables. The interpretation of these models poses no difficulties, for example, both univariate analyses indicate that women drink and smoke less than men; from the first model we infer that, on average, Flemish people consume less alcohol than elsewhere in the country, and so on.

7.6.4 The Bivariate Probit Model

If there is heterogeneity in the study population, then a single two-by-two or $I \times J$ contingency table, of the type described in Chapter 6, will give a distorted picture of the real association between the two behaviors. The

TABLE 7.6. *BIRNH Study. Univariate and bivariate probit analysis on a 10% random sample of the original set of data.*

Effect	Estimate (s.e.)	
	Univariate	Bivariate
	Alcohol	
intercept 1	-1.07 (0.14)	-1.04 (0.14)
intercept 2	-0.69 (0.14)	-0.66 (0.14)
intercept 3	0.07 (0.14)	0.09 (0.14)
sex	0.70 (0.08)	0.69 (0.08)
social 1	-0.29 (0.07)	-0.31 (0.07)
site	0.21 (0.07)	0.21 (0.07)
	Smoking	
intercept 1	-3.77 (0.35)	-3.75 (0.35)
intercept 2	-3.18 (0.34)	-3.16 (0.34)
sex	-1.15 (0.09)	-1.15 (0.09)
BMI 2	0.05 (0.01)	0.05 (0.01)
age($\times$10)	0.12(0.03)	0.11 (0.03)
social 1	0.23 (0.10)	0.21 (0.10)
social 2	0.25 (0.09)	0.24 (0.09)
	Correlation coefficient	
intercept		0.41 (0.13)
sex		-0.30 (0.09)
social 2		0.17 (0.10)
log-likelihood	-2287.06	-2281.01

reason is that part of the association can be "explained" by the confounding effect of the (un)measured variables causing the heterogeneity. The BPM takes account of this effect while calculating the tetrachoric correlation.

In Section 7.3, the multivariate probit model was presented as one member of a general class. Here, we will provide more insight into the genesis of this particular model by first focusing on the bivariate case and then consider the specific approach of an underlying (bivariate) continuous density.

Suppose that there is an underlying but unobservable latent variable $W_s(\equiv W_1)$ that expresses the resistance of an individual to smoking, and further suppose that the individual will smoke if W_s is less than a threshold θ_1. Similarly, we assume that there is a $W_a(\equiv W_2)$ that reflects an individual's attitude toward alcohol consumption and that the individual will be a drinker if W_2 is less than θ_2. We assume that $\boldsymbol{W} = (W_1, W_2)'$ has a bivariate normal density with mean vector $\boldsymbol{\mu} = (\mu_1, \mu_2)'$ and with correlation ρ. Further, assume that each subject has a p-dimensional vector

of explanatory variables, $\boldsymbol{x} = (x_0, x_1, \ldots, x_{p-1})'$ with $x_0 \equiv 1$, which has the following effect on the mean vector:

$$\mu_j = \boldsymbol{\beta}_j' \boldsymbol{x}, \quad (j = 1, 2).$$

Thus, by contrast to Section 7.6.2, we now assume that the distribution depends on $\boldsymbol{x}$ and that each individual with covariate vector $\boldsymbol{x}$ is supposed to have a latent bivariate normal attitude distribution with mean vector $(\boldsymbol{\beta}_1' \boldsymbol{x}, \boldsymbol{\beta}_2' \boldsymbol{x})'$ and correlation ρ. In other words, the covariates move the mean vector of the two-dimensional Normal density over the plane. This results in the BPM first suggested by Ashford and Sowden (1970).

The cell probabilities for the fourfold table are again given by the probability of a quadrant under a suitable normal distribution. In Figure 7.4, we show the quadrants corresponding to the four cells of the two-by-two contingency table. The probability that a particular combination occurs is then obtained from the volume under the density surface taken in by the corresponding quadrant. For example, the probability of cell $(2, 2)$ in the fourfold table for an individual with covariate vector $\boldsymbol{x}$ is equal to the volume under the normal density $N(\boldsymbol{\beta}_1' \boldsymbol{x}, \boldsymbol{\beta}_2' \boldsymbol{x}; \rho)$ for the quadrant $]-\infty, \theta_1[\times]-\infty, \theta_2[$. The quadrant probabilities are also the class of posterior probabilities for each individual once the vector of covariates $\boldsymbol{x}$, is known. Let $\boldsymbol{Y}$ be the two-dimensional vector with first component $Y_1 = 1$ or 2 corresponding to 'non-drinker' or 'drinker,' respectively. The second component Y_2 is defined similarly with respect to smoking. Let $\boldsymbol{y}$ denote the observed values. The class H_y will then contain all cases with the combination of thetwo response classes corresponding to $\boldsymbol{y}$. We will use

$$\mu_{k_1 k_2}(\boldsymbol{\beta}; \rho | \boldsymbol{x}) = P(Y_1 = k_1, Y_2 = k_2 | \boldsymbol{\beta}, \rho, \boldsymbol{x})$$

to denote the posterior probability of H_y, conditional on $\boldsymbol{x}$. Formally, the BPM assumes

$$\begin{aligned}
\mu_{11}(\boldsymbol{\beta}; \rho | \boldsymbol{x}) &= \Phi_2(\boldsymbol{\beta}_1' \boldsymbol{x}, \boldsymbol{\beta}_2' \boldsymbol{x}; \rho), \\
\mu_{12}(\boldsymbol{\beta}; \rho | \boldsymbol{x}) &= \Phi(\boldsymbol{\beta}_1' \boldsymbol{x}) - p_{11}(\boldsymbol{\beta}; \rho | \boldsymbol{x}), \\
\mu_{21}(\boldsymbol{\beta}; \rho | \boldsymbol{x}) &= \Phi(\boldsymbol{\beta}_2' \boldsymbol{x}) - p_{11}(\boldsymbol{\beta}; \rho | \boldsymbol{x}), \\
\mu_{22}(\boldsymbol{\beta}; \rho | \boldsymbol{x}) &= 1 - \mu_{12}(\boldsymbol{\beta}; \rho | \boldsymbol{x}) - \mu_{21}(\boldsymbol{\beta}; \rho | \boldsymbol{x}) - \mu_{11}(\boldsymbol{\beta}; \rho | \boldsymbol{x}),
\end{aligned} \tag{7.28}$$

with $\beta_{j0} = \theta_j - \alpha_{j0} (j = 1, 2)$, and $\beta_{js} = -\alpha_{js} (s = 1, \ldots, p; j = 1, 2)$, where $\Phi(a)$ is the standard normal distribution in a and $\Phi_2(a_1, a_2)$ the standard bivariate normal distribution with mean 0 and correlation ρ. Morimune (1979) extended this model by letting ρ depend on $\boldsymbol{x}$, so that $\rho = \rho(\boldsymbol{\alpha}' \boldsymbol{x})$. An immediate generalization of model (7.28) is obtained by allowing more than one cutpoint for each latent variable W_j $(j = 1, 2)$. This corresponds to the analysis of $r_1 \times r_2$ contingency tables. In Figure 7.5, the integration areas are shown for a 4×3 table.

For the binary response model (7.28) we get the marginal probabilities

$$\begin{aligned} \mu_{1+}(\boldsymbol{\beta}_1|\boldsymbol{x}) &= \Phi(\boldsymbol{\beta}_1^{'}\boldsymbol{x}), \\ \mu_{+1}(\boldsymbol{\beta}_2|\boldsymbol{x}) &= \Phi(\boldsymbol{\beta}_2^{'}\boldsymbol{x}), \end{aligned} \tag{7.29}$$

where the first corresponds to the probability of alcohol drinking for a specific combination of the covariates and the second to the probability of smoking. Observe that these probabilities are identical to those under a univariate probit model. However, with two univariate probit models, the joint probabilities are obtained by simple multiplication of the marginal probabilities, for example the probability of alcohol drinking and smoking is calculated as

$$\Phi(\boldsymbol{\beta}_1^{'}\boldsymbol{x}) \cdot \Phi(\boldsymbol{\beta}_2^{'}\boldsymbol{x}),$$

which corresponds to $\mu_{22}(\boldsymbol{\beta}; \rho|\boldsymbol{x})$ under the BPM only if $\rho = 0$. Thus, by employing two univariate probit models for the analysis of correlated binary response variables, we explicitly assume that $\rho = 0$ in a BPM. Clearly, the same reasoning applies to discrete ordinal responses.

To conclude the model specification, we suppose that there are N independent subsamples, where the rth subsample is characterized by the covariate vector $\boldsymbol{x}_r$. Within the rth subsample, we have N_r independent observations. The corresponding counts are Z_{ry}, the number of occurrences of response $\boldsymbol{y}$ in the rth subsample. If $S_j = \{1, 2\}$ denotes the set of levels of the jth characteristic in the binary case, then $S = S_1 \times S_2$ contains all possible combinations of characteristics. Given $\boldsymbol{x}_r$, the counts

$$(Z_{ry}, \boldsymbol{y} \in S)$$

are multinomially distributed with N_r replicates and probability vector

$$(\mu_{ry} = p_y(\boldsymbol{\beta}; \rho|\boldsymbol{x}_r), \boldsymbol{y} \in S)\,. \tag{7.30}$$

To estimate the unknown parameters $\boldsymbol{\beta}$ and ρ, the likelihood of the sample under the model is needed and is given by

$$\ell(\boldsymbol{\beta}, \rho) = \sum_{r=1}^{N} \sum_{y \in S} z_{ry} \ln \mu_{ry}. \tag{7.31}$$

A maximum likelihood estimate of $(\boldsymbol{\beta}^{'}, \rho)$, denoted by $(\widehat{\boldsymbol{\beta}}^{'}, \hat{\rho})$, is obtained by maximizing (7.31) with respect to the unknown parameters. The negative inverse of the second derivative of the log-likelihood provides the estimated covariance matrix of the parameters.

For the BIRNH study, a bivariate selection procedure, based on the score statistic, selected the variables: 'sex' ($p < 0.0001$); 'BMI' ($p < 0.0001$); 'age' ($p = 0.0003$); 'social 1' ($p = 0.0033$); 'site' ($p = 0.0071$) and 'social 2' ($p =$

0.0197). Taking these covariates into account, the polychoric correlation coefficient dropped from 0.25 to 0.059 ($p = 0.16$). Thus, it seems that all correlation between drinking and smoking was induced by the confounding effect of the demographic variables.

The score statistic to test the hypothesis of a constant correlation coefficient (that is whether or not Morimune's extension is needed) equals 13.56, which referred to a chi-squared distribution with 4 degrees of freedom, indicating dependence of the correlation on the predictors ($p = 0.035$). Based on the significance of the regression coefficients, both 'sex' ($p = 0.001$) and 'social 2' ($p = 0.048$) seem to have an impact on the polychoric correlation. Thus in the next step, besides the constant, 'sex' and 'social 2' were also included in the model for ρ; the regression coefficients are 0.45 ($p = 0.0004$) for the constant, -0.33 ($p = 0.0004$) for 'sex' and 0.18 ($p = 0.068$) for 'social 2.'

Up to now, the same covariate vector has been employed for both responses. This is not necessary and in a further step we retained only the significant ($p < 0.05$) covariates in modelling the marginal dependencies. As can be seen from Table 7.6, the bivariate probit regression coefficients are very close to those obtained from the univariate probit regressions. This model, applied to the full dataset, gave similar regression coefficients that are not reported here.

For 14 of the 16 combinations of 'sex,' 'social 1,' 'social 2,' and 'site' it was possible to calculate the polychoric correlation locally with only 'age' and 'BMI' as predictors. There is reasonable agreement between global and local estimates, except for the two outlying correlations in the ('sex=female,' 'social 2=1') combination, but these were based on relatively small numbers, 168 and 229 cases, respectively.

Thus, the BPM indicates the same dependence of the responses on the demographic variables as the two univariate probit models, but it has provided extra information about the relationship between alcohol drinking and smoking. We conclude this analysis by observing that the BPM has nicely discerned the predictors affecting the marginal risk of alcohol drinking and smoking from those which affect the relationship between these two habits.

7.6.5 Ordered Categorical Outcomes

As stated before, the probit models can be generalized from binary outcome variables to ordered categorical outcomes. In this case, $\boldsymbol{Y} = (Y_1, Y_2)'$ is a bivariate stochastic vector of discrete ordered variables. Without loss of generality, assume that $Y_j \in S_j \equiv \{1, \ldots, c_j\}$, $(j = 1, 2)$.

Again, we assume that $\boldsymbol{Y}$ is a discretized version of an unobservable latent stochastic vector $\boldsymbol{W} = (W_1, W_2)'$ with bivariate normal cumulative distribution function having mean vector $\boldsymbol{\mu} = (\mu_1, \mu_2)'$ standard deviations

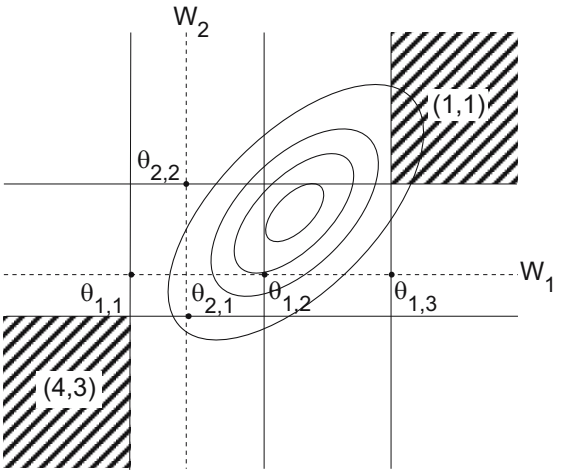

FIGURE 7.6. *Graphical representation of assumed underlying latent space of BPM. The areas for the response combinations (1,1) and (3,4) are shaded. The contours correspond to the surfaces of equal density of the bivariate normal density.*

$\sigma_1 = \sigma_2 = 1$ and correlation coefficient ρ. Then, $c_j - 1$ finite thresholds

$$-\infty \equiv \theta_{j0} < \theta_{j1} < \ldots < \theta_{j,c_j-1} < \theta_{jc_j} \equiv +\infty, \quad (j = 1, 2), \tag{7.32}$$

result in the vector $\boldsymbol{Y}$ by defining

$$Y_j = k \Longleftrightarrow \theta_{j,k-1} \leq W_j < \theta_{jk},$$

with $k \in S_j$. The $\boldsymbol{W}$-space, the bivariate normal density and its associated subdivision are graphically depicted in Figure 7.6, for $c_1 = 4$ and $c_2 = 3$. The association between Y_1 and Y_2 is expressed as the correlation between the latent variables W_1 and W_2; ρ is called the polychoric correlation (Pearson 1900).

Again, the model description is complete if we specify the link function between $\boldsymbol{x}$ and $\boldsymbol{Y}$. The probability that $\boldsymbol{Y} = (k_1, k_2)^{'}$, given $\boldsymbol{x}$, is equal to the probability $p_{k_1k_2}(\boldsymbol{x})$ that $\boldsymbol{W}$ lies in the rectangle

$$R_{k_1k_2}(\boldsymbol{x}) = [\theta_{1,k_1-1}(\boldsymbol{x}), \theta_{1k_1}(\boldsymbol{x})] \times [\theta_{2,k_2-1}(\boldsymbol{x}), \theta_{2k_2}(\boldsymbol{x})],$$

where $\theta_{jk}(\boldsymbol{x}) = \theta_{jk} - \boldsymbol{\beta}_j^{'}\boldsymbol{x}$. Specifically, for a BPM where ρ does not depend on the covariates:

$$\mu_{k_1k_2}(\boldsymbol{x}) = P\left(Y_1 = k_1, Y_2 = k_2 | \boldsymbol{x}\right) = \iint_{R_{k_1k_2}(\boldsymbol{x})} \phi_2(\boldsymbol{w}, \rho) d\boldsymbol{w}, \tag{7.33}$$

where $\phi_2(\boldsymbol{w}, \rho)$ denotes the standard bivariate normal density with correlation ρ. If ρ depends on the covariates, it is given by

$$\rho = \rho(\boldsymbol{\alpha}'\boldsymbol{x}). \tag{7.34}$$

Often, ρ is replaced by Fisher's z transform, as in the second component of (7.12), which takes values in $\mathbb{R}$:

$$\varphi = \ln\left(\frac{1+\rho}{1-\rho}\right).$$

Using such a transformation avoids estimates to jump out of the interval $[-1, +1]$ and is especially useful when covariates are allowed, as in (7.34).

