

## Loglinear Marginal Models

Loglinear models provide the most flexible tools for analyzing relationships among categorical variables in complex tables. It will be shown in this chapter how to apply these models in the context of marginal modeling. First, in Section 2.1, the basics of ordinary loglinear modeling will be explained. The main purpose of this section is to introduce terminology and notation and those aspects of loglinear modeling that will be used most in the remainder of this book. It will be assumed that the reader already has some familiarity with loglinear modeling and, therefore, the discussion will be concise. An advanced overview of loglinear models is provided by Agresti (2002); an intermediate one by Hagenaars (1990) and an introduction is given by Knoke and Burke (1980) among many others. In Section 2.2, several motivating examples will be presented showing what types of research questions can be answered by means of loglinear marginal modeling. Finally, in Section 2.3, a general ML estimation procedure will be discussed for testing and estimating loglinear marginal models.

### 2.1 Ordinary Loglinear Models

#### 2.1.1 Basic Concepts and Notation

The most simple applications of loglinear models are to two-dimensional tables such as Table 2.1, in which the self-reported *Political Orientation* ( $P$ ) and *Religion* ( $R$ ) of a sample of 911 U.S. citizens is cross-classified. Table 2.1 contains the raw frequencies as well as the vertical percentages. Variable  $P$  has seven categories, ranging from extremely liberal ( $P = 1$ ) to extremely conservative ( $P = 7$ ). *Religion* has three categories: Protestant ( $R = 1$ ), Catholic ( $R = 2$ ) and None ( $R = 3$ ). The joint probability that  $P = i$  and  $R = j$  is denoted by  $\pi_{ij}^{PR}$ . The number of categories of  $P$  is  $I = 7$  and of  $R$  is  $J = 3$ . For a first interpretation of the data, the vertical percentages in Table 2.1 are useful for comparing the conditional distributions of *Political Orientation* for the three religious groups. It can easily be seen that the nonreligious people are more liberal than the Protestants or the Catholics, but the differences between the latter two groups are less clear. Even in this simple example, a more formal approach may

**Table 2.1.** *Political Orientation and Religion in the United States in 1993* (Source: General Social Survey 1993)

<i>Political Orientation (P)</i>	<i>Religion (R)</i>			Total
	1. Protestant	2. Catholic	3. None	
1. Extremely liberal	11 (1.8%)	2 (1.0%)	4 (4.4%)	17 (1.0%)
2. Liberal	49 (8.0%)	21 (10.1%)	23 (25.3%)	93 (10.2%)
3. Slightly liberal	79 (12.9%)	23 (11.1%)	19 (20.9%)	121 (13.3%)
4. Moderate	220 (35.8%)	96 (46.4%)	30 (33.0%)	346 (38.0%)
5. Slightly conservative	112 (18.3%)	36 (17.4%)	9 (9.9%)	157 (17.2%)
6. Conservative	119 (19.4%)	27 (13.0%)	4 (4.4%)	150 (16.5%)
7. Extremely conservative	23 (3.8%)	2 (1.0%)	2 (2.2%)	27 (3.0%)
Total	613 (100%)	207 (100%)	91 (100%)	911 (100%)

*Note: The small Jewish and Other religious groups are omitted*

be needed to separate true population differences from sampling fluctuations and to arrive at a clear and parsimonious description of the data.

Saturated loglinear models decompose the observed logarithms of the cell probabilities in terms of loglinear parameters without imposing any restrictions on the data:

$$\log \pi_{ij}^{PR} = \lambda + \lambda_i^P + \lambda_j^R + \lambda_{ij}^{PR}.$$

The parameter  $\lambda$  is called the overall effect,  $\lambda_i^P$  is the effect of category  $i$  of  $P$ ,  $\lambda_j^R$  is the effect of category  $j$  of  $R$ , and  $\lambda_{ij}^{PR}$  is the two-variable effect of categories  $i$  and  $j$  of  $P$  and  $R$ . Note that the term ‘effect’ is not intended to have a causal connotation here: it simply refers to a parameter in the loglinear model and the term ‘effect’ is only used for convenience to avoid complicated and awkward phrases (see also Chapter 5).

The loglinear model can also be represented in its multiplicative form as a direct function of the cell frequencies or probabilities, rather than of the log cell frequencies or log probabilities:

$$\pi_{ij}^{PR} = \tau \tau_i^P \tau_j^R \tau_{ij}^{PR}.$$

The multiplicative parameters, denoted as  $\tau$ , have nice interpretations in terms of odds and odds ratios. However, formulas and computations are simpler in their loglinear representations, and therefore we will mostly use the additive loglinear form of the model. It is, of course, easy to switch between the two representations by means of the transformation  $\tau = e^\lambda$ .

For the purposes of this book, a somewhat different notation than this standard notation is often needed, because in marginal analyses it is generally necessary to indicate from which marginal table a particular loglinear parameter is calculated. In this new notation, the superscripts will indicate the relevant marginal table. In the loglinear equation above, all parameters are calculated from table  $PR$ . Therefore, all

parameters will get  $PR$  as their superscript. To indicate to which effect a particular symbol refers, the pertinent variable(s) will be indexed while the others get an asterisk (\*) as their subscript. For example, parameter  $\lambda_{i*}^{PR}$  (which is the same as  $\lambda_i^P$  in traditional notation) is the effect of category  $i$  of  $P$  calculated from table  $PR$ . Throughout this book, we will generally use this ‘marginal’ notation, unless its use becomes too cumbersome. In all cases, the meaning of the notation used will be made clear or will be evident from the context.

The equation for the saturated loglinear model above now looks as follows in the marginal notation:

$$\log \pi_{ij}^{PR} = \lambda_{**}^{PR} + \lambda_{i*}^{PR} + \lambda_{*j}^{PR} + \lambda_{ij}^{PR}.$$

Without further restrictions, the  $\lambda$ -parameters are not identified. For example, there are already as many unknown two-variable parameters  $\lambda_{ij}^{PR}$  as there are known cell frequencies. One common identification method is to use *effect coding* (as in traditional ANOVA models), where for all effects the loglinear parameters sum to zero over any subscript. Letting the ‘+’-sign in a subscript represents summation over that subscript, e.g.

$$\lambda_{+*}^{PR} = \sum_i \lambda_{i*}^{PR},$$

the following identifying restrictions are imposed:

$$\lambda_{+*}^{PR} = \lambda_{*+}^{PR} = 0$$

and

$$\lambda_{i+}^{PR} = \lambda_{+j}^{PR} = 0 \quad \text{for all } i, j.$$

Using effect coding, the parameters can be computed as follows:

$$\begin{aligned} \lambda_{**}^{PR} &= \frac{1}{IJ} \sum_{k=1}^I \sum_{l=1}^J \log \pi_{kl}^{PR}, \\ \lambda_{i*}^{PR} &= \frac{1}{IJ} \sum_{k=1}^I \sum_{l=1}^J \log \frac{\pi_{il}^{PR}}{\pi_{kl}^{PR}}, \\ \lambda_{*j}^{PR} &= \frac{1}{IJ} \sum_{k=1}^I \sum_{l=1}^J \log \frac{\pi_{kj}^{PR}}{\pi_{kl}^{PR}}, \\ \lambda_{ij}^{PR} &= \frac{1}{IJ} \sum_{k=1}^I \sum_{l=1}^J \log \frac{\pi_{ij}^{PR} \pi_{kl}^{PR}}{\pi_{il}^{PR} \pi_{kj}^{PR}}. \end{aligned}$$

The overall effect  $\lambda_{**}^{PR}$  is in principle always present in a loglinear model. It is a normalizing constant that guarantees that the estimated probabilities sum to 1 or the estimated cell frequencies to sample size  $N$ . The overall effect equals the mean of the log (expected) probabilities in the table, as can be seen from the way it is computed. The one-variable effect  $\lambda_{i*}^{PR}$  is the mean of the log odds (or logits) in the table that have  $\pi_{il}^{PR}$  in the numerator. Roughly speaking, it indicates how much larger the probability is that someone belongs to  $P = i$  rather than to any of the other categories of

$P$ , on average among the religious groups. The one-variable effect  $\lambda_{*j}^{PR}$  is the mean of those log odds in the table that have  $\pi_{kj}^{PR}$  in the numerator. It is especially important in the context of this book to realize that generally the one-variable parameters do not reflect the marginal distribution of  $P$  or  $R$ , but the average conditional distribution of  $P$  and  $R$ , respectively. Restrictions on the one-variable parameters are therefore not restrictions on the one-variable marginal distributions, but on the average conditional one-variable distributions (and this extends analogously to multiway tables and multiway marginals). The one-variable effects are almost always included in a loglinear model, unless one wants to explicitly test hypotheses about the average conditional distribution of a particular variable, which is rarely the case. Finally, parameter  $\lambda_{ij}^{PR}$  equals the mean of the logs of the odds ratios in the table which have  $\pi_{ij}^{PR}$  in the numerator. The two-variable parameters reflect the sizes of the log cell probability due to the association between  $P$  and  $R$ , and indicate how much bigger or smaller a particular cell probability is than expected on the basis of the lower-order effects. The variables  $P$  and  $R$  are statistically independent of each other if and only if  $\lambda_{ij}^{PR} = 0$  for all  $i$  and  $j$ .