7.6.6 The Multivariate Probit Model

When the latent vector $\boldsymbol{W}$ has an n-dimensional normal distribution, that is when there are n characteristics or repeated measures, and the probability of each diagnostic class conditional on a risk vector $\boldsymbol{x}$ is again an integral over an orthant, the n-dimensional generalization of the quadrant, as in (7.28), we apply a MPM. As for the BPM the n-dimensional response vector can also consist of ordinal discrete responses with integration areas as in Figures 7.5 and 7.6. Anderson and Pemberton (1985) employed a trivariate probit model for the analysis of data on blackbirds. They fitted the model using by fitting the univariate margins independently, supplemented with the correlation parameters assembled from fitting bivariate probit models to all pairs of outcomes. Here, a fully general approach will be presented, but the approximate solution can be a viable option when computations become too cumbersome, e.g., when dimensionality is high.

Thus, a MPM of dimension n actually consists of n marginal probability distributions each corresponding to a particular characteristic and $n(n-1)/2$ polychoric correlations expressing the association between the occurences of the n characteristics. If the correlations equal zero then the marginal probability distributions are sufficient to generate the probabilities of all combinations of characteristics, if not, then the multivariate probability distributions are needed.

In analogy with the bivariate case, we suppose that there is a sample of N independent subsamples available, where the rth subsample is characterized by the covariate vector $\boldsymbol{x}_r$. The observed response vector is denoted by

$$\boldsymbol{y} = (y_1, \ldots, y_n)' \in \prod_{j=1}^{n} S_j.$$

Within the rth subsample, we have N_r independent replications. The number of occurrences of response $\boldsymbol{y}$ in the rth subsample is denoted as $z_j\boldsymbol{y}$. Given $\boldsymbol{x}_r$, the counts

$$\left(z_r\boldsymbol{y}, \boldsymbol{y} \in \prod_j S_j\right)$$

are multinomially distributed with N_r replications and probability vector

$$\left(\mu_{ry}(\boldsymbol{\theta}) = \mu_y(\boldsymbol{\theta}|\boldsymbol{x}_r), \boldsymbol{y} \in \prod_j S_j\right),$$

where $\boldsymbol{\theta}$ is the total parameter vector containing both regression and association parameters. Finally, the log-likelihood of the sample under the specified model is given by

$$\ell(\boldsymbol{\theta}) = \sum_{j=1}^{N} \sum_{\boldsymbol{y} \in \prod_j S_j} z_{ry} \ln p_{ry}(\boldsymbol{\theta}). \tag{7.35}$$

The maximum likelihood estimate of $\boldsymbol{\theta}$, denoted by $\widehat{\boldsymbol{\theta}}$, is obtained by maximizing (7.35) with respect to the unknown parameters.

7.7 The Dale Model

7.7.1 Two Binary Responses

Suppose that for each of N subjects in a study a vector $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2})'$ of two binary responses is observed, together with a vector of covariates $\boldsymbol{x}$. The vector $\boldsymbol{x}$ can be different for each response as in longitudinal studies with time-dependent covariates. Thus, the study subjects are described by $(y_{i1}, y_{i2}, \boldsymbol{x}_{ij}), (i = 1, \ldots, N; j = 1, 2)$. Just as with the bivariate probit model, we want to establish the dependence of each of the two responses on the covariate vector(s), taking the dependence between the responses into account.

Dale (1986) proposed a family of bivariate response models arising from the decomposition of the joint probabilities $\mu_{k_1 k_2}(\boldsymbol{x}) = P(Y_1 = k_1, Y_2 = k_2|\boldsymbol{x})$, $(k_1, k_2 = 1, 2)$, into 'main effects' and 'interactions.' The marginal probabilities describe the main effect and the log cross-ratio is the interaction term. Formally, this decomposition is given by

$$h_1\left(\mu_{1+}(\boldsymbol{x})\right) = \boldsymbol{\beta}_1' \boldsymbol{x}, \tag{7.36}$$

$$h_2\left(\mu_{+1}(\boldsymbol{x})\right) = \boldsymbol{\beta}_2' \boldsymbol{x}, \tag{7.37}$$

$$h_3\left(\frac{\mu_{11}(\boldsymbol{x})\mu_{22}(\boldsymbol{x})}{\mu_{12}(\boldsymbol{x})\mu_{21}(\boldsymbol{x})}\right) = \boldsymbol{\beta}_3' \boldsymbol{x}, \tag{7.38}$$

where h_1, h_2 and h_3 are link functions in the generalized linear model terminology and $\mu_{1+}(\boldsymbol{x})$, $\mu_{+1}(\boldsymbol{x})$ are the marginal probabilities for observing $Y_1 = 1$, and $Y_2 = 1$, respectively. The most popular choice for $h_1 \equiv h_2$ is the logit function, whereas for h_3 the natural logarithmic function is commonly used. In that case, one has two marginal logistic regression models

and the logarithm of the cross-ratio

$$\ln \psi(\boldsymbol{x}) = \ln \left(\frac{\mu_{11}(\boldsymbol{x})\mu_{22}(\boldsymbol{x})}{\mu_{12}(\boldsymbol{x})\mu_{21}(\boldsymbol{x})} \right) \tag{7.39}$$

is linear in the covariates. Note that (7.39) is in line with (7.21), for the specific situation of two binary outcomes.

The joint probabilities follow from the marginal probabilities in the following way, where we have omitted the dependence of the different terms on $\boldsymbol{x}$ for the ease of notation (Plackett 1965):

$$\mu_{11} = \begin{cases} \dfrac{1 + (\mu_{1+} + \mu_{+1})(\psi - 1) - S(\mu_{1+}, \mu_{+1}, \psi)}{2(\psi - 1)} & \text{if } \psi \neq 1, \\ \mu_{1+}\mu_{+1} & \text{if } \psi = 1, \end{cases} \tag{7.40}$$

and $\mu_{12} = \mu_{1+} - \mu_{11}$, $\mu_{21} = \mu_{+1} - \mu_{11}$, and $\mu_{22} = 1 - \mu_{12} - \mu_{21} - \mu_{11}$, with the function S defined by

$$S(q_1, q_2, \psi) = \sqrt{[1 + (q_1 + q_2)(\psi - 1)]^2 + 4\psi(1 - \psi)q_1 q_2},$$

for $0 \leq q_1, q_2 \leq 1$ and $0 \leq \psi < +\infty$.

Just as in the probit case, above description can also be seen as arising from the discrete realization of a continuous bivariate distribution, the Plackett distribution (Plackett 1965) in this case. Suppose the bivariate random vector $\boldsymbol{W} = (W_1, W_2)'$ has joint distribution function $F(w_1, w_2)$, with marginal distributions $F(w_j)$ $(j = 1, 2)$. Define the (global) cross-ratio function, or global odds ratio function, $\psi(w_1, w_2)$, by

$$\psi(w_1, w_2) = \frac{\mu_{11}\mu_{22}}{\mu_{12}\mu_{21}} = \frac{F(1 - F_1 - F_2 + F)}{(F_1 - F)(F_2 - F)}, \tag{7.41}$$

with $F_j \equiv F_j(w_j), (j = 1, 2)$ and $F \equiv F(w_1, w_2)$. It is clear that $\psi(w_1, w_2)$ satisfies $0 \leq \psi \leq \infty$. The components $\mu_{k_1 k_2}$ in (7.41) are the quadrant probabilities in $\mathbb{R}^2$ with vertex at (w_1, w_2). For a Plackett distribution, the global cross-ratio $\psi(w_1, w_2) \equiv \psi$ is constant. Expression (7.41) can be seen as a defining equation for F, once F_1, F_2, and ψ are known. The Plackett distribution then gives rise to the above bivariate response model if its mean vector $\boldsymbol{\mu} = (\mu_1, \mu_2)'$ depends linearly on the covariate vector and if it is assumed that $\boldsymbol{Z}$ is a discretized version of the continuous vector $\boldsymbol{W}$ in the sense that $Y_j = 1 \Longleftrightarrow \theta_j \leq W_j$, for $j = 1, 2$. Here, θ_1, θ_2 are two a priori defined thresholds. In other words, Dale's bivariate response model is obtained if the bivariate response vector $\boldsymbol{Y}$ is a discretized version of $\boldsymbol{W}$ using the threshold vector $\boldsymbol{\theta}$, and if the covariate vector shifts the mean vector of the distribution of $\boldsymbol{W}$ over the plane, thereby possibly changing also the association parameter ψ as a function of $\boldsymbol{x}$.

7.7.2 The Bivariate Dale Model

Dale (1986) generalized above approach to model pairs of ordered categorical variables with c_1 and c_2 levels, respectively, in the presence of explanatory variables $\boldsymbol{x}$. We will refer to this as the (bivariate) *global odds ratio model*, *global cross-ratio model*, or simply *bivariate Dale model* (BDM).

Let $\boldsymbol{Y} = (Y_1, Y_2)'$ be a random vector taking on values (k_1, k_2), where $1 \leq k_j \leq c_j$ $(j = 1, 2)$. The outcomes, corresponding to a given covariate vector $\boldsymbol{x}$, can be arranged as an $c_1 \times c_2$ contingency table $(Z_{k_1 k_2})$ $(k_j = 1, \ldots, c_j; j = 1, 2)$:

$$\begin{array}{|c|c|c||c|c|c|}
\hline
z_{11} & \cdots & z_{1k_2} & z_{1,k_2+1} & \cdots & z_{1c_2} \\
\hline
\vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
\hline
z_{k_1 1} & \cdots & z_{k_1 k_2} & z_{k_1,k_2+1} & \cdots & z_{k_1 c_2} \\
\hline\hline
z_{k_1+1,1} & \cdots & z_{k_1+1,k_2} & z_{k_1+1,k_2+1} & \cdots & z_{k_1+1,c_2} \\
\hline
\vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
\hline
z_{c_1 1} & \cdots & z_{c_1 k_2} & z_{c_1,k_2+1} & \cdots & z_{c_1 c_2} \\
\hline
\end{array}. \tag{7.42}$$

Similarly, the probabilities can be represented as a $c_1 \times c_2$ table:

$$\begin{array}{|c|c|c||c|c|c|}
\hline
\mu_{11} & \cdots & \mu_{1k_2} & \mu_{1,k_2+1} & \cdots & \mu_{1c_2} \\
\hline
\vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
\hline
\mu_{k_1 1} & \cdots & \mu_{k_1 k_2} & \mu_{k_1,k_2+1} & \cdots & \mu_{k_1 c_2} \\
\hline\hline
\mu_{k_1+1,1} & \cdots & \mu_{k_1+1,k_2} & \mu_{k_1+1,k_2+1} & \cdots & \mu_{k_1+1,c_2} \\
\hline
\vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
\hline
\mu_{c_1 1} & \cdots & \mu_{c_1 k_2} & \mu_{c_1,k_2+1} & \cdots & \mu_{c_1 c_2} \\
\hline
\end{array}. \tag{7.43}$$

This map establishes a lilnk between the regression and table notations (Section 7.1). Note that sparseness of these tables is not an issue, as the essence of the approach is truly of a regression type. When the number of subjects per covariate level $\boldsymbol{x}$ is small, the number of 'tables' increases with sample size, exactly as in a regression setting. However, when the number of covariate levels is small or even bounded (e.g., two sex levels), then the tables fill up, as in ANOVA and genuine contingency tables settings.

Dichotomizing contingency table (7.42) at (k_1, k_2) (double lines) leads to a 2×2 contingency table:

$$\begin{array}{|c|c|}
\hline
\{Y_1 \leq k_1, Y_2 \leq k_2\} & \{Y_1 \leq k_1, Y_2 > k_2\} \\
\hline
\{Y_1 > k_1, Y_2 \leq k_2\} & \{Y_1 > k_1, Y_2 > k_2\} \\
\hline
\end{array}, \tag{7.44}$$

of which the probabilities are given by

$$P_{11}(k_1, k_2, \boldsymbol{x}) \quad = \quad P(Y_1 \leq k_1, Y_2 \leq k_2 | \boldsymbol{x}),$$

$$\begin{aligned} P_{12}(k_1, k_2, \boldsymbol{x}) &= P(Y_1 \leq k_1, Y_2 > k_2 | \boldsymbol{x}), \\ P_{21}(k_1, k_2, \boldsymbol{x}) &= P(Y_1 > k_1, Y_2 \leq k_2 | \boldsymbol{x}), \\ P_{22}(k_1, k_2, \boldsymbol{x}) &= P(Y_1 > k_1, Y_2 > k_2 | \boldsymbol{x}). \end{aligned}$$

Marginal probabilities are obtained by summing over subscripts: $P_{1+}(k_1, \boldsymbol{x}) = P(Y_1 \leq k_1 | \boldsymbol{x})$ and $P_{+1}(k_2, \boldsymbol{x}) = P(Y_2 \leq k_2 | \boldsymbol{x})$.

In analogy with (7.36)–(7.38), the link functions are described by

$$h_1[P_{1+}(k_1, \boldsymbol{x})] = \beta_{0,1k_1} + \boldsymbol{\beta}_1^{'} \boldsymbol{x}, \quad (k_1 = 1, \ldots, c_1 - 1), \tag{7.45}$$

$$h_2[P_{+1}(k_2, \boldsymbol{x})] = \beta_{0,2k_2} + \boldsymbol{\beta}_2^{'} \boldsymbol{x}, \quad (k_2 = 1, \ldots, c_2 - 1), \tag{7.46}$$

$$h_3[\psi(k_1, k_2, \boldsymbol{x})] = \boldsymbol{\alpha}^{'} \boldsymbol{x}, \quad (k_j = 1, \ldots, c_j - 1; j = 1, 2), \tag{7.47}$$

where the global cross-ratio $\psi(k_1, k_2, \boldsymbol{x})$ is given by

$$\psi(k_1, k_2, \boldsymbol{x}) = \frac{P_{11}(k_1, k_2, \boldsymbol{x}) P_{22}(k_1, k_2, \boldsymbol{x})}{P_{12}(k_1, k_2, \boldsymbol{x}) P_{21}(k_1, k_2, \boldsymbol{x})}.$$

Note that for every contingency table (7.42) [or, equivalently, table of probabilities (7.43)], a set of $(c_1 - 1) \times (c_2 - 1)$ global cross-ratios is obtained:

ψ_{11}	$\cdots$	ψ_{1k_2}	ψ_{1,k_2+1}	$\cdots$	ψ_{1,c_2-1}
$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$\psi_{k_1 1}$	$\cdots$	$\psi_{k_1 k_2}$	ψ_{k_1,k_2+1}	$\cdots$	ψ_{k_1,c_2-1}
$\psi_{k_1+1,1}$	$\cdots$	ψ_{k_1+1,k_2}	ψ_{k_1+1,k_2+1}	$\cdots$	ψ_{k_1+1,c_2-1}
$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$\psi_{c_1-1,1}$	$\cdots$	ψ_{c_1-1,k_2}	ψ_{c_1-1,k_2+1}	$\cdots$	ψ_{c_1-1,c_2-1}

More complex choices for the linear predictors on the right hand side of (7.45)–(7.47) are possible. For instance, h_3 can incorporate terms depending on k_1 and k_2, representing row, column, and cell effects. In principle, extensions to non-linear predictors are possible too, although this would make the updating algorithms more cumbersome.

For every table (7.44), we assume that (7.41) holds with ψ replaced by $\psi(k_1, k_2, \boldsymbol{x})$, indicating that ψ is allowed to depend on the cutpoints and on the covariates. Further, $F(.|\boldsymbol{x}) \equiv F_{k_1 k_2}(.|\boldsymbol{x}) = P_{11}(k_1, k_2, \boldsymbol{x})$, and $F(.|\boldsymbol{x})$ can also be expressed in terms of the assumed underlying Plackett distribution: $F(.|\boldsymbol{x}) = P\left(W_1 \leq \theta_{1k_1}, W_2 \leq \theta_{2k_2} | \boldsymbol{x}\right)$. Observe that for each double dichotomy of the $c_1 \times c_2$ table, a different underlying Plackett distribution is assumed. When it can be assumed that $\psi(k_1, k_2, \boldsymbol{x}) \equiv \psi(\boldsymbol{x})$, for $k_j = 1, \ldots, c_t - 1$ $(j = 1, 2)$, there is a single underlying Plackett distribution, exactly as for the binary response model.

7.7.3 Some Properties of the Bivariate Dale Model

Dale's model has appealing properties. First, there is the flexibility with which the marginal structure is modeled, i.e., the cumulative marginal probabilities can be fitted in the generalized linear models framework. Second, the marginal parameters are orthogonal onto the association parameters in the sense that the corresponding elements in the expected covariance matrix are identically zero (Palmgren 1989). Further, the associations can be modeled in a flexible way including covariate, row, column, and cell specific terms (Dale 1986).

The BDM does not require marginal scores for the responses and is essentially invariant under any monotonic transformation of the marginal response variables. Further, if adjacent marginal categories are combined, the model for the new table has fewer parameters, but they have the same interpretation as they had in the model for the original, expanded table, because the parameters pertain to cutpoints between categories. This is in contrast with models based on local association (Goodman 1981a), as discussed in Chapter 6.

7.7.4 The Multivariate Plackett Distribution

The computational basis of the BDM is the Plackett distribution. Therefore, we first generalize the bivariate Plackett distribution to n dimensions. In this section, we present a general description and some properties. The multivariate Plackett distribution will be the basis for the multivariate Dale model. The genesis of the distribution will automatically lead to an algorithmic way to compute cell probabilities and their derivatives. This is an alternative to the iterative proportional fitting algorithm presented in Section 7.12.3. Other alternatives are given by Lang and Agresti (1994) and Glonek and McCullagh (1995). This rather technical development is deferred to Appendix 7.13.

7.7.5 The Multivariate Dale Model

Given the multivariate Plackett distribution, the multivariate Dale model is a straightforward extension of the BDM. Let $\boldsymbol{W}_{ri} = (W_{ri1}, \ldots, W_{rin})'$ have a multivariate Plackett distribution with univariate marginals F_j, $(r = 1, \ldots, N; i = 1, \ldots, N_r; j = 1, \ldots, n)$ and a particular set of generalized global cross-ratios. Further, let $\boldsymbol{Y}_{ri} = (Y_{ri1}, \ldots, Y_{rin})'$ be a vector of ordered categorical variables with Y_{rij} assuming values $k_j = 1, \ldots, c_j$, $(j = 1, \ldots, n)$. Thus, in analogy with the bivariate case, $\boldsymbol{Y}_{ri}$ is a discrete realization of $\boldsymbol{W}_{ri}$. The covariates at level r are indicated by $\boldsymbol{x}_r$. Both the marginal distributions and the cross-ratios can depend on the covariates.

For each multi-index $\boldsymbol{k} = (k_1, \ldots, k_n)$ with $1 \leq k_j < c_j$, $(j = 1, \ldots, n)$, define a 2^n-*dichotomization table (multiple dichotomy)*:

$$T_{\boldsymbol{k}} = \{\mathcal{O}_{\boldsymbol{s}}(\boldsymbol{k}) | \boldsymbol{s} \in \{-1, 1\}^n\},$$

where

$$\mathcal{O}_{\boldsymbol{s}}(\boldsymbol{k}) = \{\boldsymbol{Y}_{ri} | Y_{rij} \leq k_j \text{ if } s_j = -1 \text{ and } Y_{rij} > k_j \text{ if } s_j = 1\}.$$

This means that, at every n-dimensional cutpoint, the data table is collapsed into a $2 \times 2 \times \ldots \times 2$ table. Observe the analogy with the bivariate case, as well as with the probit case (Section 7.6). For $n = 2$, $T_{\boldsymbol{k}}$ contains the four corners of the $c_1 \times c_2$ contingency table, split up at $\boldsymbol{k} = (k_1, k_2)$.

Every table is assumed to arise as a discretization of a multivariate Plackett distribution. The n marginal distributions are modeled, together with all pairs of two-way cross-ratios. In addition, three-way up to n-way interactions (generalized cross-ratios) are included to fully specify the joint distribution. Formally, we assume that for each $T_{\boldsymbol{k}}$, (7.69) holds with a cross-ratio possibly depending on $\boldsymbol{k}$ and $\boldsymbol{x}_r$, i.e., $\psi_{1\ldots n}$ is replaced by $\psi(\boldsymbol{k}; \boldsymbol{x}_r)$. Further,

$$\begin{aligned} F \equiv F_{\boldsymbol{k}}(.|\boldsymbol{x}_r) &= P(Y_{ri1} \leq k_1, \ldots, Y_{rin} \leq k_n | \boldsymbol{x}) \\ &= P(W_{ri1} \leq \theta_{1k_1}, \ldots, W_{rin} \leq \theta_{nk_n} | \boldsymbol{x}_r). \end{aligned}$$

The model description is complete by specifying link functions and linear predictors for both the univariate marginals and the association parameters. If we assume a marginal proportional odds model, then the marginal links can be written as:

$$\begin{aligned} \eta_{rijk}(\boldsymbol{x}_r) &= h_j\left[P(Y_{rij} \leq k | \boldsymbol{x}_r)\right] = \beta_{0,jk} + \boldsymbol{\beta}_j^{'} \boldsymbol{x}_r, \qquad (7.48) \\ &\quad (1 \leq j \leq n, 1 \leq k < c_j). \end{aligned}$$

Expression (7.48) can be represented in terms of the latent variables as well:

$$h_j\left[P(W_{rij} \leq \theta_{jk} | \boldsymbol{x}_r)\right] = \beta_{0,jk} + \boldsymbol{\beta}_j^{'} \boldsymbol{x}_r, \quad (1 \leq j \leq n, 1 \leq k < c_j).$$

As in the bivariate case, common choices for the link functions h_j are the logit and the probit link.

The cross-ratios are usually log-linearly modeled. Covariate terms may be included, together with row, column, and cell-specific terms. A possible choice consists of complex models for the bivariate associations and simple ones for the higher order associations. For a fixed pair of variables (j_1, j_2), where $1 \leq j_1 < j_2 \leq n$, one can model the log cross-ratio as

$$\gamma_{j_1 j_2}^{k_1 k_2}(\boldsymbol{x}_r) = \ln \psi_{j_1 j_2}(k_1, k_2, \boldsymbol{x}_r) = \nu + \rho_{k_1} + \kappa_{k_2} + \tau_{k_1 k_2} + \boldsymbol{x}_r^{'} \boldsymbol{\beta}_{j_1 j_2}. \quad (7.49)$$

Here, ν is an intercept parameter, ρ_{k_1} $(k_1 = 1, \ldots, c_1 - 1)$ are row-specific parameters, κ_{k_2} $(k_2 = 1, \ldots, c_2 - 1)$ are column parameters, and $\tau_{k_1 k_2}$ $(k_1 =$

$1, \ldots, c_1 - 1; k_2 = 1, \ldots, c_2 - 1)$ are cell-specific parameters. Uniqueness constraints need to be imposed on the row, column, and cell parameters. For instance, $\rho_1 = 0$, $\kappa_1 = 0$, $\tau_{k_1 1} = 0, (k_1 = 1, \ldots, c_1 - 1)$, and $\tau_{1 k_2} = 0, (k_2 = 1, \ldots, c_2 - 1)$. The higher order associations usually are assumed to be constant. Parameter estimates are obtained using the maximum likelihood method.

As this model description yields the BDM for $n = 2$, it follows that the attractiveness and the flexibility of the original two-dimensional version is carried over on its n-dimensional version. However, not all properties of the BDM are inherited by the MDM. As mentioned above, Palmgren (1989) shows that the estimated marginal and association parameters are orthogonal. This result does not carry over onto the MDM, although Molenberghs and Lesaffre (1994) have shown it holds approximately for lower order associations, while it fully holds for the n-way association.

Having specified the model, the links and the linear predictors, the model parameters can be estimated by the ML estimation method. The use of the multivariate Plackett distribution makes it easy to compute both the joint probabilities and their derivatives. A Fisher scoring algorithm is a good choice, as it also provides us with the asymptotic expected covariance matrix for the model parameters.

The model formulated above still fits within the general log-contrasts of probabilities framework given by (7.17), as it should be, given the presentation here is merely a more elaborate introduction of the MDM, with an alternative way to compute the cell probabilities.

7.7.6 Maximum Likelihood Estimation

Section 7.4 sketched a general framework for maximum likelihood estimation, using the iterative proportional fitting algorithm. Here, we will specialize to the MDM, using the Plackett probability formulation. Essentially, for every individual or every covariate level, the kernel of a multinomial log-likelihood can be used, considering a highly structured n-way contingency table merely as a collection of multinomial cells.

Despite the fact that the Plackett distribution is only known implicitly, its values can be computed in an efficient way using numerical algorithms. Further, the derivatives of the Plackett cumulative distribution function can be evaluated in an analytical way, using implicit derivation. Based on these results, the score functions and the expected Fisher information matrix can be used to implement a convenient Fisher scoring algorithm. Details are presented in Appendix 7.14.

7.7.7 The BIRNH Study

In this section, we reconsider the BIRNH study, analyzed before in Section 7.6.1. We compare performance of the BPM, the bivariate Dale model

TABLE 7.7. *BIRNH Study. Parameter estimates (standard errors) for the bivariate models [BPM: bivariate probit model; BDM: bivariate Dale model with normal (N) or logistic (L) margins] with constant association parameter (the correlation coefficient for the BPM and the global cross-ratio for the BDM).*

Effect	BPM	BDM-N	BDM-L
	Alcohol		
Intercept 1	-1.07(0.14)	-1.07(0.14)	-1.69(0.23)
Intercept 2	-0.68(0.14)	-0.69(0.14)	-1.07(0.23)
Intercept 3	0.07(0.14)	0.07(0.14)	0.20(0.23)
Sex	0.70(0.08)	0.70(0.07)	1.11(0.12)
Social 1	-0.29(0.07)	-0.29(0.07)	-0.49(0.12)
Site	0.21(0.07)	0.21(0.07)	0.35(0.12)
	Smoking		
Intercept 1	-3.76(0.35)	-3.76(0.35)	-6.24(0.60)
Intercept 2	-3.17(0.34)	-3.18(0.34)	-5.25(0.59)
Sex	1.15(0.09)	1.16(0.09)	1.92(0.15)
BMI	0.04(0.01)	0.04(0.01)	0.07(0.02)
Age($\times$10)	0.12(0.03)	0.12(0.03)	0.20(0.06)
Social 1	0.22(0.10)	0.22(0.10)	0.36(0.16)
Social 2	0.25(0.08)	0.24(0.09)	0.41(0.15)
Association	0.06(0.04)	0.18(0.12)	0.18(0.11)
Log-likelihood	-2286.12	-2285.87	-2286.58

(BDM) with probit (N) and logistic (L) margins, in modeling the relationship between alcohol drinking and smoking habits on the one hand and certain demographic variables on the other hand.

Tables 7.7–7.9 present the estimates for several models. The BPM column in Table 7.8 coincides with the bivariate column in Table 7.6. The three models in Table 7.7 have a very comparable fit. When comparing the BPM in Table 7.7 with the univariate probit models in Table 7.6 using the likelihood ratio test statistics, we find $G^2 = 0.94$ ($p = 0.1703$). Thus, it would seem there is no need to account for the association. However, this was different when comparing both columns in Table 7.6. It illustrates the point that sometimes careful modeling of the association is necessary, in agreement with several analyses in Chapter 6. Table 7.8 presents the same three models, with the association now depending on the covariates 'sex' and 'social 2,' in line with Table 7.6. Also here, the three models have a comparable fit. Note that in Tables 7.7 and 7.8, the marginal regression parameters for the BPM and the BDM-N are virtually identical, which is to be expected as both models have probit margins. The parameters for the BDM-L are related with the others through the well-known factor $\pi/\sqrt{3}$

TABLE 7.8. *BIRNH Study. Parameter estimates (standard errors) for the bivariate models [BPM: bivariate probit model; BDM: bivariate Dale model with normal (N) or logistic (L) margins] with association depending on the covariates (the correlation coefficient for the BPM and the global cross-ratio for the BDM).*

Effect	BPM	BDM-N	BDM-L
	Alcohol		
Intercept 1	-1.04(0.14)	-1.05(0.14)	-1.65(0.23)
Intercept 2	-0.66(0.14)	-0.67(0.14)	-1.03(0.23)
Intercept 3	0.09(0.14)	0.09(0.14)	0.23(0.23)
Sex	0.69(0.08)	0.70(0.08)	1.09(0.13)
Social 1	-0.31(0.07)	-0.31(0.07)	-0.53(0.12)
Site	0.21(0.07)	0.21(0.07)	0.35(0.12)
	Smoking		
Intercept 1	-3.75(0.35)	-3.75(0.35)	-6.21(0.59)
Intercept 2	-3.16(0.34)	-3.16(0.34)	-5.22(0.58)
Sex	1.15(0.09)	1.14(0.09)	1.91(0.15)
BMI	0.05(0.01)	0.05(0.01)	0.08(0.02)
Age($\times$10)	0.11(0.03)	0.11(0.03)	0.19(0.06)
Social 1	0.21(0.10)	0.21(0.10)	0.34(0.16)
Social 2	0.24(0.09)	0.25(0.09)	0.41(0.15)
	Association parameters		
Constant	0.41(0.13)	1.15(0.36)	1.15(0.35)
Sex	-0.30(0.09)	-0.82(0.27)	-0.82(0.27)
Social 2	0.17(0.10)	0.46(0.28)	0.46(0.28)
Log-likelihood	-2281.01	-2280.90	-2281.64

(see also Section 3.4), the standard deviation of the logistic distribution. A similar phenomenon will be observed in Section 7.10. In the three models the association between alcohol and smoking is small but perhaps a bit higher for the Dale models. The association parameters of the BDM-N and BDM-L are similar, as both are framed in terms of odds ratios, in contrast to the correlation-based association in the BPM. The coefficients of the association measures for the variable dependence models are more difficult to compare because of the different reparameterizations used. For the BPM, the Fisher z transform of the correlation ρ depends linearly on the covariates, while for BDM $\log \psi$ depends linearly on $\boldsymbol{x}$. Nevertheless, from the log-likelihoods it is apparent that again the three models explain the data in virtually the same manner. Table 7.9 further includes row and column effects in the association structure of the BDM models. However, this does not significantly improve the fit of the model.

TABLE 7.9. *BIRNH Study. Parameter estimates (standard errors) for the bivariate Dale model (BDM) with normal (N) and logistic (L) margins, where the association depends both on the covariates and on the cutpoints.*

Effect	BDM-N	BDM-L
	Alcohol	
Intercept 1	-1.05(0.14)	-1.66(0.24)
Intercept 2	-0.67(0.14)	-1.03(0.23)
Intercept 3	0.09(0.14)	0.23(0.23)
Sex	0.70(0.08)	1.10(0.13)
Social 1	-0.31(0.07)	-0.52(0.12)
Site	0.21(0.07)	0.35(0.12)
	Smoking	
Intercept 1	-3.70(0.35)	-6.13(0.60)
Intercept 2	-3.12(0.34)	-5.14(0.59)
Sex	1.13(0.09)	1.88(0.15)
BMI	0.05(0.01)	0.08(0.02)
Age($\times$10)	0.11(0.03)	0.18(0.06)
Social 1	0.20(0.10)	0.33(0.16)
Social 2	0.24(0.09)	0.41(0.15)
	Association parameters	
Intercept	1.15(0.36)	1.13(0.36)
Sex	-0.72(0.28)	-0.72(0.28)
Social 2	0.45(0.28)	0.45(0.28)
Row 1	-0.19(0.18)	-0.19(0.18)
Row 2	-0.03(0.16)	-0.02(0.16)
Column 1	-0.03(0.11)	-0.03(0.12)
Log-likelihood	-2279.47	-2280.19

7.8 Hybrid Marginal-conditional Specification

The fully specified models in most of this chapter are of a marginal nature. The previous chapter presented marginal models alongside conditionally specified ones, to make a number of points about the advantages of marginal models. Chapter 11 zooms in on conditionally specified models. In this section, we will present a hybrid model family, in the sense that it combines aspects of marginal and conditional models. Because the lower order moments, usually of principal scientific interest, are marginally specified, we have chosen to present it here, rather than in Part III.