A second common method for obtaining an identified model is *dummy coding*, where each parameter that refers to any of the reference categories of the variables is set equal to zero. Using the first category of each variable as the reference category, this method amounts to setting

$$\lambda_{1*}^{PR} = \lambda_{*1}^{PR} = 0$$

and

$$\lambda_{i1}^{PR} = \lambda_{1j}^{PR} = 0 \quad \text{for all } i, j.$$

The choice of the first category of each variable as the reference category is arbitrary, and for each variable any of its categories could, in principle, be used as the reference category.

It is important to note that the values of the parameters will differ from each other depending on the kinds of identifying restrictions chosen: effect coding yields different parameter values from dummy coding and, when using dummy coding, selecting the first category as the reference category leads to different parameter values than using the last category. However, whatever reference category is used, the substantive interpretations in terms of what goes on in the table will remain the same, provided the appropriate interpretation of the parameters is employed, taking the nature of the chosen identifying restrictions into account. The values of the odds and the odds ratios estimated under a particular model will be the same regardless of the particular identification constraints chosen. In this book, we will use effect coding unless stated otherwise. The two-variable parameter estimates for Table 2.1 are given in Table 2.2, using effect coding.

Like the vertical percentages in Table 2.1, the estimates  $\hat{\lambda}$  of the saturated loglinear model presented in Table 2.2 clearly indicate that nonreligious people are more liberal than Protestants or Catholics. But it is harder to discover a clear pattern for the differences between the two religious groups, i.e., between Catholics and Protestants.

**Table 2.2.** *Political Orientation and Religion in the United States in 1993: Estimates  $\hat{\lambda}_{ij}^{PR}$  for Table 2.1. Effect coding is used; \* significant at .05 level*

<i>Political Orientation (P)</i>	<i>Religion (R)</i>		
	1. Protestant	2. Catholic	3. None
1. Extremely liberal	-.20	-.52	.72*
2. Liberal	-.57*	-.04	.61*
3. Slightly liberal	-.22	-.07	.29
4. Moderate	-.17	.39*	-.22
5. Slightly conservative	.11	.36*	-.47*
6. Conservative	.52*	.42*	-.93*
7. Extremely conservative	.52*	-.54	.01

A comparison of particular restricted, nonsaturated models might provide some better insights into what is going on.

Nonsaturated loglinear models are usually tested by means of two well-known test statistics: the likelihood ratio test statistic

$$G^2 = -2N \sum_i p_i \log \frac{\hat{\pi}_i}{p_i}$$

and Pearson's chi-square test statistic

$$X^2 = N \sum_i \frac{(p_i - \hat{\pi}_i)^2}{\hat{\pi}_i}.$$

If the postulated model is true, these test statistics have an asymptotic chi-square distribution. The degrees of freedom ( $df$ ) equal the number of independent restrictions on the nonredundant loglinear parameters (often the number of nonredundant parameters that are set to zero) or, equivalently, to the number of independent constraints on the cell probabilities. In many circumstances,  $G^2$  can be used to obtain a more powerful conditional test by testing a particular model, not (as implied above) against the saturated model, but against an alternative that is more restrictive than the saturated model (but less restrictive than the model to be tested). Given that interest lies in a model  $M_1$  with  $df_1$  degrees of freedom, a conditional test requires that an alternative hypothesis  $M_2$  is considered with  $df_2$  degrees of freedom that contains model  $M_1$  as a special case, i.e.,  $M_1 \subset M_2$ . The conditional test statistic is then defined as

$$G^2(M_1|M_2) = G^2(M_1) - G^2(M_2)$$

and has an asymptotic chi-square distribution with  $df = df_1 - df_2$  if  $M_1$  is true. This conditional testing procedure is valid only under the condition that the more general model  $M_2$  is (approximately) valid in the population.

To indicate (non)saturated hierarchical loglinear models, use will be made of the standard short-hand notation. This short-hand notation can be most easily described

in terms of the standard, nonmarginal notation for the  $\lambda$  parameters. In this shorthand notation then, a loglinear model is denoted by the superscripts of all its highest order interaction terms. Because of the hierarchical nature of the model, all lower-order effects that can be formed from these superscripts are also included in the loglinear model. For Table 2.1, the saturated model is denoted as model  $\{PR\}$  and the independence model with  $\lambda_{ij}^{PR} = 0$  as  $\{P, R\}$ .

The hypothesis of statistical independence between  $P$  and  $R$  in Table 2.1 is definitely not a viable hypothesis: likelihood ratio chi-square  $G^2 = 54.9$ ,  $df = 12$  ( $p = .000$ ; Pearson chi-square  $X^2 = 57.4$ ). However, a partial independence model can be formulated in which the conditional probability distributions of *Political Orientation* are the same for Protestants and Catholics, but different for the nonreligious people. In loglinear terms, this form of partial independence is identical to the restriction that the two-variable  $\lambda$ 's are the same for Protestants and Catholics:  $\lambda_{i1}^{PR} = \lambda_{j2}^{PR}$  for all  $i = j$ . Several programs are available for handling such restrictions, e.g., Vermunt's free software LEM (Vermunt, 1997a). The test results are  $G^2 = 15.0$ ,  $df = 6$  ( $p = .024$ ; Pearson chi-square  $X^2 = 13.8$ ). This is a somewhat inconclusive result, the interpretation of which strongly depends on the (arbitrarily) chosen significance level of .05 or .01. Assuming that the partial independence model is valid in the population, independence model  $\{P, R\}$  can be tested conditionally against this less restricted alternative:  $G^2 = 54.9 - 15.0 = 39.9$ ,  $df = 12 - 6 = 6$ ,  $p = .000$ . The complete independence model definitely has to be rejected in favor of the partial independence model.

A more powerful investigation of what goes on in the table might be obtained by explicitly taking into account the ordered nature of variable  $P$ . There are essentially three partly overlapping ways in which we can deal explicitly with loglinear models for ordered data. If the ordered nature of the data is considered to be the result of strictly ordinal measurement, it makes sense to assume (weakly) monotonically increasing or declining relationships between the variables and impose inequality restrictions on the loglinear association parameters. If the ordered data are considered as interval-level data, fixed numerical scores can be assigned to the interval-level variables and the loglinear parameters may be linearly restricted to obtain linear relationships in the loglinear models. In the third approach, the scores for the variables are not fixed, but linear relationships are assumed to be true and the variable scores are estimated in such a way that the relationships in the loglinear model will be linear. An extensive literature on loglinear modeling of ordinal data exists (see Croon, 1990; Vermunt, 1999; Hagenaaars, 2002; Clogg & Shihadeh, 1994).

By way of example, variable  $P$  in table  $PR$  may be considered as an interval-level variable with fixed scores  $P_i$  and  $R$  may be treated as nominal-level variable, resulting in an *interval by nominal* loglinear model (also called a column association model). The model has the form

$$\log \pi_{ij}^{PR} = \lambda_{**}^{PR} + \lambda_{i*}^{PR} + \lambda_{*j}^{PR} + P_i \alpha_j^R$$

in which the two-variable effect  $\lambda_{ij}^{PR}$  is replaced by the more parsimonious term  $P_i \alpha_j^R$ . In terms of ordinary regression analysis, the term  $\alpha_j^R$  is similar to a regression

coefficient: in this case, one for each category of  $R$ , and the scores  $P_i$  define the independent interval-level variable  $X$ . To maintain the identifying effect coding restrictions for the restricted  $\lambda_{ij}^{PR}$  effects, the scores  $P_i$  must sum to 0 (we will use the equal unit distance interval scores  $-3, -2, \dots, 2, 3$ ) and we need  $\sum_j \alpha_j^R = 0$ . This linear model implies that the log odds of belonging to religious group  $j$  rather than  $j'$  increase or decrease linearly with an increasing score on  $P$ . Or formulated the other way around, the log odds of belonging to category  $i$  of  $P$  rather than  $i + 1$  are systematically larger (or systematically smaller) for  $R = j$  than for  $R = j'$ , where these log odds differences between religious groups  $j$  and  $j'$  are the same for all values of  $i$ :

$$\alpha_j^R - \alpha_{j'}^R = \log \frac{\pi_{ij}^{PR} / \pi_{i+1,j}^R}{\pi_{i,j'}^{PR} / \pi_{i+1,j'}^R}.$$

As can be seen from this formula, the odds ratio on the right-hand side has the same value for all  $i$ . The difference  $\alpha_j^R - \alpha_{j'}^R$  indicates how much higher the log odds of scoring one category higher on the political orientation scale is for people of religion  $j$  than for people of religion  $j'$ . Since there are three religious denominations here, there are three relevant (and two independent) differences of this kind.