Fitzmaurice and Laird (1993) model the marginal mean parameters, together with the canonical interaction parameters in the multivariate ex-

ponential family distribution of Cox (1972). Their model is related to the quadratic exponential model of Zhao and Prentice (1990). The distribution of Fitzmaurice and Laird (1993) differs from the previously described distributions because it is specified in terms of a mixture of marginal and conditional parameters.

Molenberghs and Ritter (1996) and Molenberghs and Danielson (1999) proposed a model that combines important advantages of a full marginal model and a mixed marginal-conditional model. The model is parameterized using marginal means, pairwise marginal odds ratios, and higher order conditional odds ratios. These conditional odds ratios are the canonical parameters of the exponential family described by Cox (1972) of which it is known that their interpretation is difficult, especially when the number of measurements per unit is variable. More details on the fully conditional model can be found in Section 11. The mixed parameterization has important advantages. First, it produces lower order parameter estimators that are robust against misspecification of the higher order structure. Second, the likelihood equations are less complex and easier to fit than the ones for the fully marginally specified models of Chapter 7. As such, a hybrid specification is an attractive alternative specification for a full likelihood method. However, one can set higher order association parameters equal to zero, whence they provide an appealing alternative to generalized estimating equations, in particular GEE2, as well (Section 8.7). This last observation was also employed by Heagerty and Zeger (1996), who consider a mixed marginal-conditional parameterization for clustered ordinal data, with the first and the second moments specified through marginal parameters, and who 8propose estimating the model parameters through GEE2, GEE1, or alternating logistic regressions.

7.8.1 A Mixed Marginal-conditional Model

We will use the regression notation. For each individual, subject, or experimental unit i in a study, a series of n categorical measurements Y_{ij}, grouped into a vector $\boldsymbol{Y}_i$ is recorded, together with covariate information $\boldsymbol{x}_i$. The parameters of primary interest are the first- and second-order marginal parameters. The covariate vector can include both time-dependent and time-stationary covariates. Covariate information can be used to model the marginal means, the associations, or both. In this section, we will restrict ourselves to binary outcomes. Section 7.8.2 considers the extension to categorical outcomes. The use of this modeling framework to derive GEE is discussed in Section 8.7 and exemplified in Sections 8.10 and 8.11.

Model building is based on the quadratic version of the joint distribution proposed by Cox (1972) and used by Zhao and Prentice (1990) and Fitzmaurice and Laird (1993). In particular, we write

$$f(\boldsymbol{y}_i|\boldsymbol{\Psi}_i, \boldsymbol{\Omega}_i) = \exp\left\{\boldsymbol{\Psi}_i'\boldsymbol{v}_i + \boldsymbol{\Omega}_i'\boldsymbol{w}_i - A(\boldsymbol{\Psi}_i, \boldsymbol{\Omega}_i)\right\}, \tag{7.50}$$

with outcomes and pairwise cross-products thereof grouped into

$$\boldsymbol{v}_i = (\boldsymbol{y}_i'; y_{i1}y_{i2}, \ldots, y_{i,n-1}y_{in})',$$

and third and higher order cross-products collected in

$$\boldsymbol{w}_i = (y_{i1}y_{i2}y_{i3}, \ldots, y_{i1}y_{i2}\ldots y_{in})',$$

and $\boldsymbol{\Psi}_i$ and $\boldsymbol{\Omega}_i$ the corresponding canonical parameter vectors. Further, let $\boldsymbol{\mu}_i = E(\boldsymbol{V}_i)$ and $\boldsymbol{\nu}_i = E(\boldsymbol{W}_i)$. The distribution is fully parameterized by modeling $\boldsymbol{\Psi}_i$ and $\boldsymbol{\Omega}_i$. However, we choose to model $\boldsymbol{\mu}_i$ and $\boldsymbol{\Omega}_i$, enabling us to describe the marginal means and the pairwise marginal odds ratios.

A model for $\boldsymbol{\mu}_i$ is specified via a vector of link functions

$$\boldsymbol{\eta}_i = \boldsymbol{\eta}_i(\boldsymbol{\mu}_i), \tag{7.51}$$

An important class of link functions, due to McCullagh and Nelder (1989), is given by (6.2). In particular, the marginal logit link and marginal log odds ratios can be used. The marginal part of the model formulation is complete by specifying the dependence on the covariates. From the covariate vector $\boldsymbol{x}_i$ a design matrix $\boldsymbol{X}_i$ is derived, such that $\boldsymbol{\eta}_i = \boldsymbol{X}_i\boldsymbol{\beta}$, with $\boldsymbol{\beta}$ a vector of parameters of interest.

Similarly, a model for the conditional higher order parameters needs to be constructed. In agreement with Fitzmaurice and Laird (1993), and because the components of $\boldsymbol{\Omega}_i$ can be interpreted as conditonal higher order log odds ratios, we assume an identity link and specify the covariate dependence as $\boldsymbol{\Omega}_i = \boldsymbol{X}_i'\boldsymbol{\alpha}$, with $\boldsymbol{X}_i'$ another design matrix and $\boldsymbol{\alpha}$ a parameter vector. A simple model is found by holding the components of $\boldsymbol{\Omega}_i$ constant.

In principle, $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ could be allowed to overlap, making the model slightly more general, but there would typically be little practical relevance to this.

Following derivations in Fitzmaurice and Laird (1993), Fitzmaurice, Laird, and Rotnitzky (1993), and Molenberghs and Ritter (1996), the likelihood equations can be written as:

$$\begin{aligned}\frac{\partial \ell}{\partial(\boldsymbol{\beta},\boldsymbol{\alpha})} &= \sum_{i=1}^{N} \begin{pmatrix} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} & 0 \\ 0 & \frac{\partial \boldsymbol{\Omega}_i}{\partial \boldsymbol{\alpha}} \end{pmatrix}' \begin{pmatrix} \boldsymbol{M}_i^{-1} & 0 \\ -\boldsymbol{N}_i\boldsymbol{M}_i^{-1} & I \end{pmatrix} \\ &\quad \times \begin{pmatrix} \boldsymbol{v}_i - \boldsymbol{\mu}_i \\ \boldsymbol{w}_i - \boldsymbol{\nu}_i \end{pmatrix}, \end{aligned} \tag{7.52}$$

with $\boldsymbol{M}_i = \mathrm{cov}(\boldsymbol{V}_i)$ and $\boldsymbol{N}_i = \mathrm{cov}(\boldsymbol{V}_i, \boldsymbol{W}_i)$.

The form of the derivatives in the first matrix of (7.52) depends on the choice of link functions and linear predictors. Under the assumed linear model for $\boldsymbol{\Omega}_i$, the derivative reduces to $\boldsymbol{X}_i'$. The computation of $\partial\boldsymbol{\mu}_i/\partial\boldsymbol{\beta}$ is

particularly straightforward for link functions of the form (6.2), in agreement with (7.61):

$$\left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}\right)' = \boldsymbol{X}_i'(\boldsymbol{D}_i')^{-1}$$

with

$$\boldsymbol{D}_i = \left(\frac{\partial \boldsymbol{\eta}_i}{\partial \boldsymbol{\mu}_i}\right) = \boldsymbol{C}_i \left\{\text{diag}(\boldsymbol{A}_i\boldsymbol{\mu}_i)\right\}^{-1} \boldsymbol{A}_i.$$

As the model is a mixed parameterization of an exponential family model (Barndorff-Nielsen 1978), the parameter vectors $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are orthogonal in the sense of Cox and Reid (1987). This implies that $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are asymptotically independent. Indeed, the inverse of the expected information matrix equals:

$$\begin{pmatrix} \Gamma_1^{-1} & 0 \\ 0 & \Gamma_2^{-1} \end{pmatrix}$$

with

$$\begin{aligned} \Gamma_1 &= \sum_{i=1}^{N} \boldsymbol{X}_i'(\boldsymbol{D}_i')^{-1}\boldsymbol{M}_i^{-1}\boldsymbol{D}_i^{-1}\boldsymbol{X}_i, \\ \Gamma_2 &= \sum_{i=1}^{N} (\boldsymbol{X}_i')'(\boldsymbol{P}_i - \boldsymbol{N}_i\boldsymbol{M}_i^{-1}\boldsymbol{N}_i')\boldsymbol{X}_i', \end{aligned}$$

and $\boldsymbol{P}_i = \text{cov}(\boldsymbol{W}_i)$.

Calculating the joint probabilities can be done in various ways. Fitzmaurice and Laird (1993) proposed the use of the iterative proportional fitting (IPF) algorithm to avoid the computation of $\boldsymbol{\Psi}_i$. We will proceed similarly. First, the components of $\boldsymbol{\mu}_i$ are computed. Let us focus on logit and log odds ratio links. Inverting the logit links, like in (7.18) yields μ_{ij} $(j = 1, \ldots, n)$. Given μ_{ij_1}, μ_{ij_2}, and $\psi_{ij_1j_2} = \exp(\eta_{ij_1j_2})$, $\mu_{i_1j_2}$ can be calculated using Plackett's expression (7.40). To obtain higher order probabilities, an initial contingency table is constructed satisfying the third- and higher order conditional odds ratio structure. Then, the set of $n(n-1)/2$ bivariate marginal probabilities is fitted iteratively. This is similar to but different from the IPF algorithm outlined in Section 7.12.3. Although in Section 7.12.3 the algorithm had to be adapted to a marginally specified model for ordinal data, we are faced here with a more conventional application, the higher-order model being specified conditionally and the outcomes of a binary type. The standard algorithm is described in Agresti (2002).

Parameter estimation can be performed using a standard Fisher scoring iteration procedure. The inverse of the Fisher information, with the parameter estimates substituted, provides a variance estimator for $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}})$. As pointed out in Fitzmaurice, Laird, and Rotnitzky (1993), the consistency of the estimator for $\boldsymbol{\beta}$ only depends on the correct specification of the marginal part of the model, and not on $\boldsymbol{\alpha}$. If the $\boldsymbol{\alpha}$ part is misspecified, the

model based variance will be inconsistent, so the empirically corrected or 'robust' variance should be used. Apart from inferential advantages (Fitzmaurice, Laird, and Rotnitzky 1993), there are also computational advantages in terms of stability (Cox and Reid 1987). These points are taken up in Section 8.7.

As an alternative to the use of the robust variance estimator, a model checking procedure can be performed to assess whether the model specification is acceptable. If not, the model for the higher order associations can be made more complex in order to improve the fit. When there are only a few categorical covariate levels and the sample size within each level is sufficiently large, a classical model checking procedure such as the Pearson X^2 or the deviance G^2 test can be used (Agresti 1990).

7.8.2 Categorical Outcomes

Like the multivariate probit (Section 7.6) and Dale (Section 7.7) models, the hybrid model can accommodate categorical outcomes just as easily as dichotomous ones.

Let Y_{ij} again be a categorical outcome with c_j (possibly ordered) categories and use the dummy variables formally defined by (7.1) and (7.2). In particular, because we are not making use of the design level indicator r, Z^*_{ijk} indicates outcomes and Z_{ijk} cumulative outcomes. These indicator variables are again grouped into vectors $\boldsymbol{Z}^*_i$ and $\boldsymbol{Z}_i$. The corresponding sets of univariate and pairwise probabilities are

$$\begin{aligned} \mu^*_{ijk} &= \P(Y_{ij} = k|\boldsymbol{X}_i, \boldsymbol{\beta}) = P(Z^*_{ijk} = 1|\boldsymbol{X}_i, \boldsymbol{\beta}), \\ \mu^*_{i,jh,k\ell} &= P(Y_{ij} = k, Y_{ih} = \ell|\boldsymbol{X}_i, \boldsymbol{\beta}) = P(Z^*_{ijk} = 1, Z^*_{ih\ell} = 1|\boldsymbol{X}_i, \boldsymbol{\beta}). \end{aligned}$$

The cumulative probabilities are

$$\begin{aligned} \mu_{ijk} &= P(Y_{ij} \leq k|\boldsymbol{X}_i, \boldsymbol{\beta}), \\ \mu_{i,jh,k\ell} &= P(Y_{ij} \leq k, Y_{ih} \leq \ell|\boldsymbol{X}_i, \boldsymbol{\beta}) \end{aligned}$$

which are grouped in $\boldsymbol{\mu}^*_i$ and $\boldsymbol{\mu}_i$, respectively. The higher order probabilities $\boldsymbol{\nu}^*_i$ and $\boldsymbol{\nu}_i$ are defined similarly. Exponential models, similar to (7.50), are

$$f(\boldsymbol{y}_i|\boldsymbol{\Psi}^*_i, \boldsymbol{\Omega}^*_i) = \exp\left\{(\boldsymbol{\Psi}^*_i)'\boldsymbol{v}^*_i + (\boldsymbol{\Omega}^*_i)'\boldsymbol{w}^*_i - A^*(\boldsymbol{\Psi}^*_i, \boldsymbol{\Omega}^*_i)\right\}, \tag{7.53}$$

and

$$f(\boldsymbol{y}_i|\boldsymbol{\Psi}_i, \boldsymbol{\Omega}_i) = \exp\left\{\boldsymbol{\Psi}_i\boldsymbol{v}_i + \boldsymbol{\Omega}_i\boldsymbol{w}_i - A(\boldsymbol{\Psi}_i, \boldsymbol{\Omega}_i)\right\}, \tag{7.54}$$

where $\boldsymbol{V}^*_i$ contains the components of $\boldsymbol{Z}^*_i$ and the pairwise cross-products thereof, and $\boldsymbol{W}^*_i$ contains all higher order cross-products. The vectors $\boldsymbol{V}_i$ and $\boldsymbol{W}_i$ are defined similarly. Observe that (7.53) and (7.54) are overparameterized, as sum constraints apply to (7.53) and the variable $Z_{ijc_j} = 1$

in (7.54), which necessitates the use of identifying restrictions. In the case of a single nominal variable, (7.53) is called the multigroup logistic model (Albert and Lesaffre 1986).

With nominal outcomes, the marginal mean functions $\boldsymbol{\mu}_i^*$ will be modeled, together with the higher order conditional association parameters $\boldsymbol{\Omega}_i^*$. A vector of link functions $\boldsymbol{\eta}_i^* = \boldsymbol{\eta}_i^*(\boldsymbol{\mu}_i^*)$ has to be chosen and form (6.2) provides a convenient subclass. Baseline category logits seem very natural, together with local odds ratios. If the outcomes are measured on an ordinal scale it is more convenient to model $\boldsymbol{\mu}_i$, rather than $\boldsymbol{\mu}_i^*$, i.e., link functions $\boldsymbol{\eta}_i = \boldsymbol{\eta}_i(\boldsymbol{\mu}_i)$ are chosen and model (7.54) can be used. Note that this description is equally compatible with (7.53), as $\boldsymbol{\mu}_i^* = \boldsymbol{B}_i\boldsymbol{\mu}_i$ for an appropriate transformation matrix $\boldsymbol{B}_i$, as in (7.4).

The likelihood equations are of the form (7.52). Even more than with binary outcomes, the number of parameters proliferates rapidly with an increasing number of measurements, calling for parsimonious modeling. A simple, but often satisfactory model for the pairwise association is the constant global odds ratio model for ordinal outcomes: the global odds ratio for a pair of variables $(Z_{ijk}, Z_{ih\ell})$ is independent of the 'row' and 'column' indices k and ℓ. Further, one should exploit any additional structure in the outcomes. For exchangeable outcomes, the odds ratios are usually assumed constant for all pairs of variables, whereas for time-ordered measurements, association structures taking into account the time dependence can be investigated. A similar reasoning could be made to simplify the higher order conditional associations. In many instances this effort will be considered of no real benefit, whence one can set $\boldsymbol{\Omega}_i = \mathbf{0}$. In order to compute the variance matrix $\boldsymbol{M}_i$, we only need to compute the third- and fourth-order probabilities, which is particularly easy using the iterative proportional fitting algorithm.

7.9 A Cross-over Trial: An Example in Primary Dysmenorrhoea

The data are taken from a cross-over trial that appeared in the paper of Kenward and Jones (1991). Eighty-six subjects were enrolled in a cross-over study that compared placebo (A) with an analgesic at low and high doses (B and C) for the relief of pain in primary dysmenorrhoea. The three treatments were administered in one of six possible orders: ABC, ACB, BAC, BCA, CAB, and CBA. The primary outcome score was the amount of relief coded as none (1), moderate (2), and complete (3). There are 27 possible outcome combinations: $(1,1,1), (1,1,2), \ldots, (3,3,3)$, where (a_1, a_2, a_3) denotes outcome a_j in period j. The data, analyzed before by Kenward and Jones (1991), can be found in Table 7.10. For the analysis of the cross-over data, these authors suggested a subject-specific approach

TABLE 7.10. *Primary Dysmenorrhoea Data.*

Response	ABC	ACB	BAC	BCA	CAB	CBA
(1,1,1)	0	2	0	0	3	1
(1,1,2)	1	0	0	1	0	0
(1,1,3)	1	0	1	0	0	0
(1,2,1)	2	0	0	0	0	0
(1,2,2)	3	0	1	0	0	0
(1,2,3)	4	3	1	0	2	0
(1,3,1)	0	0	1	1	0	0
(1,3,2)	0	2	0	0	0	0
(1,3,3)	2	4	1	0	0	1
(2,1,1)	0	1	1	0	0	3
(2,1,2)	0	0	2	0	1	1
(2,1,3)	0	0	1	0	0	0
(2,2,1)	1	0	0	6	1	1
(2,2,2)	0	2	1	0	0	0
(2,2,3)	1	0	0	0	0	0
(2,3,1)	0	0	0	1	0	2
(2,3,2)	0	0	0	0	0	0
(2,3,3)	0	2	0	0	1	0
(3,1,1)	0	0	0	1	0	2
(3,1,2)	0	0	2	0	2	1
(3,1,3)	0	0	3	0	4	1
(3,2,1)	0	0	0	1	0	0
(3,2,2)	0	0	0	1	0	0
(3,2,3)	0	0	0	0	0	0
(3,3,1)	0	0	0	0	0	1
(3,3,2)	0	0	0	0	0	0
(3,3,3)	0	0	0	0	0	0

based on the Rasch model. Here too, it was of interest to estimate the treatment, period- and carry-over effects.

7.9.1 Analyzing Cross-over Data

Consider a cross-over trial where each patient subsequently receives each of three treatments (A, B, C) in a random order. There are 6 treatment sequences: ABC, ACB, BAC, BCA, CAB, and CBA. Suppose the outcome at time j (corresponding to treatment t) is an ordered categorical variable Y_{jt} with c levels. Then, to each sequence a $c \times c \times c$ table is assigned,

containing the joint outcomes for the patients, allocated to that particular sequence. The multivariate Dale model can be used to fit such data. The marginal parameters are used to describe the overall treatment effects, the period- and the carry-over effects. The cross-ratios play a role, similar to the subject specific parameters in the paper of Kenward and Jones (1991).

Given a particular sequence s, let $L^s_{jtk} = \text{logit}\,[\text{P}(Y_{jt} \leq k)]$ be the cumulative logit for cutpoint k ($k = 1, \ldots, c-1$), and time j, which, for sequence s, corresponds to treatment t. In full detail, we have

$$\begin{array}{c} L^{ABC}_{11k}, L^{ABC}_{22k}, L^{ABC}_{33k}; \\ L^{ACB}_{11k}, L^{ACB}_{23k}, L^{ACB}_{32k}; \\ L^{BAC}_{12k}, L^{BAC}_{21k}, L^{BAC}_{33k}; \\ L^{BCA}_{12k}, L^{BCA}_{23k}, L^{BCA}_{31k}; \\ L^{CAB}_{13k}, L^{CAB}_{21k}, L^{CAB}_{32k}; \\ L^{CBA}_{13k}, L^{CBA}_{22k}, L^{CBA}_{31k}. \end{array}$$

The following model for the logits is adopted: $L^s_{jtk} = \mu_k + \tau_t + \rho_j + \lambda_{s(j-1)}$, where μ_k are intercept parameters, τ_t are treatment effects, ρ_j are period effects. $\lambda_{s(j-1)}$ stands for the carry-over effect, corresponding to the treatment at time $j-1$ in sequence s. Given, for instance, sequence CAB, we get

$$\begin{aligned} L_{13k} &= \mu_k + \tau_3 + \rho_1, \\ L_{21k} &= \mu_k + \tau_1 + \rho_2 + \lambda_3, \\ L_{32k} &= \mu_k + \tau_2 + \rho_3 + \lambda_1. \end{aligned}$$

To avoid overparameterization, the following uniqueness constraints are set:

$$\tau_1 = \rho_1 = \lambda_1 = 0.$$

Let $\gamma^s_{jt,j't'} = \ln \psi^s_{jt,j't'}$ be the log cross-ratio, for the marginal $c \times c$ table, formed by the responses at times j and j' for sequence s (corresponding to treatments t and t' respectively). The simplest model for the cross-ratios is given by

$$\gamma^s_{jt,j't'} = \mu.$$

The most complex model assumes all 18 cross-ratios to be different, which is one by Jones and Kenward (1989) and by Becker and Balagtas (1993). In between these two models there is room for parsimonious modeling. One can think of the following linear models in the log cross-ratios

$$\gamma^s_{jt,j't'} = \mu + \tau_{tt'}, \tag{7.55}$$

$$\gamma^s_{jt,j't'} = \mu + \rho_{jj'}, \tag{7.56}$$

$$\gamma^s_{jt,j't'} = \mu + \tau_{tt'} + \rho_{jj'}, \tag{7.57}$$

where μ is an intercept parameter, $\tau_{tt'}$ are parameters for the joint (t, t')th treatments effects, and $\rho_{jj'}$ describe effects for periods j and j'. In the first model, the log cross-ratio only depends on the treatments, irrespective of their order and the periods they were administered. In the second model, only the periods are of importance. In the third model, the two effects are linearly combined. For instance, for sequence CAB, we get

$$\begin{aligned}\gamma_{13,21} &= \mu + \tau_{13} + \rho_{12},\\ \gamma_{13,32} &= \mu + \tau_{23} + \rho_{13},\\ \gamma_{21,32} &= \mu + \tau_{12} + \rho_{23}.\end{aligned}$$

Possible uniqueness constraints are $\tau_{12} = \rho_{12} = 0$. The model with association structure (7.56) corresponds to the model introduced in Section 7.7.5. In a model with association structure (7.55) or (7.57), the two-way cross-ratios change with the treatment combination, which is a time-dependent covariate. Finally, the three-way association depends in all six cases on the same periods and treatments, the only difference being the order in which the treatments occur. So the most natural choice is $\gamma^s_{123} = \mu + \mu^s$, $(\mu^{ABC} = 0)$, however in most cases it is reasonable to assume $\gamma^s_{123} = \gamma_{123}$ constant over sequences.

No carry-over effects are incorporated in the cross-ratios, as the marginal carry-over parameters have no straightforward generalization. As usual, the different nested models can be tested using the likelihood ratio test statistic, denoted G^2.

7.9.2 Analysis of the Primary Dysmenorrhoea Data

Table 7.11 gives the details concerning the selection of effects for the primary dysmenorrhoea data. As can be seen from this table, the marginal logit modeling yields a highly significant treatment effect. The period and carry-over effects are not significant. The model retained (model I in Table 7.12), consists of two cutpoints μ_k and two treatment parameters τ_t; the estimates are shown in Table 7.12. Up to now no two-way or three-way association is assumed.

Let us turn to the association structure; the three-way association is assumed constant in all cases. First the minimal model is fitted. This model will serve as the basic model against which the other models will be compared. The three different models mentioned in (7.55), (7.56), and (7.57) were fitted to the data. There seems to be evidence that both the treatment terms as well as the period terms are necessary. The maximal model, i.e., with 18 cross-ratios, has a G^2 statistic of 16.27 (13 d.f., $p = 0.2349$) compared to the last model. Model II in Table 7.12 shows the parameter estimates when treatment parameters are included in the two-way cross-ratios. Model III, contains as association parameters: the intercept μ, treatment

TABLE 7.11. *Primary Dysmenorrhoea Data. Selection of effects. The columns describe the model number, the effects included, the log-likelihood of the model, the number of the model to which this model is compared, the likelihood ratio G^2 statistics with the number of degrees of freedom, and the corresponding p-value.*

	Effects	log-lik	Comp.	G^2	d.f.	p-value
		Marginal effects				
1	μ_k	-279.74				
2	μ_k, τ_t	-245.53	1	68.42	2	< 0.0001
3	μ_k, τ_t, ρ_j	-243.78	2	3.50	2	0.1740
4	$\mu_k, \tau_t, \lambda_{s(j-1)}$	-245.40	2	0.26	2	0.8790
		Model 2 + association effects				
5	$\mu_k, \tau_t; \mu, \psi_{123}$	-244.40				
6	$\mu_k, \tau_t; \mu, \tau_{tt'}, \psi_{123}$	-239.54	5	9.66	2	0.0080
7	$\mu_k, \tau_t; \mu, \rho_{jj'}, \psi_{123}$	-239.50	5	9.73	2	0.0077
8	$\mu_k, \tau_t; \mu, \tau_{tt'}, \rho_{jj'}, \psi_{123}$	-236.44	5	15.87	4	0.0032
			6	6.21	2	0.0448
			7	6.14	2	0.0465

effects $\tau_{tt'}$, period parameters $\rho_{jj'}$ and the three-way interaction $\ln \psi_{123}$. This model will be chosen.

Parameter interpretation is as follows. The odds of observing $Y_{jt} \leq k$ ($k = 1, 2$) decreases with factor $\exp(-1.98)$ when the patient is treated with the analgesic at low dose rather than with placebo. A further decrease with factor $\exp(-2.37 + 1.98)$ is observed if the patient is treated with the analgesic at high dose. Further, the association between responses is higher if they are close to each other in time ($\widehat{\rho}_{13} = -1.12$). Also, responses from the two analgesic treatments are more associated than responses from one analgesic treatment and placebo ($\widehat{\tau}_{23} = 1.32$).

Thus, our analysis confirms the results found by Kenward and Jones (1991). However, the marginal approach here allows the estimation of treatment effects that now are easily interpretable, in contrast with Kenward and Jones (1991) and in contrast with the conditional approach in Jones and Kenward (1989) as well. Confidence intervals for the effects can be found from the estimated standard errors, shown in Table 7.12 for Model III. Finally, the method allows flexible modeling of the association.

7.10 Multivariate Analysis of the POPS Data

The POPS data were introduced in Section 2.6. We will compare the Bahadur model (BAH), introduced in Section 7.2 and applied earlier to

TABLE 7.12. *Primary Dysmenorrhoea Data. Fitted models. Each entry represents the parameter estimates (standard error). The absence of a standard errors corresponds to a preset value.*

Effect	Par.	Model I	Model II	Model III
Marginal effects				
Intercept 1	μ_1	1.07(0.25)	1.07(0.24)	1.08(0.24)
Intercept 2	μ_2	2.71(0.29)	2.70(0.29)	2.72(0.29)
Treatment effect	τ_2	-2.03(0.33)	-2.02(0.35)	-1.98(0.34)
Treatment effect	τ_3	-2.41(0.33)	-2.37(0.36)	-2.37(0.35)
Two-way association effects				
Intercept	μ	0(-)	-0.62(0.47)	-0.46(0.56)
Treatment effect	τ_{13}	0(-)	-0.16(0.65)	-0.10(0.58)
Treatment effect	τ_{23}	0(-)	1.51(0.64)	1.32(0.61)
Period effect	ρ_{13}	0(-)	0(-)	-1.12(0.55)
Period effect	ρ_{23}	0(-)	0(-)	0.51(0.66)
Three-way association				
		1(-)	1.59(0.75)	0.63(0.88)
Log-likelihood				
		-245.53	-239.54	-236.43

the clustered NTP data and the longitudinal fluvoxamine study, with the trivariate probit model (TPM, Section 7.6) and the trivariate Dale model (TDM, Section 7.7), both with probit (normally based, N) and logistic (L) margins. Note that several comparisons are possible: (1) the Bahadur model and the TPM capture the association by means of correlations, whereas the TDM features odds ratios; (2) the Bahadur model and TDM-L have logistic margins, while the TPM and the TDM-N have univariate marginal regressions of a probit type. Finally, the log-likelihood at maximum, or the AIC can be used to compare the models with each other.

From the 8 candidate predictor variables, neonatal seizures, congenital malformations, and highest bilirubin value since birth were retained for analysis. They were selected using a stepwise logistic analysis for each response separately, at significance level 0.05. The first two regressors are dichotomous, the third one is continuous.

Table 7.13 contains the estimated parameters under all four models. In all models, transformed correlation parameters are used to reduce parameter space violations. We present both the transformed parameter (Fisher z transformed correlation and log odds ratio) as well as the parameter expressed on the original scale.

It is seen that the presence of neonatal seizures and/or of congenital malformation significantly decreases the probability of successfully performing

TABLE 7.13. *POPS Study. Parameter estimates (standard errors) for the trivariate Bahadur (BAH), probit (TPM) and Dale models (with probit, TPM-N, or logit, TPM-L, margins). For the associations, correlations [ρ and transformed correlations, using (7.12)] (BAH, TPM) and cross-ratios (ψ) and log cross-ratios for the TDM.*

	BAH	TPM	TDM-N	TDM-L
First ability score				
Intercept	3.67(0.49)	2.01(0.26)	2.03(0.27)	3.68(0.52)
Neonatal seiz.	-1.94(0.42)	-1.12(0.26)	-1.16(0.26)	-2.06(0.44)
Congenital malf.	-1.21(0.31)	-0.61(0.18)	-0.62(0.18)	-1.17(0.33)
100× Bilirubin	-0.69(0.25)	-0.32(0.14)	-0.32(0.14)	-0.64(0.27)
Second ability score				
Intercept	4.03(0.51)	2.19(0.27)	2.21(0.27)	4.01(0.54)
Neonatal seiz.	-2.26(0.43)	-1.27(0.26)	-1.29(0.26)	-2.28(0.44)
Congenital malf.	-1.08(0.32)	-0.56(0.19)	-0.59(0.19)	-1.11(0.34)
100× Bilirubin	-0.85(0.26)	-0.42(0.14)	-0.41(0.14)	-0.80(0.27)
Third ability score				
Intercept	3.32(0.50)	1.84(0.27)	1.91(0.27)	3.49(0.54)
Neonatal seiz.	-1.55(0.44)	-0.88(0.27)	-0.93(0.27)	-1.70(0.46)
Congenital malf.	-0.96(0.32)	-0.47(0.19)	-0.49(0.19)	-0.96(0.35)
100× Bilirubin	-0.44(0.26)	-0.21(0.14)	-0.24(0.14)	-0.49(0.28)
Association parameters				
	ρ	ρ	ψ	ψ
(1,2): ρ or ψ	0.27(0.05)	0.73(0.05)	17.37(5.19)	17.35(5.19)
(1,2): $z(\rho)$ or $\ln\psi$	0.55(0.11)	1.85(0.23)	2.85(0.30)	2.85(0.30)
(1,3): ρ or ψ	0.39(0.05)	0.81(0.04)	30.64(9.78)	30.61(9.78)
(1,3): $z(\rho)$ or $\ln\psi$	0.83(0.12)	2.27(0.25)	3.42(0.32)	3.42(0.32)
(2,3): ρ or ψ	0.23(0.05)	0.72(0.05)	17.70(5.47)	17.65(5.47)
(2,3): $z(\rho)$ or $\ln\psi$	0.47(0.10)	1.83(0.23)	2.87(0.31)	2.87(0.31)
(1,2,3): ρ or ψ	—	—	0.91(0.69)	0.92(0.69)
(1,2,3): $z(\rho)$ or $\ln\psi$	—	—	-0.09(0.76)	-0.09(0.76)
Log-likelihood	-598.44	-570.69	-567.11	-567.09

any of the three ability tests. A similar effect of bilirubin on the first and second ability score is observed.