For the data in Table 2.1, the ordinal by nominal model fits well:  $G^2 = 14.4$ ,  $df = 10$  ( $p = .16$ ,  $X^2 = 15.16$ ). The estimates of the regression coefficients are  $\hat{\alpha}_1^R = .226$ ,  $\hat{\alpha}_2^R = .096$ , and  $\hat{\alpha}_3^R = -.322$ , which shows that the Protestants are the most conservative and the nonreligious people are the most liberal, while the Catholics occupy an intermediate position very close to the Protestants. As explained below, for the fitted sample data, the odds of scoring one category higher on the liberal-conservative scale is just 1.14 times higher for Protestants than for Catholics, but 1.73 times higher for Protestants than for nonreligious people, and, finally, 1.52 times higher for Catholics than for nonreligious people. Coefficient 1.14 for the comparison Protestants-Catholics is computed as follows:  $1.14 = \exp(.226 - .096)$ ; using the estimated standard errors of  $\hat{\alpha}^R$  (not reported here) its 95% confidence interval (CI) equals  $[1.01, 1.29]$ ; the coefficient for the comparison of Protestants-nonreligious is  $1.73 = \exp(.226 + .322)$  and its CI equals  $[1.45, 2.06]$ ; the coefficient for the comparison Catholic-nonreligious is  $1.52 = \exp(.10 + .32)$  and its CI equals  $[1.25, 1.84]$ . The difference in political orientation of Catholics and Protestants is not very large and the reported confidence interval for the pertinent odds ratio almost includes the no difference value of 1. To test whether the (linear) difference between Catholics and Protestants is significant, the same interval by nominal model can be defined, but now with the extra restriction that  $\alpha_1^R = \alpha_2^R$ . The test outcomes for this model are  $G^2 = 18.8$ ,  $df = 11$  ( $p = .07$ ,  $X^2 = 18.84$ ). In this restricted interval by nominal model,  $\hat{\alpha}_1^R = \hat{\alpha}_2^R = .17$  and  $\hat{\alpha}_3^R = -.34$ . On the basis of the unconditional test outcome against the alternative hypothesis that the saturated model is true ( $p = .07$ ), one might decide to accept the restricted model and conclude that Protestants and Catholics have the same political orientation. The more powerful conditional test for the thus restricted model against the alternative hypothesis that the original interval by nominal model holds yields  $G^2 = 18.8 - 14.4 = 4.4$ ,  $df = 11 - 10 = 1$  ( $p = .04$ ).

**Table 2.3.** *Political Orientation, Religion, and Opinion on teenage birth control in the United States in 1993 (Source: General Social Survey 1993)*

		<i>Opinion on Teenage Birth Control (B)</i>											
		1. Strongly agree			2. Agree			3. Disagree			4. Strongly disagree		
<i>Religion (R)</i>		1	2	3	1	2	3	1	2	3	1	2	3
<i>Pol. or. (P)</i>	1	5	1	3	4	0	0	0	0	1	2	1	0
	2	18	6	10	15	6	10	9	6	3	7	3	0
	3	24	7	7	29	11	7	18	5	4	8	0	1
	4	61	31	13	69	30	7	54	20	4	36	15	6
	5	19	11	5	32	11	3	37	8	0	24	6	1
	6	13	6	2	31	8	1	32	6	0	43	7	1
	7	5	0	1	5	1	0	4	0	1	9	1	0

At the 5% significance level, the unrestricted interval by nominal model is accepted but its restricted version is rejected.

From all these test outcomes, it can be clearly concluded that first, there is no reason to reject the linear nature of the relationships in the interval by nominal model; second, that nonreligious people are definitely more liberal than Catholics or Protestants, and third, that the differences in political orientation between Catholics and Protestants are small. For the time being, it may be accepted that Catholics are slightly more liberal than Protestants but new data are needed to conform this outcome. Suspension of judgement is the best option here (Hays, 1994, p. 281).

### 2.1.2 Modeling Association Among Three Variables

Basic loglinear modeling for two-way tables can easily be extended to tables of much higher dimensions. As a simple example, the data in three-way Table 2.3 will be used to investigate how the two variables dealt with so far, *Religion (R)* and *Political Orientation (P)*, affect opinion on teenage *Birth Control (B)*. Variable *B* has  $K = 4$  categories, ranging from strongly agree to strongly disagree.

For the three-dimensional table *PRB*, saturated loglinear model  $\{PRB\}$  decomposes the log probabilities as follows:

$$\log \pi_{ijk}^{PRB} = \lambda_{***}^{PRB} + \lambda_{i**}^{PRB} + \lambda_{*j*}^{PRB} + \lambda_{**k}^{PRB} + \lambda_{i j*}^{PRB} + \lambda_{i*k}^{PRB} + \lambda_{*jk}^{PRB} + \lambda_{ijk}^{PRB}.$$

Note that here the superscripts of the  $\lambda$  parameters are all *PRB*, indicating that the parameters refer to table *PRB* rather than to table *PR* from the previous subsection. Where the highest order effect parameter in the previous subsection was a two-variable effect, now we have also a three-variable parameter  $\lambda_{ijk}^{PRB}$  that indicates to what extent the conditional associations between any of the two variables vary among the categories of the third variable.



The effect coding identifying restrictions are

$$\begin{aligned}\lambda_{+**}^{PRB} &= \lambda_{*+*}^{PRB} = \lambda_{**+}^{PRB} = 0 \\ \lambda_{i+*}^{PRB} &= \lambda_{+j*}^{PRB} = \lambda_{i**}^{PRB} = \lambda_{+*k}^{PRB} = \lambda_{*j+}^{PRB} = \lambda_{**k}^{PRB} = 0 \\ \lambda_{i j+}^{PRB} &= \lambda_{i+k}^{PRB} = \lambda_{+jk}^{PRB} = 0\end{aligned}$$

for all  $i, j$ , and  $k$ . Because there are sampling zeroes in the observed table  $PRB$ , the sample values of the loglinear parameters of the saturated model are either plus or minus infinity or undefined.

In general, as indicated above, the loglinear effects pertaining to the same variables are different when calculated in different (marginal) tables; even their signs may be different. For the data in Table 2.3, we have  $\hat{\lambda}_1^P = -1.65$ ,  $\hat{\lambda}_{1*}^{PR} = -1.50$ , and  $\hat{\lambda}_{1**}^{PRB} = -\infty$  (minus infinity). All three parameters pertain to the distribution of  $P$  and represent the effect of the first category of  $P$ , but are calculated in the marginal tables  $P$ ,  $PR$ , and the full table  $PRB$ , respectively (and assuming saturated models for the pertinent tables). Parameter  $\lambda_1^P$  reflects the cell size of  $P = 1$  in the univariate marginal distribution of  $P$ ; parameter  $\lambda_{1*}^{PR}$  indicates the cell size of  $P = 1$  on average in the  $J$  conditional distributions of  $P$  in table  $PR$ ; and parameter  $\lambda_{1**}^{PRB}$  mirrors the average cell size  $P = 1$  in the  $J \times K$  conditional distributions of  $P$  in table  $PRB$ . The two-variable parameters for table  $PRB$  are now partial coefficients indicating the direct relationship between two variables on average within the categories of the third variable, in this way controlling for the third variable.

Table 2.4 contains the test outcomes of a few relevant hierarchical loglinear models for Table 2.3 concerning the influence of  $P$  and  $R$  on  $B$ . The models in Table 2.3 can also be seen as logit models for the effects of  $P$  and  $R$  on  $B$  (Agresti, 2002, Section 8.5). The models are again represented in the usual short-hand notation by means of which hierarchical models are indicated by their highest order interactions, implying the presence of all pertinent lower order effects. The second column in Table 2.3 gives an interpretation of the model in terms of (conditional) independence relations ( $\perp$ ) or the absence of interaction terms. The final four columns summarize the results of the testing procedures.

In the last row of Table 2.4, the results of the no three-factor interaction model are given. The no three-factor interaction model has the form

$$\log \pi_{ijk}^{PRB} = \lambda_{***}^{PRB} + \lambda_{i**}^{PRB} + \lambda_{*j*}^{PRB} + \lambda_{**k}^{PRB} + \lambda_{i j*}^{PRB} + \lambda_{i *k}^{PRB} + \lambda_{*jk}^{PRB}.$$

As can be seen in Table 2.4, this model  $\{PR, PB, RB\}$  fits the data well. However, the estimated (and observed) table is sparse, which may invalidate the approximation of the chi-square distribution. It is not certain whether the reported  $p$ -value for the model is correct. One may become more confident that the model can be accepted by observing that the value of Pearson's chi-square statistic is 35.2, which is not too different from  $G^2$ . The no three-factor interaction model will be accepted here and used as an alternative hypothesis for testing more parsimonious models: especially with sparse tables, conditional tests more readily approximate the theoretical chi-square distribution and are in many circumstances more powerful than unconditional tests.