The marginal regression parameters agree in pairs: the logit-based Bahadur and TDM-L models on the one hand and the TPM and TDM-N models on the other hand. There is as light tendency for the Bahadur pa-

rameter estimates and standard errors to be a bit smaller. This should not be seen as resulting from a higher efficiency, but rather as downward bias resulting from the model's stringent parameter space restrictions (Declerck, Aerts, and Molenberghs 1998). Upon multiplying the TPM and TDM-N coefficients with the factor $\pi/\sqrt{3}$, the standard deviation of the logistic distribution, all coefficient become very close to each other.

When comparing the association parameters, the (log) odds ratios are clearly very similar between both TDM models. This is less the case when the correlation estimates, obtained from the Bahadur, are compared with their probit model counterparts. A strong downward bias is seen. This is due, again, to the strong parameter space restrictions in the Bahadur case. This effect is magnified by setting the three-way correlation in the Bahadur model equal to zero. Recall that there is no such thing in the TPM, since this model is based on discretizing a multivariate standard normal distribution, which is completely described in terms of its two-way correlations, without the need to separately specifying three-way correlations.

A slight preference for the TDM could be inferred if based either on doubling the negative log-likelihood, or on the AIC. It is not possible to express a preference for either of the TDM models, based on this example. This confirms the well-known univariate result, that logistic and probit regression are hard to distinguish from each other, except when datasets become very large and response probabilities approach zero or one. While the TPM's performance is somewhat worse, the difference is around 5. Bahadur's model, on the other hand, lags behind by about 55 in deviance or AIC. Considering the strength of the association, there is a strong association between each pair of dichotomous responses, but no significant three-way association, as seen from the TDM.

An important feature of the likelihood method is that calculation of individual probabilities can be performed. For example, the method allows to calculate the joint probability of failing at the three tests. This can be quite different from the joint probability obtained by assuming independent responses, as is shown in Figure 7.7, where the probability that the child will fail on all three ability scores is calculated for different bilirubin values, given that both CGM and NSZ are one.

7.11 Longitudinal Analysis of the Fluvoxamine Study

The relationship between the severity of the side effects at the three visits and some baseline characteristics of the patients was established. The response is a trivariate ordered categorical vector with 4 classes, measured at three visits. For the selection of predictors, *age* and *sex* were included by default into the model. The other baseline characteristics were then consid-

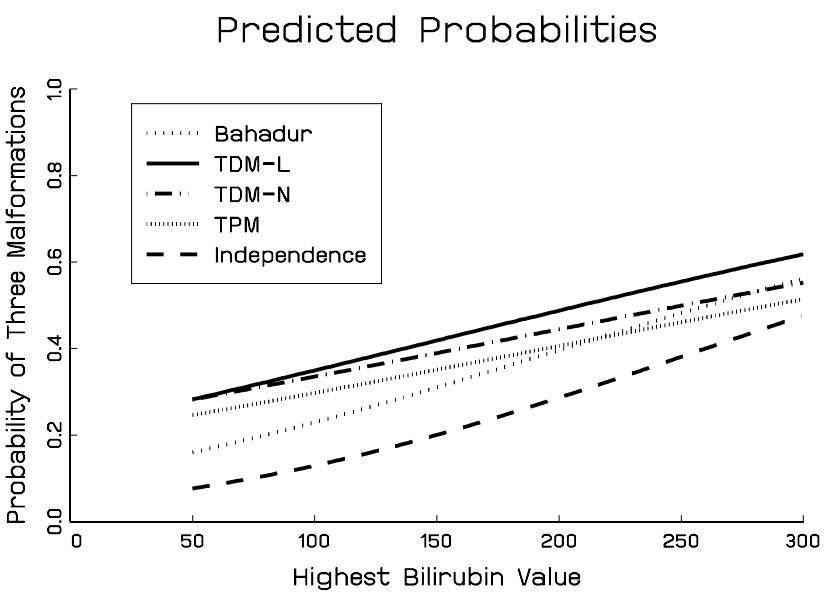

FIGURE 7.7. *POPS Study. probability that a child fails on all three ability scores for a range of bilirubin values, evaluated under five fitted models: the trivariate Bahadur model, the Dale model (TDM) with logistic (L) and normal (N) margins, the trivariate probit model (TPM), and a model assuming independent responses (three logistic regressions).*

ered for selection. Only the *duration* (months) of the disease and the *initial severity* (measured on a 7-point scale) turned out to significantly influence the severity of side effects.

At the second and third visit, a non-negligible portion of the patients dropped out from the study (20%). An ordinary contingency table analysis, as well as a logistic regression of the variable *dropout* on potential covariates showed that the dropout mechanism is heavily depending on the severity of the side-effect reported at the preceding visit. We refer to Part VI for several analysis explicitly addressing the missingness issue.

From the parameter estimates shown in Table 7.14 (Model I), it is seen that the effect of some covariates is almost constant over time. The G^2 test statistic for the hypothesis that both the intercepts and parameters for 'age' and 'sex' are time invariant is 5.37 (10 d.f., $p = 0.8654$). However, 'duration' and 'initial severity' depend on time. ($G^2 = 37.58$, 4 d.f, $p < 0.0001$). This leads to the more parsimonious Model II. The odds of observing high side-effects increases with 'age' and 'duration' and decreases with 'initial severity.' The influence of 'initial severity' increases over time. There is a strong association between side-effects measured at successive visits. Although significant, the association is less strong between the first and third visit.

TABLE 7.14. *Fluvoxamine Trial. Longitudinal analysis. The side effects at three successive times are regressed on age, duration, initial severity, and sex, using the multivariate Dale model. In Model I, the parameters are assumed to be different over time. In Model II, only* duration *and* initial severity *have a time-dependent effect. The entries represent the parameter estimates (standard errors).*

Effect	Side 1	Side 2	Side 3
		Model I	
		Marginal parameters	
Intercept 1	-0.41(0.90)	-0.45(0.95)	-0.79(1.06)
Intercept 2	1.78(0.90)	1.64(0.96)	1.64(1.07)
Intercept 3	2.94(0.92)	2.97(0.99)	2.85(1.13)
Age	-0.19(0.09)	-0.22(0.09)	-0.25(0.10)
Duration	-0.14(0.05)	-0.20(0.05)	-0.24(0.06)
In. Severity	0.29(0.14)	0.28(0.15)	0.42(0.17)
Sex	-0.23(0.24)	0.09(0.24)	0.16(0.27)
		Association parameters	
12	13	23	123
3.20(0.27)	2.49(0.28)	3.71(0.33)	-0.38(0.76)
		Model II	
		Marginal parameters	
Intercept 1		-0.52(0.82)	
Intercept 2		1.67(0.82)	
Intercept 3		2.89(0.84)	
Age		-0.21(0.07)	
Duration	-0.14(0.05)	-0.21(0.05)	-0.24(0.06)
In. Severity	0.27(0.13)	0.33(0.13)	0.42(0.13)
Sex		-0.06(0.22)	
		Association parameters	
12	13	23	123
3.13(0.26)	2.43(0.27)	3.74(0.33)	-0.29(0.74)

7.12 Appendix: Maximum Likelihood Estimation

We present details on a general expression for the likelihood in marginal models, the corresponding score equations, and how to solve them.

7.12.1 Score Equations and Maximization

Under a multinomial sampling scheme, the kernel of the log-likelihood, in terms of the counts obtained at design level r, $\boldsymbol{Z}_r^*$, and the corresponding

cell probabilities $\boldsymbol{\mu}_r^*$ is

$$\ell(\boldsymbol{\beta}; \boldsymbol{Z}^*) = \sum_{r=1}^{N} \boldsymbol{Z}_r^{*'} \ln[\boldsymbol{\mu}_r^*(\boldsymbol{\beta})].$$

When working with the cumulative counts $\boldsymbol{Z}_r$ and the cumulative probabilities $\boldsymbol{\mu}_r$, and knowing that relations (7.4) hold, we can rewrite the log-likelihood as

$$\ell(\boldsymbol{\beta}; \boldsymbol{Z}) = \sum_{r=1}^{N} \ell_r(\boldsymbol{\beta}; \boldsymbol{Z}_r) = \sum_{r=1}^{N} (B_r \boldsymbol{Z}_r)' \ln[B_r \boldsymbol{\mu}_r(\boldsymbol{\beta})]. \tag{7.58}$$

The derivative of the contribution of group r to (7.58) with respect to $\boldsymbol{\mu}_r$ is then given by

$$\begin{aligned} \frac{\partial \ell}{\partial \boldsymbol{\mu}_r} &= \frac{\partial \ell_r}{\partial \boldsymbol{\mu}_r} \\ &= \left\{ B_r' [\mathrm{diag}(\boldsymbol{\mu}_r^*)]^{-1} B_r \right\} (\boldsymbol{Z}_r - N_r \boldsymbol{\mu}_r) \\ &= \left\{ B_r' \mathrm{cov}(\boldsymbol{Z}_r^*)^{-1} B_r \right\} (\boldsymbol{Z}_r - N_r \boldsymbol{\mu}_r) \\ &= \mathrm{cov}(\boldsymbol{Z}_r)^{-1} (\boldsymbol{Z}_r - N_r \boldsymbol{\mu}_r). \end{aligned} \tag{7.59}$$

Given (7.59), the score function becomes

$$\boldsymbol{U}(\boldsymbol{\beta}) = \frac{\partial \ell}{\partial \boldsymbol{\beta}} = \sum_{r=1}^{N} \left(\frac{\partial \boldsymbol{\eta}_r}{\partial \boldsymbol{\beta}} \right)' \left[\left(\frac{\partial \boldsymbol{\eta}_r}{\partial \boldsymbol{\mu}_r} \right)' \right]^{-1} V_r^{-1} \boldsymbol{S}_r, \tag{7.60}$$

with $\boldsymbol{S}_r = \boldsymbol{Z}_r - N_r \boldsymbol{\mu}_r$, and $V_r = \mathrm{cov}(\boldsymbol{Z}_r)$. A typical element of V_r is

$$\begin{aligned} &\mathrm{cov}\left(z_r(k_1 \ldots k_{n_r}), z_r(\ell_1, \ldots, \ell_{n_r})\right) \\ &\quad = \mu_r(m_1, \ldots, m_{n_r}) - \mu_r(k_1, \ldots, k_{n_i}) \cdot \mu_r(\ell_1, \ldots, \ell_{n_r}), \end{aligned}$$

where $m_j = \min(k_j, \ell_j)$.

Computation of the matrix $Q_r = \partial \boldsymbol{\eta}_r / \partial \boldsymbol{\mu}_r$ is extremely simple if the link is of the form (7.17), because then (Grizzle, Starmer, and Koch 1969)

$$Q_r = C \left\{ \mathrm{diag}(A \boldsymbol{\mu}_r) \right\}^{-1} A. \tag{7.61}$$

This motivates the choice to compute Q_r and invert it, rather than computing Q_r^{-1} directly, as was done by Molenberghs and Lesaffre (1994) and detailed in Section 7.7.

When we use cumulative probabilities, the component $\mu_r(c_1, \ldots, c_{n_r}) = 1$, whence it can be omitted. This implies that the matrix Q_r is square and can easily be inverted. In case one chooses to use cell probabilities, all

components of $\boldsymbol{\mu}_r$ contain information whence the length of $\boldsymbol{\mu}_r^*$ is one more than the length of $\boldsymbol{\eta}_r$, but the probabilities sum to one. This additional equation needs to be added to the list of $\boldsymbol{\eta}_r$, making Q_r square again (McCullagh and Nelder, 1989).

Replacing the univariate marginal link functions in (7.17), $\boldsymbol{\eta}_r^{(1)}$ say, by any other inverse cumulative distribution function F^{-1} with probability density function f, and retaining the specification of the association in terms of a form satisfying (7.17), yields the expression

$$\boldsymbol{\eta} = \boldsymbol{\eta}(\boldsymbol{\mu}) = \begin{pmatrix} F^{-1}(\boldsymbol{\mu}^{(1)}) \\ \hline C_2 \ln(A\boldsymbol{\mu}) \end{pmatrix},$$

with corresponding derivative

$$Q_r = \left(\begin{array}{c|c} \operatorname{diag}\left\{ f(\boldsymbol{\eta}^{(1)}) \right\}^{-1} & \mathbf{0} \\ \hline C_2 \left\{ \operatorname{diag}(A\boldsymbol{\mu}) \right\}^{-1} A \end{array} \right). \tag{7.62}$$

The matrix C_2 is similar to the matrix C in (7.17) but now only applies to the association part of the model. Choosing $F = \Phi$ and $f = \phi$, the standard normal distribution and density functions, we obtain a global odds ratio model with univariate probit links.

As discussed in the previous section, the multivariate probit model also fits within the proposed framework. In this case, it might be preferable to compute the matrix Q_r^{-1}, rather than its inverse, unlike with the global odds ratio model, or most other models of the form (7.17). Although in the probit case the matrix Q_r^{-1} is easier to compute than Q_r, the computations are still more complex than calculating (7.62). The components are the derivatives of multivariate standard normal distribution functions. The evaluation of multivariate normal integrals is required. Lesaffre and Molenberghs (1991) chose to use the algorithm proposed by Shervish (1984). In the common case of linear predictors, the derivative of the link vector with respect to $\boldsymbol{\beta}$ is the design matrix X_r. See also Section 7.6.

The maximum likelihood estimator satisfies $\boldsymbol{U}(\widehat{\boldsymbol{\beta}}) = 0$. Two popular fitting algorithms are Fisher scoring and the Newton-Raphson algorithm. In the case of Fisher scoring, one starts with a vector of initial estimates $\boldsymbol{\beta}^{(0)}$ and updates the current value of the parameter vector $\boldsymbol{\beta}^{(t)}$ by

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + W(\boldsymbol{\beta}^{(t)})^{-1} \boldsymbol{U}(\boldsymbol{\beta}^{(t)}), \tag{7.63}$$

with

$$W(\boldsymbol{\beta}) = \sum_{i=1}^{N} N_r \left(\frac{\partial \boldsymbol{\eta}_r}{\partial \boldsymbol{\beta}} \right)' \left[\left(\frac{\partial \boldsymbol{\eta}_r}{\partial \boldsymbol{\mu}_r} \right)' \right]^{-1} V_r^{-1} \left[\left(\frac{\partial \boldsymbol{\eta}_r}{\partial \boldsymbol{\mu}_r} \right) \right]^{-1} \left(\frac{\partial \boldsymbol{\eta}_r}{\partial \boldsymbol{\beta}} \right).$$

The expected information matrix assumes the form $W(\boldsymbol{\beta})$, estimated by $W(\widehat{\boldsymbol{\beta}})$. A Newton-Raphson iteration scheme is found by substituting the

matrix $W(\boldsymbol{\beta})$ in (7.63) by $H(\boldsymbol{\beta})$, the matrix of second-order derivatives of the log-likelihood. An outline of this procedure for cumulative counts is given next.

7.12.2 Newton-Raphson Algorithm with Cumulative Counts

Replacing the matrix $W(\boldsymbol{\beta})$ in (7.63) by the matrix of second-order derivatives $H(\boldsymbol{\beta})$ of the log-likelihood (7.58) implements a Newton-Raphson algorithm. We present an expression for $H = H(\boldsymbol{\beta})$. It is useful to borrow some notation from McCullagh's (1987) book on tensor methods in statistics. From McCullagh (1987), it follows that the (p, q) element of H is

$$\begin{aligned} H_{pq} &= \sum_{r=1}^{N} \sum_{a,b,c,d} \frac{\partial \eta_{ra}}{\partial \beta_p} \frac{\partial \mu_{rb}}{\partial \eta_{ra}} \frac{\partial^2 \ell}{\partial \mu_{rb} \mu_{rc}} \frac{\partial \mu_{rc}}{\partial \eta_{rd}} \frac{\partial \eta_{rd}}{\partial \beta_q} \\ &\quad + \sum_{r=1}^{N} \sum_{a,d,b} \left[\frac{\partial \eta_{ra}}{\partial \beta_p} \frac{\partial \eta_{rd}}{\partial \beta_q} \frac{\partial^2 \mu_{rb}}{\partial \eta_{ra} \partial \eta_{rd}} + \frac{\partial^2 \eta_{ra}}{\partial \beta_p \partial \beta_q} \frac{\partial \mu_{rb}}{\partial \eta_{ra}} \right] \frac{\partial \ell}{\partial \mu_{rb}}. \end{aligned}$$

Observing

$$\begin{aligned} \frac{\partial \ell}{\partial \mu_{rb}} &= \sum_k (V_r^{-1})_{bk} S_{rk}, \\ \frac{\partial^2 \ell}{\partial \mu_{rb} \mu_{rc}} &= -N_r * (V_r^{-1})_{bc} - \sum_{e,f,k} (V_r^{-1})_{be} J_{r,c,ef} (V_r^{-1})_{fk} S_{rk}, \\ J_{r,c,ef} &= \delta_{c,\iota(e,f)} - \delta_{ce} \mu_{rf} - \delta_{cf} \mu_{re}, \end{aligned}$$

where δ is the Kronecker delta function and $\iota(a, d) = c$ if $\min(a_j, d_j) = c_j$ for all components of the index vectors, we can separate the terms involving S_r in the expression for $H(\boldsymbol{\beta})$:

$$H_{pq} = -W_{pq} + \sum_{r=1}^{N} \boldsymbol{\alpha}'_{rpq} \boldsymbol{S}_r,$$

for some vector $\boldsymbol{\alpha}_{rpq}$. Obviously, the second term has expectation zero.

The first and second derivatives of μ_r with respect to ν_r follow from the identities

$$\begin{aligned} \delta_{bc} &= \frac{\partial \mu_{rb}}{\partial \eta_{ra}} \frac{\partial \eta_{ra}}{\partial \mu_{rc}}, \\ \frac{\partial^2 \mu_{rb}}{\partial \eta_{ra} \partial \eta_{rd}} &= -\sum_{c,a,v} \frac{\partial^2 \eta_{rc}}{\partial \mu_{ra} \partial \mu_{rv}} \frac{\partial \mu_{rb}}{\partial \eta_{rc}} \frac{\partial \mu_{ra}}{\partial \eta_{ra}} \frac{\partial \mu_{rv}}{\partial \eta_{rd}}. \end{aligned}$$

Note that the first identity merely rephrases that $Q_r Q_r^{-1} = I$.

Opting for linear predictors, we obtain:

$$\frac{\partial \boldsymbol{\eta}_i}{\partial \boldsymbol{\beta}} = X_i \qquad \frac{\partial^2 \eta_{it}}{\partial \beta_p \partial \beta_q} = 0.$$

We are now able to rewrite the Hessian in a concise matrix form

$$H(\boldsymbol{\beta}) = \sum_{r=1}^{N} X_r^{'} \left[F_r + (Q_r^{'})^{-1} G_r (Q_r)^{-1} \right] X_r$$

with

$$\begin{aligned} F_r &= \left(\sum_b \frac{\partial^2 \mu_{rb}}{\partial \eta_{ra} \partial \eta_{rd}} \frac{\partial \ell}{\partial \mu_{rb}} \right)_{a,d}, \\ G_r &= \frac{\partial^2 \ell}{\partial \boldsymbol{\mu}_r \partial \boldsymbol{\mu}_r^{'}}. \end{aligned}$$

Finally, if we again choose a link function of the type (7.17) we can use simple forms

$$Q_r = \frac{\partial \boldsymbol{\eta}_r}{\partial \boldsymbol{\mu}_r} = C \left\{ \operatorname{diag}(A \boldsymbol{\mu}_r) \right\}^{-1} A = C B_r A$$

and

$$\frac{\partial^2 \eta_{ra}}{\partial \mu_r \mu_r^{'}} = -A' B_{ra}^{(2)} A,$$

where the matrix $B_{ra}^{(2)}$ is obtained by multiplying all rows of B_r^2 with the ath row of C.

7.12.3 Determining the Joint Probabilities

To compute the score equations and to implement the updating algorithm, knowledge of the multivariate cumulative probabilities $\boldsymbol{\mu}_r$ is required. The choice of a fitting technique will strongly depend on the choice of link functions. For multivariate odds ratio models (multivariate Dale models, see also Section 7.7) several proposals have been made, such as the use of multivariate Plackett probabilities (Plackett 1965, Molenberghs and Lesaffre 1994), the use of Lagrange multipliers (Lang and Agresti 1994), and a Newton iteration mechanism (Glonek and McCullagh 1995). With the Plackett probability approach, we found that for four and higher dimensional problems, the derivatives of high dimensional polynomials can become numerically unstable. Here, the iterative proportional fitting (IPF) algorithm is adapted to produce a quick and reliable tool to compute the cumulative probabilities. A similar use of the IPF algorithm was proposed by Kauermann (1993). Due to the use of score function (7.60), there is no

need to compute the *derivatives* of the probabilities directly since Q_r easily follows from (7.61), leaving only the probabilities to be computed.

Given the marginal probabilities and the odds ratio parameters, our IPF algorithm produces a multidimensional table of cell probabilities. The IPF algorithm is known from its use in fitting log-linear models (Bishop, Fienberg, and Holland 1975), where the association is described using *conditional* odds ratios. The algorithm was also applied by Fitzmaurice and Laird (1993) for their mixed marginal-conditional models (Section 7.8). In our fully marginal models, marginal odds ratios are used. We distinguish between two types. Global odds ratios, given in (7.21)–(7.23), are relevant for ordinal responses (Dale 1984), and local odds ratios as in (7.24) are a natural choice for nominal outcomes. Of course, both sets coincide for binary responses.

We will describe our algorithm for global odds ratios first, and then discuss the local odds ratio version in the concluding paragraph of this section. We need to determine the cumulative probabilities $\mu_r(k_1, \ldots, k_{n_r})$ which correspond to cumulative cell count $Z_r(k_1, \ldots, k_{n_r})$. Recall that this notation encompasses not only n_r-way classifications, but also one-way, two-way,... classifications, by setting an appropriate set of indices $k_j = c_j$. Omitting indices for which $k_j = c_j$, we assume without loss of generality that we need to determine a K-way probability $\mu_r(k_1, \ldots, k_K)$, with $k_j < c_j$ for all j.

We will proceed recursively. First, note that the cumulative probabilities $\mu_r(\ell_1, \ldots, \ell_K)$, with $\ell_j \in \{k_j, c_j\}$ for $j = 1, \ldots, K$, completely describe a 2^K contingency table. Second, as soon as at least one $\ell_j = c_j$, we obtain a lower order probability. Our recursion will be based on the assumption that these lower order probabilities have been determined. The starting point of the inductive construction is obtained by setting all but one $\ell_j = c_j$, whence we obtain univariate probabilities μ_{rjk_j} which are of course easy to determine from the marginal links η_{rjk_j}. Drop the index r from notation.

From the cumulative probabilities, we easily determine the cell probabilities $\mu^{z_1 \ldots z_K}_{k_1 \ldots k_K}$, with $z_j = 1, 2$ and adopt the convention that the K-way cumulative cell probabilities are incorporated as:

$$\mu^{1 \ldots 1}_{k_1 \ldots k_K} = \mu(k_1, \ldots, k_K). \tag{7.64}$$

We will explicitly need the cell probabilities of dimension $K - 1$:

$$\mu^{z_1 \ldots z_{j-1} z_{j+1} \ldots z_K}_{k_1 \ldots k_{j-1} k_{j+1} \ldots k_K} = \sum_{z_j=1}^{2} \mu^{z_1 \ldots z_K}_{k_1 \ldots k_K}.$$

The IPF algorithm is started by choosing a table of initial values, e.g.,

$$\mu^{z_1 \ldots z_K}_{k_1 \ldots k_K}(0) = \begin{cases} \psi_r(k_1, \ldots, k_K) & \text{if } (z_1, \ldots, z_K) = (1, \ldots, 1), \\ 1 & \text{otherwise.} \end{cases}$$

with $\psi_r(k_1,\ldots,k_K) = \exp[\eta_r(k_1,\ldots,k_K)]$, the corresponding global odds ratio. This table clearly has the correct association structure, but the marginals are incorrect and the sum of the cell counts is not equal to one. Updating cycle $(m+1)$ requires K substeps, to match each of the $K-1$ dimensional marginal tables:

$$\mu^{z_1\ldots z_K}_{k_1\ldots k_K}\left(m+\frac{j}{K}\right) = \mu^{z_1\ldots z_K}_{k_1\ldots k_K}\left(m+\frac{j-1}{K}\right).\frac{\mu^{z_1\ldots z_{j-1}z_{j+1}\ldots z_K}_{k_1\ldots k_{j-1}k_{j+1}\ldots k_K}}{\mu^{z_1\ldots z_{j-1}z_{j+1}\ldots z_K}_{k_1\ldots k_{j-1}k_{j+1}\ldots k_K}\left(m+\frac{j-1}{K}\right)},$$

$(j=1,\ldots,K)$, the argument of μ indicating the iteration subcycle. Upon convergence, (7.64) can be used to identify the required K-way probability.

Convergence of the IPF algorithm is established in Csiszar (1975). However, the parameter space of the marginal odds ratios is constrained, unless in the special case of a constant odds ratio for a bivariate outcome (Liang, Zeger, and Qaqish 1992). A violation of the constraints will be revealed by a cumulative probability vector with negative entries. If this occurs in the course of an updating algorithm, appropriate action (e.g., step halving) has to be taken. We never encountered problems of this kind, suggesting that the constraints are very mild. Practice suggests that these restrictions are much milder than those for the Bahadur model with which a fully satisfactory analysis of the fluvoxamine data (Section 7.2.4) was not possible.

For marginal local odds ratios a slightly adapted and simpler procedure is proposed. Instead of considering subsets of binary variables, we now consider the whole marginal multi-way table directly. With a similar recursive argument, we assume that the full set of marginal tables up to dimension $K-1$ is determined. Then, we construct a K-dimensional initial table

$$\mu^*_r(k_1,\ldots,k_K)(0) = \prod_{c_j>\ell_j\geq k_j}\psi^*_r(\ell_1,\ldots,\ell_K),$$

for all cells $(k_1,\ldots,k_K)$. This table clearly has got the required K-way local association structure. The updating algorithm matches the entries to the K sets of $K-1$ dimensional marginal tables.

7.13 Appendix: The Multivariate Plackett Distribution

Let us start from the bivariate case first. Given the marginal distributions $F_1(w_1)$, $F_2(w_2)$ and the cross-ratio ψ, the Plackett distribution is the solution of the second degree polynomial equation

$$\psi(F-a_1)(F-a_2)-(F-b_1)(F-b_2)=0, \tag{7.65}$$

where $a_1 = F_1, a_2 = F_2, b_1 = 0, b_2 = F_1+F_2-1$. The solution of this equation is given by (7.40). To yield a genuine distribution function, the

solution F of (7.65) should satisfy the Fréchet inequalities (Fréchet 1951):

$$\max(b_1, b_2) \leq F \leq \min(a_1, a_2).$$

Now, this approach can be generalized to n dimensions. To define the multivariate Plackett distribution, consider the set of $2^n - 1$ generalized cross-ratios with values in $[0, +\infty]$:

$$\begin{aligned}
\psi_j, &\qquad (1 \leq j \leq n)\\
\psi_{j_1 j_2}, &\qquad (1 \leq j_1 < j_2 \leq n)\\
&\vdots\\
\psi_{j_1 \ldots j_k}, &\qquad (1 \leq j_1 < \ldots < j_k \leq n)\\
&\vdots\\
\psi_{1\ldots n}.
\end{aligned}$$

The one-dimensional ψ_j's are precisely the odds of the univariate probabilities, i.e.,

$$\psi_j = \frac{\mu_1^j}{\mu_2^j} = \frac{F_j}{1 - F_j}, \tag{7.66}$$

$(1 \leq j \leq n)$. Note that we put the response level in the subscript to μ and the occasions to which they pertain the superscript. Thus, μ_1^j is the probability to observe a '1' at occasion j and μ_2^j is the probability to observe a '2' at this occasion. Similar conventions will be used for the higher orders. The bivariate associations $\psi_{j_1 j_2}$ are defined as in (7.41):

$$\psi_{j_1 j_2} = \frac{\mu_{11}^{j_1 j_2} \mu_{22}^{j_1 j_2}}{\mu_{12}^{j_1 j_2} \mu_{21}^{j_1 j_2}} = \frac{F_{j_1 j_2}(1 - F_{j_1} - F_{j_2} + F_{j_1 j_2})}{(F_{j_1} - F_{j_1 j_2})(F_{j_2} - F_{j_1 j_2})}, \tag{7.67}$$

$(1 \leq j_1 < j_2 \leq n)$. As soon as $\psi_{j_1}, \psi_{j_2}, \psi_{j_1 j_2}$ are known, $F_{j_1 j_2}$ can be calculated. The cross-ratio $\psi_{j_1 j_2}$ can also be viewed as the odds ratio of $\psi_{j_1(1)}, \psi_{j_2(2)}$, computed as in (7.66), within the first and second level of dimension j_2, respectively.

The three-dimensional cross-ratios can be defined in a similar way as the three factor interactions in loglinear models (Agresti 1990) and is analogous to the above extension. They have been considered already in, for example, (7.21), (7.22), and (7.23). Thus, the cross-ratio $\psi_{j_1 j_2 j_3}$ is defined as the ratio of two conditional cross-ratios $\psi_{j_1 j_2(1)}$ and $\psi_{j_1 j_2(2)}$. These are the two-dimensional cross-ratios defined within the first and second level of dimension j_3 respectively. The numerator of $\psi_{j_1 j_2 j_3}$ contains $F_{j_1 j_2 j_3}$ with a positive sign and the denominator contains $F_{j_1 j_2 j_3}$ with a negative sign. Again, the knowledge of the cross-ratios enables one to determine $F_{j_1 j_2 j_3}$.

However, care has to be taken when specifying the cross-ratios, since not every combination leads to a valid solution. This is not surprising, and occurred earlier with the Bahadur model (Section 7.2). Also the multivariate

probit model of Section 7.6 is subject to such constraints, since the correlation matrix has to be positive definite. In fact, such constraints will show up for every marginal model, because specifying marginal models implies specifying overlapping information, in contrast to conditional models, the genesis of which can be viewed as specifying new model components, conditional upon ones already in the model. Although this may seem a drawback, it is largely compensated by ease of interpretation for the corresponding model parameters, marginal regression functions, etc.

The n-dimensional probabilities can be computed if all lower dimensional probabilities together with the global cross-ratio of dimension n are known. Let $\mu_{k_1\ldots k_m}^{j_1\ldots j_m}$ be the $(k_1,\ldots,k_m)$-orthant probability of the m-dimensional marginal table, formed by dimensions $(j_1,\ldots,j_m)$. We present the defining equation for $F_{m_1\ldots m_k}$:

$$\psi_{j_1\ldots j_m} = \frac{\prod\limits_{(k_1,\ldots,k_m)\in A_m^+} \mu_{k_1\ldots k_m}^{j_1\ldots j_m}}{\prod\limits_{(k_1,\ldots,k_m)\in A_m^-} \mu_{k_1\ldots k_m}^{j_1\ldots j_m}}, \tag{7.68}$$

where

$$A_m^+ = \{(k_1,\ldots,k_m)\in\{1,2\}^m | 2 \text{ divides } \sum_{\ell=1}^m k_\ell - m\}$$

and

$$A_m^- = \{1,2\}^m \backslash A_m^+,$$

'\' indicating set difference. In particular, for $F_{1\ldots n}$:

$$\psi_{1\ldots n} = \frac{\prod\limits_{(j_1,\ldots,j_n)\in A_n^+} \mu_{j_1\ldots j_n}}{\prod\limits_{(j_1,\ldots,j_n)\in A_n^-} \mu_{j_1\ldots j_n}}. \tag{7.69}$$

For example, for $n=3$:

$$\begin{aligned} A_1^+ &= \{1\}, \\ A_2^+ &= \{(1,1),(2,2)\}, \\ A_3^+ &= \{(1,1,1),(1,2,2),(2,1,2),(2,2,1)\}. \end{aligned}$$

Based on these expressions, (7.68) yields (7.66), (7.67), and the three-dimensional odds-ratio

$$\psi_{123} = \frac{\mu_{111}\mu_{122}\mu_{212}\mu_{221}}{\mu_{112}\mu_{121}\mu_{211}\mu_{222}}.$$

The orthant probabilities $\mu_{k_1\ldots k_n}$ are determined by the distribution F. A general expression can be derived, which will be useful for the automated

computation of the orthant probabilities. Some notation is needed. Let $\lambda(\boldsymbol{k}) \equiv \lambda(k_1, \ldots, k_n)$ be the set of places for which k_j is equal to 1, (e.g., $\lambda(1,2,1,1) = \{1,3,4\}$), then

$$\mu_{k_1 \ldots k_n} = \sum_{\boldsymbol{s} \supset \lambda(\boldsymbol{k})} \operatorname{sgn}(\boldsymbol{s}) F_{\boldsymbol{s}}, \tag{7.70}$$

where

$$\operatorname{sgn}(\boldsymbol{s}) = \begin{cases} 1 & \text{if } \#\boldsymbol{s} - \#\beta(\boldsymbol{k}) \text{ is even,} \\ -1 & \text{otherwise,} \end{cases}$$

and $F_{\boldsymbol{s}} = F_{s_1 \ldots s_m}$, with $s_1 \leq \ldots \leq s_m$. In the three-dimensional case, the octant probabilities are

$$\begin{aligned} \mu_{111} &= F_{123}, \\ \mu_{112} &= F_{12} - F_{123}, \\ \mu_{121} &= F_{13} - F_{123}, \\ \mu_{211} &= F_{23} - F_{123}, \\ \mu_{122} &= F_1 - F_{12} - F_{13} + F_{123}, \\ \mu_{212} &= F_2 - F_{12} - F_{23} + F_{123}, \\ \mu_{221} &= F_3 - F_{13} - F_{23} + F_{123}, \\ \mu_{222} &= 1 - F_1 - F_2 - F_3 + F_{12} + F_{13} + F_{23} - F_{123}. \end{aligned} \tag{7.71}$$

As an example, consider μ_{212}. In this case, $\lambda(2,1,2) = \{2\}$ and there are 4 possible vectors $\boldsymbol{s}$: (2), (1,2), (2,3) and (1,2,3). Therefore, (7.70) yields the expression for μ_{212} in (7.71).