**Table 2.4.** Goodness of fit of various hierarchical loglinear models for Table 2.3

Model	Interpretation	$G^2$	$df$	$p$ -value	$X^2$
1. $\{PR, B\}$	$PR \perp\!\!\!\perp B$	120.2	60	.000	105.3
2. $\{PR, RB\}$	$P \perp\!\!\!\perp B   R$	95.6	54	.000	84.7
3. $\{PR, PB\}$	$R \perp\!\!\!\perp B   P$	53.7	42	.107	46.8
4. $\{PR, PB, RB\}$	No 3-factor interaction	39.2	36	.328	35.2

The model in which neither  $P$  nor  $R$  have an effect on  $B$  does not fit the data (model 1 in Table 2.3). The same is true for conditional independence model 2 in which  $P$  has no direct influence on  $B$ , but  $R$  has. At first sight, conditional independence model 3, in which *Religion* has no direct effect on opinion of teenage *Birth Control*, provides an acceptable fit to the data ( $p = .107$ ). However, testing this model against the no three-factor interaction model yields  $G^2 = 14.7$  with  $df = 6$  ( $p = .025$ ). This conditional test has more power to detect the (possibly small) effects of  $R$  on  $B$  in the population than the corresponding unconditional test. Although the conclusion regarding  $p = .025$  again strongly depends on the chosen significance level .01 or .05, we will proceed cautiously and at least for the time being accept the possibility of (small) effects of  $R$  on  $B$  in the population. Model 3 will be rejected in favor of the no three-factor interaction model 4.

In the no three-factor interaction model 4, the conditional association between  $P$  and  $B$  given  $R$  is described by the 28 parameters  $\lambda_{i* k}^{PRB}$  and the conditional association between  $R$  and  $B$  given  $P$  is described by the 12 parameters  $\lambda_{* j k}^{PRB}$ . A simpler description of the models might be obtained by taking the ordered character of variables  $P$  and  $B$  into account. More precisely, variable  $B$  will be treated as an interval-level variable with scores  $B_k = -1.5, -.5, +.5, +1.5$  and also variable  $P$  will be considered (as before) as an interval-level variable with scores  $P_i = -3, -2, \dots, +2, +3$ . Further, the following restrictions will be applied:

$$\begin{aligned}\lambda_{i* k}^{PRB} &= \vartheta P_i B_k \\ \lambda_{* j k}^{PRB} &= \rho \gamma_j^R B_k.\end{aligned}$$

These restrictions are similar to the ones used above in the interval by nominal model for Table  $PR$ , but a slightly different notation than above is employed to indicate somewhat different aspects of models for ordered data. The resulting loglinear model has the form

$$\log \pi_{i j k}^{PRB} = \lambda_{***}^{PRB} + \lambda_{i**}^{PRB} + \lambda_{* j*}^{PRB} + \lambda_{***}^{PRB} + \lambda_{i j*}^{PRB} + \vartheta P_i B_k + \rho \gamma_j^R B_k. \quad (2.1)$$

The parameter  $\vartheta$  is a kind of regression coefficient for the linear effect of  $P$  on  $B$ , and  $\rho \gamma_j^R$  is the regression coefficient for the linear relationship between  $R$  and  $B$ , one for each category of  $R$ . One might also say that  $\rho$  is the regression coefficient and consider  $\gamma_j^R$  as scores to be estimated for  $R$ , given a linear relationship between  $R$  and  $B$ . In order to guarantee model identification, and more specifically to guarantee identification of the product  $\rho \gamma_j^R$ , the additional constraint  $\sum_j \gamma_j^R = 0$  is imposed.

If  $\rho$  and  $\gamma_j^R$  have to be identified separately, which is not necessary here for our purposes, the variance of the estimated scores  $\gamma_j^R$  has to be fixed, e.g., by means of the restriction  $\sum_j (\gamma_j^R)^2 = 1$ .

According to this model, the direct relationship between  $P$  and  $B$  is linear in the sense that the log odds of choosing category  $k$  of  $B$  rather than  $k'$  increase (or decrease) linearly with increasing values for  $P$ , or vice versa, but less appropriate here given the assumed ‘causal’ order of the variables: the log odds of choosing category  $i$  of  $P$  rather than  $i'$  increase (or decrease) linearly with increasing values for  $B$ . The relation between  $R$  and  $B$  is linearly restricted in the sense of the interval by nominal model discussed in the previous section: for two religions  $j$  and  $j'$ , there is a linear relationship between *Religion* and the opinion about teenage *Birth Control*. One way to clarify the meanings of the effects of  $P$  and  $R$  on  $B$ , i.e., of the consequences of having different scores on  $P$  or  $R$  for the scores on  $B$  is the following:

$$\vartheta(P_i - P_{i'}) = \log \frac{\pi_{ijk}^{PRB} / \pi_{ijk+1}^{PRB}}{\pi_{i'jk}^{PRB} / \pi_{i'jk+1}^{PRB}} \quad (2.2)$$

$$\rho(\gamma_j^R - \gamma_{j'}^R) = \log \frac{\pi_{ijk}^{PRB} / \pi_{ijk+1}^{PRB}}{\pi_{i'jk}^{PRB} / \pi_{i'jk+1}^{PRB}}. \quad (2.3)$$

The odds ratio on the right-hand side of (2.2) is the conditional odds ratio indicating the direct relationship between  $P$  and  $B$  for  $R = j$ . It turns out to be the same for all values of  $j$ , a necessary consequence of the no three-variable-interaction model in (2.1). Further, the conditional odds ratio on the right-hand side of (2.3) indicating the direct relationship between  $R$  and  $B$  for  $P = i$  is the same for all values of  $i$ . The left-hand side element  $\rho(\gamma_j^R - \gamma_{j'}^R)$  shows how much higher the log odds of scoring one category higher on  $B$  is for people in category  $j$  of  $R$  than for people in category  $j'$  of  $R$ , conditionally on  $P$ . The left-hand side element  $\vartheta(P_i - P_{i'})$  indicates how much higher the log odds of scoring one category higher on  $B$  is for people in category  $i$  of  $P$  than for people in category  $i'$  of  $P$ , conditionally on  $R$ . The linearly restricted model for the direct relation between  $P$  and  $B$  is called an ‘interval by interval’ or a ‘linear by linear’ model, and also a ‘uniform association model’, because all local partial odds ratios for the direct relation between  $P$  and  $B$  are the same, their logarithm being  $\vartheta$ . In terms of the original lambda parameters,

$$\vartheta = \lambda_{i*k}^{PRB} - \lambda_{i+1*k}^{P R B} - \lambda_{i*k+1}^{PR B} + \lambda_{i+1*k+1}^{P R B},$$

for all values of  $i$  and  $k$ .

Testing model (2.1) yields  $G^2 = 57.7$  with  $df = 57$  ( $p = .449$ ,  $X^2 = 52.7$ ); testing it against the no three-factor interaction model, it is found  $G^2 = 57.7 - 39.2 = 18.5$  with  $df = 57 - 42 = 15$  ( $p = .247$ ). There is no reason to reject this very parsimonious model for describing the association structure between the three variables. The relevant estimated effects are

$$\begin{aligned} \hat{\vartheta} &= .155, \\ \widehat{\rho\gamma_1^R} &= .182, \end{aligned}$$

$$\begin{aligned}\widehat{\rho\gamma}_2^R &= .026, \\ \widehat{\rho\gamma}_3^R &= -.208.\end{aligned}$$

The estimated direct linear effect of *Political Orientation* on the opinion on teenage *Birth Control* is significant (estimated standard errors not reported here) and in the expected direction: the more conservative one is, the more one is opposed to teenage birth control. The (significant) direct effects of *Religion* indicate that nonreligious people are less opposed to birth control than the Protestants with the Catholics in an intermediate position.

The effects of *Political Orientation* on the opinion on teenage *Birth Control* are much stronger than the effects of *Religion*. One way to see this clearly is to estimate the maximum effects for the variables, that is, the largest (log) odds ratios that can be obtained in the pertinent tables. Because of the linear relationship and the number of categories of the variables, the log of the maximum odds ratio for the effect of *P* on *B* turns out to be  $6 \times 3 \times \hat{\vartheta} = \hat{\lambda}_{1*1}^{PRB} - \hat{\lambda}_{1*4}^{PRB} - \hat{\lambda}_{7*1}^{PRB} + \hat{\lambda}_{7*4}^{PRB} = 6 \times 3 \times .1548 = 2.786$ . The corresponding maximum odds ratio equals  $\exp(2.786) = 16.22$ . Similar computations lead to a maximum (log) effect of *R* on *B* of 1.169 ( $= 3 \times (.182 - (-.208))$ ) and a corresponding odds ratio of 3.220: the effect of *P* on *B* is five times stronger than the effect of *R* on *B*.

## 2.2 Applications of Loglinear Marginal Models

The loglinear models applied in the previous section to analyze a joint probability distribution can also be employed for jointly analyzing two or more marginal distributions. In this section, several concrete research problems and designs will be discussed that require marginal-modeling methods, and for which loglinear marginal models are very useful to answer the pertinent research questions. Real-world examples and data will be used to illustrate these kinds of research questions and the ways they can be translated into the language of loglinear modeling. Maximum likelihood estimates of the parameters and significance tests for these examples will be given, along with their substantive explanation. In the last section, a general algorithm will be presented to obtain maximum likelihood estimates for loglinear marginal models. The contents of some parts of this section will be more demanding from a statistical point of view and are indicated by \*\*\*.

### 2.2.1 Research Questions and Designs Requiring Marginal Models

As discussed before, marginal modeling is about the simultaneous analyses of marginal distributions where the different marginal distributions involve dependent observations, but where the researcher is in principle not interested in the nature of the dependencies. As the remainder of this book will show, this is a research situation that actually occurs a lot in practice. For a simple concrete example, let us turn to the analysis of family data. Family data are of interest for social scientists studying such diverse topics as social mobility, political change, changing family relations or

the societal role of generational differences. In this respect, social scientists want to compare family members, wives and husbands, children and parents, and sisters and brothers regarding their political preferences, social and occupational status, education, religious beliefs, etc. These comparisons usually involve comparing dependent marginal tables, not only one-way, but also higher-way marginal tables that involve dependent observations. For example, clustered family data are needed to answer concrete research questions such as

- Are the relative direct influences of religion and social class on political preference the same for the children and their parents?
- Is the agreement in attitudes between fathers and sons of the same size and nature as between mothers and their daughters, and is the agreement less in pairs of opposite sex, i.e., between fathers and their daughters or between mothers and their sons?
- Are sisters more like each other than sisters and brothers?

Standard analysis techniques that ignore the dependencies in the data, i.e., ignore the hierarchical or clustered nature of the data are not appropriate here. Especially if an investigator wants answers to these research questions without at the same time wanting to make assumptions about the nature of the dependencies in the data, marginal-modeling methods provide an excellent way to analyze the family data reckoning with the fact that the observations are dependent.

Marginal modeling is also needed to answer particular kinds of research questions that make use of data that are seemingly not clustered. This happens, for instance, when a political scientist has conducted a one-shot cross-sectional survey based on simple random sampling, in which respondents are asked to state their degree of sympathy for different political parties on a five point scale. To answer the question of whether the distributions of the sympathy scores are the same for all political parties, standard chi-square tests cannot be used because the comparisons of the several one-variable distributions pertain to the same respondents. The data are actually clustered within individuals given the research question of interest, despite the cross-sectional design and the simple random sampling scheme.

Many other research questions in similar situations require marginal modeling. The same political scientist may also want to measure political interest by means of several items in the form of seven-point rating scales. These items are supposed to form a summated (Likert) scale. In its strictest form, it is assumed in Likert scaling that the items are parallel measurements of the same underlying construct, having independently distributed error terms with identical error variances (for a more precise technical definition of parallel measurements, see Lord & Novick, 1968). This strict measurement model implies that all marginal distributions of all items are the same, as are all pairwise associations. Again, marginal modeling is needed to test such implications.

In the following two subsections of this chapter, empirical illustrations will be provided for exactly the above kinds of research topics, showing how to translate these questions into the language of loglinear modeling. In the next chapter, the same

data and general research questions will be used but then formulated in terms of nonloglinear marginal models.

### 2.2.2 Comparing One Variable Distributions

#### Comparing One Variable Distributions in the Whole Population

To gain a practical insight into the nature of marginal modeling, the best starting point is the comparison of a number of simple one-way marginals. Our example concerns a study into the way people perceive their body. A group of 301 university students (204 women and 97 men) answered questions about their degrees of satisfaction with different parts or aspects of their body by completing the *Body Esteem Scale* (Franzoi & Shields, 1984; Bekker, Croon, & Vermaas, 2002). This scale consisted of 22 items (not counting the items concerning gender-specific body parts), seven of which will be considered here. These seven items loaded highest on the first unrotated principal component, with loadings higher than .70. Principal component analysis was used to discover whether the separate expressions of satisfaction with the different body aspects can be seen as just an expression of the general underlying satisfaction with the body as a whole or whether more underlying dimensions are needed (for the interested reader: two rotated factors were needed to explain the correlations among all the 22 items, one having to do with the general appearance of the body and the other with the satisfaction with the parts of one's face; the items chosen here all belong to the first factor). Such dimensional analyses tell the researcher how strongly satisfaction on one particular item goes together with satisfaction with other parts or aspects of the body. However, even if a correlation between two particular items is positive (and it turned out that the correlations among all 22 items had positive values), it does not follow automatically that the respondents react in the same way to these items in all respects. Despite the positive correlation, people may be much more satisfied on average with one part of their body than with another, or the disagreement among the respondents regarding the satisfaction with one bodily aspect may be much larger than the disagreement with another one. Such differences may be important and may reveal, in addition to correlational analyses, relevant details about how the body is perceived. To investigate these kinds of differences, one must compare the overall reactions of the respondents to the different body parts, in other words, compare the marginal distributions. For the selected seven items, the marginal distributions are shown in Table 2.5 (the complete seven-way table can be found on the website mentioned in the last chapter). The response categories are 1 = very dissatisfied; 2 = moderately dissatisfied; 3 = slightly satisfied; 4 = moderately satisfied; and 5 = very satisfied.

The items in Table 2.5 will be denoted by  $I_i$ :  $I_1$  refers to item Thighs,  $I_2$  to Build, etc. To see whether the seven response distributions differ significantly from each other, one could start from the marginal homogeneity model (MH) that states that the marginal distributions of the variables  $I_i$  are the same for all  $i$ , i.e.,

$$\pi_{i+++++}^{I_1 I_2 I_3 I_4 I_5 I_6 I_7} = \pi_{+i++++}^{I_1 I_2 I_3 I_4 I_5 I_6 I_7} = \dots = \pi_{+++++i}^{I_1 I_2 I_3 I_4 I_5 I_6 I_7} = \pi_{++++++i}^{I_1 I_2 I_3 I_4 I_5 I_6 I_7}$$

**Table 2.5.** *Body Esteem Scales*

	1. Thighs	2. Body Build	3. Buttocks	4. Hips	5. Legs	6. Figure	7. Weight	MH
1	22	10	22	22	18	15	20	21.12
2	67	45	59	51	57	45	48	52.08
3	79	78	93	88	79	70	74	76.32
4	105	127	95	111	110	140	104	110.56
5	28	41	32	29	37	31	55	40.91
Total	301	301	301	301	301	301	301	301.00
Mean	3.17	3.48	3.19	3.25	3.30	3.42	3.42	3.33
SD	1.099	1.010	1.093	1.075	1.093	1.024	1.152	1.124

*Notes: See text for explanation. Source: Franzoi and Shields, 1984*

for all  $i$ . When it does not cause any misunderstandings, a shorter and simpler notation will be used, omitting the variables over which the multidimensional table is marginalized:

$$\pi_i^{I_1} = \pi_i^{I_2} = \pi_i^{I_3} = \pi_i^{I_4} = \pi_i^{I_5} = \pi_i^{I_6} = \pi_i^{I_7}.$$

To test this MH model, application of the standard chi-square test to Table 2.5 is not appropriate since it is not an ordinary two-way contingency table. The columns of table  $SI$  (*Satisfaction*  $\times$  *Item*) contain the marginal score distributions for the seven items that are all based on the same sample of respondents. Nevertheless, we will often refer to such a table as an  $SI$  table for the sake of obtaining a simpler notation and a much simpler indication of relevant models. The MH model for Table 2.5 must be fitted using the marginal-modeling methods that are explained in the following section (and implemented in the programs described in the last chapter). The MH model turns out to fit badly ( $G^2 = 55.76$ ,  $df = 24$ ,  $p = .000$ ,  $X^2 = 45.42$ ) and it must be concluded that the seven marginal item distributions are not identical.

To investigate the differences among the marginal distributions, several paths may be followed. First, adjusted or standardized residual frequencies can be calculated comparing the observed and expected frequencies. The expected frequency distribution for all items under marginal homogeneity is also reported in Table 2.5 in the last column labeled MH. Note that the estimated frequencies under MH are not simply the average of the seven column frequencies. Among other things, the adjusted residual frequencies (not reported here) clearly indicate that more people are dissatisfied with their Thighs and Buttocks than estimated under the MH model, while more people are satisfied with their general Build than expected under MH. None of the residuals for Legs were significant.