The set of $2^n - 1$ generalized cross-ratios fully specifies the n-dimensional Plackett distribution. However, from the above reasoning it is not clear whether such a distribution always exists. Further, if existence and uniqueness is guaranteed it is not yet clear how to calculate the distribution since it is only implicitly specified by (7.68). These matters are discussed next.

Let us turn to some computational details. Note that the probabilities in the numerator (denominator) of (7.69) involve $+F_{12\ldots n}$ ($-F_{12\ldots n}$) and that both numerator and denominator contain an even number of factors. Thus, (7.69) may be abbreviated as

$$\psi = \frac{\prod_{i=1}^{2^{n-1}} (F - b_i)}{\prod_{i=1}^{2^{n-1}} (F - a_i)}, \tag{7.72}$$

where $\psi \equiv \psi_{1\ldots n}$ and $F \equiv F_{1\ldots n}$. The a_i and b_i are functions of the $(n-1)$- and lower-dimensional probabilities (or, equivalently, cross-ratios). A valid solution must satisfy

$$\max_i b_i \leq F \leq \min_i a_i. \tag{7.73}$$

However, this condition is not satisfied for all choices of a_i and b_i. To see this, take the three-way Plackett distribution. Then, according to (7.73), the one- and two-dimensional marginal distributions have to satisfy the following inequalities:

$$\begin{aligned} F_{j_1 j_2} + F_{j_1 j_3} &\leq F_{j_1} + F_{j_2 j_3}, \qquad (j_1 \neq j_2 \neq j_3 \neq j_1) \\ F_1 + F_2 + F_3 &\leq 1 + F_{12} + F_{13} + F_{23}. \end{aligned}$$

Now, as a counterexample, if

$$\begin{aligned} F_1 = F_2 = F_3 &= \frac{1}{2}, \\ \psi_{12} &= 0.05, \\ \psi_{13} &= 1, \\ \psi_{23} &= 20, \end{aligned}$$

then $F_{13} + F_{23} > F_3 + F_{12}$ and (7.73) cannot be satisfied.

In spite of this, the constraints for this model never were burdensome, neither in the analyses reported in this book, nor for others done by the authors and reported elsewhere. The same holds for the multivariate probit model. This is in contrast to the Bahadur model, where the analysis of the fluvoxamine trial (Section 7.2.4) already posed insurmountable problems.

In case (7.73) is satisfied, existence and uniqueness of a solution is guaranteed by the next lemma. The verification of (7.73) is straightforward, as the functions b_i and a_i are linear functions of the lower order marginal probabilities.

Lemma 7.1 *Let*

$$P(C) = \psi \prod_{i=1}^{m} (C - a_i) - \prod_{i=1}^{m} (C - b_i),$$

where m is even, $0 < \psi < +\infty$, and

$$b_1 = \max_{1 \leq i \leq m} b_i < \min_{1 \leq i \leq m} a_i = a_1,$$

then the interval $]b_1, a_1[$ contains exactly one real root of $P(C)$.

Proof. The inequalities

$$P(a_1) = -\prod_{i=1}^{m} (a_1 - b_i) < 0$$

and

$$P(b_1) = \psi \prod_{i=1}^{m} (b_1 - a_i) > 0,$$

together with the continuity of $P(C)$, establish the existence.

Now,

$$\frac{\partial P}{\partial C} = \psi \sum_{i=1}^{m} \prod_{j \neq i} (C - a_j) - \sum_{i=1}^{m} \prod_{j \neq i} (C - b_j) = \psi \sum_i T_i - \sum_i S_i.$$

T_i is a product of $(m-1)$ negative factors, whence T_i is negative. S_i is positive, so $P(C)$ is strictly decreasing in $]b_1, a_1[$, establishing the result.

It follows from the proof that the regula falsi method with starting points a_1 and b_1 always leads to the solution. Though in general a_1 and b_1 are close to each other and convergence is quickly reached, it is desirable to look for even faster methods. It is our experience that a Newton iteration with starting point, for example, $\frac{1}{2}(a_1 + b_1)$ converges to the root, generally in 3 or 4 steps (with convergence criterion: $|c_{k+1} - c_k| < 10^{-8}$).

An algebraic solution to the two-dimensional problem is given by Mardia (1970) and Dale (1986). The three-way Plackett distribution can also be solved algebraically using Ferrari's method for solving fourth-degree polynomials. However, the solution cannot be written in a mathematically elegant way. From the four-way Plackett distribution on, one has to rely on numerical techniques. It is a fundamental result of algebra that a polynomial of degree higher than 5 has no algebraic solution. This is not a major disadvantage, since numerical methods for the multivariate Dale model are usually much faster than for the multivariate probit model, which necessitates the calculation of multivariate normal integrals.

7.14 Appendix: Maximum Likelihood Estimation for the Dale Model

We present the basic tools for the computations. We distinguish between the following parts: model description, likelihood function and cell probabilities, and score functions and information matrix.

It is, again, convenient to adopt the contingency table notation, assuming that subjects $i = 1, \ldots, N_r$ are grouped within covariate or design levels $r = 1, \ldots, N$ (Section 7.1). Thus, observations, sharing covariate vector $\boldsymbol{x}_r$, are combined into a single $c_1 \times \ldots \times c_n$ contingency table. The dimension of this table will be abbreviated by $\boldsymbol{c}$. In other words, we adopt the table notation. Denote the entries of this table by $z_{r\boldsymbol{k}}$. Here, $\boldsymbol{k}$ indicates a multi-index: $\boldsymbol{k} = (k_1, \ldots, k_n)$, $(1 \leq k_j \leq c_j, l = 1, \ldots, n)$. In vector notation: $\mathbf{1} \leq \boldsymbol{k} \leq \boldsymbol{c}$. A particular table is indicated by $(z_{r\boldsymbol{k}})_{\boldsymbol{k}}$.

We assume that the tables are sampled from a multinomial distribution, with cell probabilities $(\mu^*_{r\boldsymbol{k}})_{\boldsymbol{k}}$, $(r = 1, \ldots, N)$, given by the MDM. They are derived from the orthant probabilities, defined by (7.70). The model is fully specified by link functions $\eta_{rjk} = \eta_{rj}(\boldsymbol{x}_r)$ given by (7.48), $\gamma^{k_1 k_2}_{j_1 j_2}(\boldsymbol{x}_r)$

given by (7.49), together with the higher-order association parameters. If we denote them by ϕ with an appropriate subscript, then we obtain in vector notation $\ln \psi_{\boldsymbol{h}} = \phi_{\boldsymbol{h}}$, with $\boldsymbol{h}$ a vector running through all higher order associations. The parameters γ and ϕ determine the association structure.

Assume that all parameters form a column vector $\boldsymbol{\theta}$. The log-likelihood takes the form:

$$\ell(\boldsymbol{\theta}) = \sum_{r=1}^{N} \sum_{\boldsymbol{k}=\boldsymbol{1}}^{\boldsymbol{c}} z_{r\boldsymbol{k}} \ln \mu^*_{\boldsymbol{k}}(\boldsymbol{\theta}, \boldsymbol{x}_r), \tag{7.74}$$

and is fully determined if we indicate in what way the cell probabilities $\mu_{r\boldsymbol{k}}(\boldsymbol{\theta}) = \mu_{\boldsymbol{k}}(\boldsymbol{\theta}, \boldsymbol{x}_r)$ arise from the link functions. Let $\mu_{r\boldsymbol{k}} = \mu_{\boldsymbol{k}}(\boldsymbol{x}_r)$, denote the n-dimensional cumulative Plackett distribution function F, evaluated in the appropriate links:

$$\mu_{r\boldsymbol{k}} = F(\boldsymbol{\eta}_r, \boldsymbol{\gamma}_r, \phi), \tag{7.75}$$

where the arguments are appropriately vectorized forms of the links. Note that $\mu_{r\boldsymbol{k}}$ is the orthant probability of $[-\infty, \eta_{r1k_1}] \times \ldots \times [-\infty, \eta_{rnk_n}]$. To compute the cell probabilities, write the cutpoints for dimension j as:

$$-\infty = \eta_{rj0} < \eta_{rj1} < \ldots < \eta_{rj,c_j-1} < \eta_{rjc_j} = +\infty.$$

If one or more components k_j of $\boldsymbol{k}$ are equal to zero, the corresponding orthant probability $\mu_{r\boldsymbol{k}}$ vanishes. If one or more components of $\boldsymbol{k}$ equal c_j, then $\mu_{r\boldsymbol{k}}$ is an orthant probability of a lower dimensional marginal distribution.

The cell probabilities $\mu^*_{r\boldsymbol{k}}$ can be expressed in terms of $\mu_{r\boldsymbol{k}}$:

$$\mu^*_{r\boldsymbol{k}} = \sum_{\boldsymbol{h}} (-1)^{S(\boldsymbol{k},\boldsymbol{h})} \mu_{r\boldsymbol{h}}.$$

Summation goes over all indices $\boldsymbol{h}$ satisfying $\boldsymbol{0} \leq \boldsymbol{k} - \boldsymbol{h} \leq \boldsymbol{1}$, and the function S is defined by $S(\boldsymbol{k}, \boldsymbol{h}) = \sum_{j=1}^{n} k_j - h_j$. The computation of $\mu_{\boldsymbol{k}}$ in (7.75) involves the evaluation of the cumulative Plackett distribution. The derivatives are computed by implicit derivation of (7.72).

The derivative of the log-likelihood ℓ with respect to a marginal parameter θ can be written as:

$$\frac{\partial \ell}{\partial \theta} = \sum_{r=1}^{N} \sum_{\boldsymbol{k}=\boldsymbol{1}}^{\boldsymbol{c}} z_{r\boldsymbol{k}} \frac{1}{\mu^*_{r\boldsymbol{k}}} \sum_{j=1}^{n} \sum_{m=1}^{c_j-1} \frac{\partial \mu^*_{r\boldsymbol{m}}}{\partial \eta_{jm}(\boldsymbol{x}_r)} \frac{\partial \eta_{jm}(\boldsymbol{x}_r)}{\partial \theta}. \tag{7.76}$$

A few conventions will simplify notation. First, assume there is only one covariate vector $\boldsymbol{x}$, thereby dropping the index r. Second, due to model (7.48), a marginal parameter pertains to only one margin, j say. For such a parameter, summation over all $j = 1, \ldots, n$ is replaced by a single j. In principle, we need to distinguish between intercepts $\beta_{0,jm}$, corresponding to

only one cutpoint m, and covariate parameters β, common to all cutpoints $k = 1, \ldots, c_j - 1$ of dimension j. However, we assume that *every* marginal parameter pertains to only one cutpoint, m_j say. The correct formula can be obtained by summing over all cutpoints, if needed. In conclusion, j and $m = m_j$ will be assumed to be fixed. Finally, note that in most formulas, some indices k_j of $\boldsymbol{k}$ will play a particular role and need being mentioned explicitely. The remaining indices will be denoted by $\boldsymbol{k'}$. Accordingly, the upper bound is denoted by $\boldsymbol{c'}$. In subscripts (e.g., $\mu^*_{\boldsymbol{k}}$), only the relevant indices will be mentioned. Applying these conventions to (7.76) yields

$$\frac{\partial \ell}{\partial \theta} = \frac{\partial \eta_{jm}}{\partial \theta} \sum_{\boldsymbol{k'}=\boldsymbol{1}}^{\boldsymbol{c'}} \left(\frac{z_m}{\mu^*_m} - \frac{z_{m+1}}{\mu^*_{m+1}} \right) \sum_{\boldsymbol{h}, h_j = m} (-1)^{S(\boldsymbol{k'}_m, \boldsymbol{h})} \frac{\partial \mu_{\boldsymbol{h}}}{\partial \eta_{jm}}.$$

For an intercept or covariate parameter in the two-way association model, we deduce

$$\frac{\partial \ell}{\partial \theta} = \frac{\partial \gamma_{j_1 j_2}}{\partial \theta} \sum_{\boldsymbol{k}} z_{\boldsymbol{k}} \frac{1}{\mu^*_{\boldsymbol{k}}} \psi^{k_1 k_2}_{j_1 j_2} \sum_{\boldsymbol{h}} (-1)^{S(\boldsymbol{k}, \boldsymbol{h})} \frac{\partial \mu_{\boldsymbol{h}}}{\partial \psi_{j_1 j_2}}.$$

Note that a similar form is obtained for higher order associations. For a parameter θ in (7.49) pertaining to a row category m, the score equation is

$$\frac{\partial \ell}{\partial \theta} = \frac{\partial \gamma_{j_1 j_2}}{\partial \theta} \sum_{\boldsymbol{k'}=\boldsymbol{1}}^{\boldsymbol{c'}} \left(\frac{z_m}{\mu^*_m} - \frac{z_{m+1}}{\mu^*_{m+1}} \right) \psi^{m\ k_2}_{j_1 j_2} \sum_{\boldsymbol{h}, h_{j_1} = m} (-1)^{S(\boldsymbol{k'}_m, \boldsymbol{h})} \frac{\partial \mu_{\boldsymbol{h}}}{\partial \psi_{j_1 j_2}},$$

while for a cell-specific parameter we find

$$\begin{aligned} \frac{\partial \ell}{\partial \theta} \quad = \quad & \frac{\partial \gamma_{j_1 j_2}}{\partial \theta} \sum_{\boldsymbol{k'}=\boldsymbol{1}}^{\boldsymbol{c'}} \left(\frac{z_{m_1 m_2}}{\mu^*_{m_1 m_2}} - \frac{z_{m_1+1, m_2}}{\mu^*_{m_1+1, m_2}} - \frac{z_{m_1, m_2+1}}{\mu^*_{m_1, m_2+1}} + \frac{z_{m_1+1, m_2+1}}{\mu^*_{m_1+1, m_2+1}} \right) \\ & \times \psi^{m_1 m_2}_{j_1 j_2} \sum_{\boldsymbol{h}, h_{j_1} = m_1, h_{j_2} = m_2} (-1)^{S(\boldsymbol{k}_{m_1 m_2}, \boldsymbol{h})} \frac{\partial \mu_{\boldsymbol{h}}}{\partial \psi_{j_1 j_2}}. \end{aligned}$$

Straightforward but lengthy computations lead to expressions for the elements of the expected information matrix. We do not present them here; they are available as a technical report from the first author. They are used to implement a Fisher scoring algorithm, to maximize (7.74).