Another way of investigating the differences among the marginal item distributions is to compare certain aspects of the marginal item distributions, e.g., the means or standard deviations. These two characteristics are reported in the last two rows



**Table 2.6.** *Body Esteem Scales.* Loglinear parameters  $\hat{\lambda}_{ij}^{SI}$  and their standard errors for data in Table 2.5

	1. Thighs	2. Build	3. Buttocks	4. Hips	5. Legs	6. Figure	7. Weight
1	.181 (.123)	-.482 (.167)	.167 (.109)	.196 (.102)	-.013 (.121)	-.010 (.120)	.051 (.141)
2	.213 (.090)	-.060 (.109)	.072 (.095)	-.045 (.092)	.059 (.096)	-.083 (.102)	-.155 (.107)
3	-.039 (.085)	.074 (.090)	.110 (.085)	.084 (.078)	-.032 (.086)	-.058 (.088)	-.139 (.099)
4	-.096 (.069)	.220 (.070)	-.210 (.073)	-.026 (.061)	-.042 (.070)	.294 (.067)	-.140 (.079)
5	-.258 (.100)	.249 (.092)	-.139 (.098)	-.208 (.096)	.028 (.093)	-.054 (.100)	.383 (.098)

of Table 2.5. However, their comparisons involve nonloglinear marginal models and will be dealt with in the next chapter.

Finally, within the loglinear context, the differences among the marginal item distributions can be described by means of the loglinear parameters. Essentially, the two-variable parameters and their standard errors in saturated model  $\{SI\}$  applied to Table 2.5 are estimated, but then in the correct way by taking the dependencies among the observations into account by means of marginal model estimation procedures (see also below). The results in terms of the two-variable parameters  $\hat{\lambda}_{ij}^{SI}$  are reported in Table 2.6, with their respective standard errors between parentheses.

The  $\hat{\lambda}_{ij}^{SI}$  estimates provide a detailed description of the differences among the marginal distributions. In agreement with the adjusted residuals, they show that there is a relative overrepresentation of dissatisfied people for Thighs and Buttocks (and a corresponding underrepresentation of satisfied people); the opposite tendency is noted for Build. Satisfaction with Legs resembles the average (log)distribution the most. Relative outlying cell frequencies can be seen for Figure and Weight, where the categories *moderately satisfied* and *very satisfied*, respectively, are comparatively strongly overrepresented.

To gain further insight into the properties of marginal modeling, it is useful to compare these results with analyses directly applied to the data in Table 2.5, incorrectly ignoring the dependencies among the column distributions. Table 2.5 is then treated as if it were a normal  $SI$  two-way table and the equality of the item distributions is tested by applying (inappropriately) the standard chi-square test for independence (model  $\{S, I\}$ ) directly to this table. The test results are  $G^2 = 45.50$  for  $df = 24$  ( $p = .005$ ,  $X^2 = 46.41$ ). The test statistic  $G^2$  for the independence hypothesis is smaller than  $G^2$  for MH obtained above which had the value 55.77 and the same degrees of freedom. This was to be expected. The item distributions in Table 2.5 are the result of repeated measurements of the same 301 respondents. As is well-known for repeated measurements, if the dependencies among the repeated measurements are positive, then in general, the standard errors of the estimates will be smaller than for independent observations and consequently test statistics will be larger (Hagenaars, 1990, p. 205-210). For negative dependencies, the situation will be reversed: larger standard errors and smaller test statistics. Because the body items here all correlate positively, the MH test will have more power to detect the differences in the population than the (inappropriately applied) independence test.



As implicitly indicated above, it is important to note explicitly that the two sets of expected frequencies estimated using maximum likelihood methods under the independence hypothesis (as defined above, treating the data incorrectly as coming from independent observations) and under MH will generally not be the same. The estimated frequencies for each item distribution under MH in Table 2.5, Column MH result in the following proportions for the satisfaction categories 1 through 5: .070, .173, .254, .367, and .136. However, the corresponding estimated proportions stemming from the (inappropriate) application of model  $\{S, I\}$  directly to Table 2.5 are: .061, .177, .266, .376, and .120. The latter distribution simply results from adding all cell frequencies in a particular row of Table 2.5 and dividing it by the total number of observations (here:  $7 \times 301$ ). In other words, the inappropriate independence model  $\{S, I\}$  reproduces the observed marginal distribution of  $S$  since the marginal distribution of  $S$  is a sufficient statistic for model  $\{S, I\}$  when the observations are independent. However, this is not true when independence in Table 2.5 is tested in the correct way by means of the model. Note that the MH model has no simple sufficient statistics like the ordinary nonmarginal independence model has. Ignoring the dependencies between the observations for the different items may not only distort the chi-square values, but also the estimates of the item distributions. Finally, regarding the application of saturated model  $\{SI\}$  directly to Table 2.5, it does not matter whether we compute the  $\hat{\lambda}_{ij}^{SI}$ -estimates just from Table 2.5 or use marginal-modeling procedures. The two methods yield the same values for the  $\hat{\lambda}_{ij}^{SI}$  estimates, that describe the observed differences between the item distributions. However, the estimated standard errors of the  $\hat{\lambda}_{ij}^{SI}$ -estimates are different for the two procedures. As expected, the inappropriate estimates, obtained directly from Table 2.5 assuming independent observations, are all larger than the ones obtained from the correct marginal-modeling procedures. Most differences are within the range .02–.04, which makes a difference for the significance level of several  $\hat{\lambda}_{ij}^{SI}$  estimates.

### Subgroup Comparisons of One Variable Distributions

The satisfaction with body parts is not only known for the whole sample, but also separately for men and women. As conventional (and scientifically based) wisdom holds, women perceive their body in ways different from men. The observed marginal distributions of satisfaction with the body parts and aspects are shown separately for men and women in Table 2.7. What strikes one immediately in Table 2.7 is that men seem to be more satisfied with their body than women. These and other differences between men and women will be investigated using loglinear marginal models. Most of these models will be indicated by referring to Table 2.7 as if it were a normal table  $GSI$  ( $Gender \times Satisfaction \times Item$ ) without repeating every time that this table is not a normal table and that the models must be tested and its parameters estimated by means of the appropriate marginal-modeling procedures.

As seen above, the MH model had to be rejected for the total group. But perhaps this is because one subgroup (maybe women) expresses different degrees of satisfaction with different body parts, while the other group (maybe men) is equally (dis)satisfied with all body parts. Testing the MH hypothesis among the men yields:

**Table 2.7.** *Body Esteem Scales for Men and Women.* See also Table 2.5

Men							
	1. Thighs	2. Body Build	3. Buttocks	4. Hips	5. Legs	6. Figure	7. Weight
1	1	3	2	2	1	3	4
2	8	9	9	8	6	11	13
3	20	26	25	24	18	18	15
4	47	40	42	48	49	50	35
5	21	19	19	15	23	15	30
Total	97	97	97	97	97	97	97
Mean	3.81	3.65	3.69	3.68	3.90	3.65	3.76
SD	.901	.995	.956	.903	.867	.974	1.064

  

Women							
	1. Thighs	2. Body Build	3. Buttocks	4. Hips	5. Legs	6. Figure	7. Weight
1	21	7	20	20	17	12	16
2	59	36	50	43	51	34	35
3	59	52	68	64	61	52	59
4	58	87	53	63	61	90	69
5	7	22	13	14	14	16	25
Total	204	204	204	204	204	204	204
Mean	2.86	3.40	2.95	3.04	3.02	3.31	3.25
SD	1.050	1.007	1.072	1.088	1.075	1.029	1.117

$G^2 = 29.41$ ,  $df = 24$  ( $p = .205$ ,  $X^2 = 20.66$ ); for women, the results are  $G^2 = 59.95$ ,  $df = 24$  ( $p = .000$ ,  $X^2 = 44.33$ ). Testing the overall hypothesis that there is MH in both (independently observed) subgroups is possible by simply summing the test statistics:  $G^2 = 29.41 + 59.95 = 89.35$ ,  $df = 24 + 24 = 48$  ( $p = .000$ ). This overall hypothesis can also be indicated as conditional independence between S and I in table  $GSI$ , i.e., model  $\{GS, GI\}$  for table  $GSI$ . The test results point out that it is clearly not true that there is MH in both subgroups. Looking at the separate subgroup tests, one might be inclined to conclude that there is definitely no MH among women, but that MH might be accepted for men. Note, however, that there are 204 women but only 97 men. If we had observed the proportional data in a sample of 204 men instead of only 97, given these results the expected value of  $G^2$  would have been  $29.41 \times \frac{204}{97} = 61.85$ , which is about the same value as obtained for the women. In other words, the number of men is too small and the test does not have enough power to definitely draw the conclusion that MH is true for men but not for women.

In light of this, the observed frequencies in Table 2.7 might be taken for granted as the best guesses of the population values, and they might be parameterized by means of the parameters of saturated model  $\{GSI\}$  for table  $GSI$ . These parameters can be used to describe how much more men are (dis)satisfied with their body (parts) than women and how this difference between men and women varies among the

items. It can also be formulated the other way around: how much more satisfied the respondents are with particular body parts than with others and how these item differences vary between men and women.

But before carrying out such detailed descriptions, other models might be considered that are more parsimonious than the saturated model, but less parsimonious than MH for both subgroups. The no three-variable interaction model  $\{GS, GI, SI\}$  for table  $GSI$  is a first interesting candidate. According to this model the items may have different satisfaction distributions, and men may be more or less satisfied than women, but the item differences are the same for men and women and the gender differences are the same for all items. Model  $\{GS, GI, SI\}$  then implies that the odds ratios in subtable  $SI$  for men are equal to the corresponding odds ratios in subtable  $SI$  for women. In other words, the corresponding conditional loglinear parameters for the association between  $S$  and  $I$  in the subgroups Men and Women are all equal:  $\lambda_{ijm}^{SI|G} = \lambda_{ijw}^{SI|G}$  in which the conditional loglinear parameters (Hagenaars, 1990, p. 43-44) are defined as

$$\lambda_{ijk}^{SI|G} = \lambda_{ij*}^{SIG} + \lambda_{ijk}^{SIG}.$$

Formulated from the viewpoint of the relationship between  $G$  and  $S$ , the no three-variable interaction model similarly implies (with obvious notation):  $\lambda_{kij}^{GS|I} = \lambda_{kij'}^{GS|I}$ .

Model  $\{GS, GI, SI\}$  will be treated here as a logit model for the investigation of the effects of  $G$  and  $I$  on  $S$ , conditioning on the observed distribution of  $GI$ . The test results for model  $\{GS, GI, SI\}$  (using the correct marginal-modeling estimation procedures) are:  $G^2 = 40.20$ ,  $df = 24$  ( $p = .020$ ,  $X^2 = 28.66$ ). Where the previous analyses for testing MH among men and women may have led to the conclusion that men are equally (dis)satisfied with their different body parts, while women react differently to different bodily aspects, acceptance of model  $\{GS, GI, SI\}$  would imply that both men and women are differently satisfied with different body parts but that these item differences are the same for both subgroups. It is, however, not clear what to do: reject or accept model  $\{GS, GI, SI\}$  given the  $p$ -value and the rather large discrepancy between  $G^2$  and  $X^2$ . The parameter estimates in model  $\{GS, GI, SI\}$  (not reported here) suggest a linear relationship between Gender and Satisfaction and an *interval by nominal* association between Satisfaction and Item. Degrees of freedom might be gained from imposing these linear restrictions obtaining a more parsimonious no three-variable interaction model. The test outcomes for such a linearly restricted model are  $G^2 = 59.88$ ,  $df = 45$  ( $p = .068$ ,  $X^2 = 43.50$ ). However, one should be careful applying such data dredging procedures, certainly in the light of the  $p$ -values obtained and the discrepancies between test statistics  $G^2$  and  $X^2$ . Actually, more data are needed to arrive at firm conclusions.

The adjusted residuals of model  $\{GS, GI, SI\}$  or, for that matter, the parameter estimates of the saturated model  $\{GSI\}$ , give interesting clues as to why model  $\{GS, GI, SI\}$  fails to fit unequivocally. They show that the biggest discrepancies regarding men's and women's satisfaction are for items Thighs and Build. On average, men are much more satisfied with all their body parts than women. The parameter estimates for the average two-variable relation  $GS$  in model  $\{SGI\}$  for all items go from  $\hat{\lambda}_{m1*}^{GSI} = -.529$  for men in the very dissatisfied category almost linearly to  $\hat{\lambda}_{m5*}^{GSI} = .623$

for men in the very satisfied category; the corresponding estimates for women are .529, and  $-.623$  (remembering effect coding is being used). From this, the (extreme) odds ratio for the average relation between  $G$  and  $S$  among the body items can be computed:  $\exp(-.529 - .623 - .623 - .529) = \exp(-2.304) = .100$ . The odds that a man is very satisfied with a body part rather than very dissatisfied is ten times ( $1/.100$ ) higher than the corresponding odds for a woman. The three-variable interaction parameter estimates indicate that this average difference between men and women is even very much (and statistically significantly) stronger for Thighs (inverse extreme odds ratio  $1/.0159 = 63.02$ ), while it is very much (and statistically significantly) weaker for Build (inverse extreme odds ratio  $1/.532 = 1.88$ ). Regarding these men-women differences, Buttocks and Hips follow Thighs, although the men-women differences are much smaller for Buttocks and Hips than for Thighs, while Legs, Figure and Weight follow Build, but also with smaller differences.

Finally, as in the analyses for the whole group, it is seen here again that the parameter estimates for the nonsaturated loglinear models when (inappropriately) estimated directly from Table 2.7 differ from the correct estimates, taking the dependencies of the observations into account. Further, the standard errors of the estimates of the loglinear parameters are all smaller when computed in the correct way and, consequently, the test statistics larger. Finally, the observed marginal proportions that are exactly reproduced in the inappropriate analyses assuming independence of the observations are not exactly reproduced by the marginal-modeling procedures.

These consequences and tendencies have been found many times in later analyses and will therefore not always be reported anymore; attention will mainly be paid to exceptions or special cases. One must keep in mind that in the relative simple cases such as discussed here, standard errors will be larger and test statistics smaller, when the correlations between the dependent observations are negative or the data in the joint table lie mainly outside the main diagonal (Hagenaars, 1990, p. 208-209). However, in situations with more complex dependencies patterns, the consequences of incorrectly assuming independent observations may not be this simple (see also Berger, 1985; Verbeek, 1991).

### 2.2.3 More Complex Designs and Research Questions

#### Complex Dependency Patterns

The analyses of the previous subsection dealt with data coming from a simple random sampling design and pertained to substantive research questions regarding the comparison of simple (conditional) one-way marginal tables. How to handle more complicated research designs and more complex marginal tables will be illustrated in this subsection. The data that will be used as an example come from the Netherlands Kinship Panel Study (NKPS), a unique in-depth large-scale study into (changing) kinship relationships covering a large number of life domains (Dykstra et al., 2004). NKPS contains several different modules with different modes of data collection. The data used here come from a module in which essentially a random sample from

the Dutch population above 18 years old has been interviewed, as well as one randomly chosen parent of each respondent. Because of selective nonresponse, there are many more women than men in the sample. This will be ignored here, but a possible approach for dealing with nonresponse in marginal models will be briefly discussed in the last chapter. We will also ignore the fact that within the module we use, just one child and one parent is selected from each family, regardless of the size of the (nuclear) family. To get a representative sample of individual family members, weights must be applied to correct for the smaller chances of children from large families to get included in the sample. We will ignore this issue in order to not complicate things further, but discuss it very briefly in the last chapter. Further details on the study's design, fieldwork and nonresponse are provided on the NKPS website, [www.nkps.nl](http://www.nkps.nl).

Among many other things, family members were questioned about their traditional sex role attitude. A scale was constructed from these questions with (mean) scores running from 0 to 4. Here, an index will be used with three categories:

- 1 = less traditional attitude (more in favor of an egalitarian division of tasks between men and women), corresponding with original scale scores between 0 and 1
- 2 = moderately traditional, corresponding to original scale scores between 1 and 2
- 3 = traditional attitude (in favor of a traditional division of labor, such as women taking care of the house and the children, but men working outside the house earning the household's income, etc.), corresponding with original mean scale scores between 2 and 4.

The data for 1,884 families (parent-child pairs) are shown in Table 2.8. There are four variables in this table: variable *P* represents *Sex of the parent* (1 = father, 2 = mother), variable *C* *Sex of the child* (1 = son, 2 = daughter). Variable *A* is the *Sex role attitude of the parent*, whereas variable *B* is the *Sex role attitude of the child*. This table is the joint (or fully) observed table since its entries are based on independently sampled households. Therefore, for many research questions regarding the relationships among the four variables, no special marginal-modeling techniques are required. However, this is certainly not true for many other questions.

A relevant question about these data might be whether there are any overall differences between parents and children with respect to their sex role attitudes. The relevant marginals are shown in the first two columns of Table 2.9. These marginals are obtained by summing the row totals (for the parents' attitude *A*) and summing the column totals (for the children's attitude *B*) in the four subtables of Table 2.8. The first two columns of Table 2.9 will now be denoted as table *TG*, where *T* stands for *Traditionalism* (with three categories) and *G* for *Generation* (parent/child). Given the sampling design of the NKPS data in the module used here, the data in this table *TG* are not independently observed as each parent is coupled with one child (from the same family). Marginal modeling methods must be used to take this dependency (clustering, matching) into account when comparing the overall parent and child distributions.

Complete homogeneity of parents' and children's distributions clearly has to be rejected:  $G^2 = 343.11$ ,  $df = 2$  ( $p = .000$ ,  $X^2 = 297.98$ ). Ignoring the partial

**Table 2.8.** *Traditional sex role attitudes*; source NKPS, see text

<i>P</i> = 1, <i>C</i> = 1: Father–Son				
	<i>B. Child's Attitude</i>			
<i>A. Parent's Attitude</i>	1. Nontrad.	2. Mod. trad.	3. Trad.	Total
1. Nontraditional	37	26	3	66
2. Moderately traditional	60	62	13	135
3. Traditional	19	41	11	71
Total	116	129	27	272
<i>P</i> = 1, <i>C</i> = 2: Father–Daughter				
	<i>B. Child's Attitude</i>			
<i>A. Parent's Attitude</i>	1. Nontrad.	2. Mod. trad.	3. Trad.	Total
1. Nontraditional	101	25	3	129
2. Moderately traditional	108	62	2	172
3. Traditional	26	37	5	68
Total	235	124	10	369
<i>P</i> = 2, <i>C</i> = 1: Mother–Son				
	<i>B. Child's Attitude</i>			
<i>A. Parent's Attitude</i>	1. Nontrad.	2. Mod. trad.	3. Trad.	Total
1. Nontraditional	92	55	5	152
2. Moderately traditional	91	123	18	232
3. Traditional	30	49	23	102
Total	213	227	46	486
<i>P</i> = 2, <i>C</i> = 2: Mother–Daughter				
	<i>B. Child's Attitude</i>			
<i>A. Parent's Attitude</i>	1. Nontrad.	2. Mod. trad.	3. Trad.	Total
1. Nontraditional	204	65	6	275
2. Moderately traditional	222	114	11	347
3. Traditional	63	63	9	135
Total	489	242	26	757

**Table 2.9.** *Traditional sex role attitudes*; marginal distributions; source NKPS, see text

<i>Sex Role Attitude</i>	1. Parent	2. Child	1. Men	2. Women
1. Nontraditional	622	1053	524	1151
2. Moderately traditional	886	722	663	945
3. Traditional	376	109	212	273
Total	1884	1884	1399	2369

dependency and inappropriately applying the standard independence model  $\{T, G\}$  directly to the first two columns of Table 2.9 yields different values for the test statistics:  $G^2 = 284.41$ ,  $df = 2$ ,  $p = .000$ ,  $X^2 = 274.62$ . All adjusted residuals under the MH model (not reported here) are statistically significant. For parents' attitude, the vector with observed marginal proportions in Table 2.9 is  $(.330, .470, .200)$ ; for

children's attitude, it is (.559, .383, .058). Looking at these observed marginal proportional distributions of traditionalism, it is evident that children have much less traditional views on sex roles than their parents. The differences in sex roles attitude between parents and children can be expressed in terms of the loglinear parameters in saturated model  $\{TG\}$  applied to table  $TG$  using marginal-modeling methods. All two-variable parameters are statistically significant and show an almost perfectly linear relationship between  $T$  and  $G$ :  $\hat{\lambda}_{11}^{GT} = -.416$ ,  $\hat{\lambda}_{12}^{GT} = -.050$ , and  $\hat{\lambda}_{13}^{GT} = .466$ . From these estimates, it can be computed that the extreme odds ratio equals  $\exp(-.416 - .466 - .416 - .466) = .171$  (and its inverse  $1/.171 = 5.836$ ): the odds of being traditional ( $T = 1$ ) rather than nontraditional ( $T = 3$ ) are almost six times larger for parents than for children.

Another possibly relevant question that gives rise to a more complex dependency pattern is about the overall differences between men and women. To answer this question, the overall marginal distributions of *Traditionalism* for men and women have to be obtained by summing appropriate row totals and column totals of the subtables in Table 2.8. For example, to obtain the marginal distribution for men, one takes the sum of the row and column totals of subtable 1, the row totals of subtable 2 and the column totals of subtable 3. The resulting marginal distributions are reported in the last two columns of Table 2.9, denoted as table  $TS$ , where  $T$  has three categories as before and  $S$  has two (men, women). Note that now the totals of these two columns are no longer equal to each other and not equal to the 1,884 parent-child pairs. In total, there are  $2 \times 1884 = 3,768$  responses to the questions about sex roles, 1,399 of which were given by men (fathers and sons) and 2,369 by women (mothers and daughters). These answers come partly from matched observations, viz. when originating from the same subtable in Table 2.8, i.e., from father-son or mother-daughter pairs. For the other part, they are independent observations. Again, marginal-modeling methods have to be used to take the (partial) dependency into account.

Marginal homogeneity for men and women in table  $TS$  must now explicitly refer to the equality of the probability distributions because of the different column totals. The test results are  $G^2 = 46.91$ ,  $df = 2$  ( $p = .000$ ,  $X^2 = 46.33$ ). Inappropriately assuming completely independent observations and applying model  $\{T, S\}$  directly to table  $TS$  yields  $G^2 = 45.38$ ,  $df = 2$ ,  $X^2 = 45.11$ . Men appear to have different opinions than women regarding the roles of the sexes. The observed distribution for men is given by vector (.375, .474, .152), for women by (.486, .399, .115). So, men are more traditional regarding sex roles than women. This is confirmed by the (statistically significant) loglinear parameters of saturated model  $\{TS\}$  for table  $TS$ . The relationship between  $S$  and  $T$  is approximately linear:  $\hat{\lambda}_{11}^{ST} = -.161$ ,  $\hat{\lambda}_{12}^{ST} = .055$ , and  $\hat{\lambda}_{13}^{ST} = .106$ . The overall differences in traditionalism between parents and children as found above in table  $TG$  are definitely larger than the overall differences found here between men and women in table  $TS$ . The extreme odds ratio for the relationship between  $S$  and  $T$  in table  $TS$ , computed analogously as above for the relation  $G - T$  equals .534 and its inverse  $1/.534 = 1.873$ : the odds that a man gives a traditional answer ( $T = 3$ ) rather than a nontraditional one ( $T = 1$ ) are almost two times



**Table 2.10.** *Traditional sex role attitudes*; men–women and parents–children marginal distributions; source NKPS, see text

<i>P.</i> Parent's sex		Male			
<i>C.</i> Child's sex		Male		Female	
<i>R.</i> Respondent's status		Parent	Child	Parent	Child
<i>T.</i> Sex role attitude	1. Nontraditional	66	116	129	235
	2. Moderately traditional	135	129	172	124
	3. Traditional	71	27	68	369
	Total	272	272	369	369

  

<i>P.</i> Parent's sex		Female			
<i>C.</i> Child's sex		Male		Female	
<i>R.</i> Respondent's status		Parent	Child	Parent	Child
<i>T.</i> Sex role attitude	1. Nontraditional	152	213	275	489
	2. Moderately traditional	232	227	347	242
	3. Traditional	102	46	135	26
	Total	486	486	757	757

larger than for a woman. It is also possible to test whether the difference in strengths of the marginal association between  $S - T$  and  $G - T$  is statistically significant using marginal-modeling methods; this will not be done here, but a similar research question will be illustrated below.

Given the overall differences in traditionalism between generations, and between men and women, a next logical question is to ask whether or not the generational differences in traditionalism are larger among men than among women; or, formulated the other way around, are the sex differences in traditionalism larger between fathers and mothers than between sons and daughters. To answer this question, all row and column totals of the subtables in Table 2.8 have to be considered. They have been put together in Table 2.10 in the form of a *PCRT* table (where the symbols *P*, *C*, *R*, and *T* are explained in the table). Here the observations are again partially dependent. The loglinear models of interest can most easily be formulated in terms of the variables in Table 2.10, using the short-hand notation for hierarchical models. All models will be logit models for the investigation of the effects on *T*, conditioning on the distribution of *PCR*.

It was found above that the overall differences between parents and children (now variable *R*) concerning their sex roles attitudes (variable *T*) were very large. But how large are these differences when we control for the sex of parents and children using table *PCRT*? The most parsimonious model in this respect is the model in which there are no differences left, that is, model  $R \perp\!\!\!\perp T | PC$ , which states that *R* and *T* are conditionally independent given *P* and *C*. This model is equivalent with loglinear



model  $\{PCR, PCT\}$  for Table 2.10, and identical to the hypothesis of simultaneous MH in each of the subtables of Table 2.8. For each of the subtables of Table 2.8, the hypothesis of marginal homogeneity has to be rejected (all  $p = .000$ ), as well as the MH hypothesis for all four subgroups simultaneously:  $G^2 = 354.89$ ,  $df = 8$  ( $p = .000$ ,  $X^2 = 303.56$ ). The next logical step is then to ask whether the apparently existing marginal differences between parents and children are the same in all four subtables in Table 2.8. In loglinear terms, is model  $\{PCR, PCT, RT\}$  for Table 2.10 true in the population? However, this model has to be rejected too:  $G^2 = 30.00$ ,  $df = 6$  ( $p = .000$ ,  $X^2 = 29.44$ ).

In order to find out why model  $\{PCR, PCT, RT\}$  did not fit the data, the parameters for the effects on  $T$  in saturated model  $\{PCRT\}$  for Table 2.10 were estimated. The first striking finding was that all loglinear parameter estimates  $\hat{\lambda}$  that had  $P$  and  $T$  among their superscripts ( $PT, PCT, PRT, PCRT$ ) had insignificant and very small values: all absolute  $\hat{\lambda}$ -values were smaller than .05. This means that there are no effects at all of  $P$  on  $T$  and it must be concluded, somewhat unexpectedly, that fathers and mothers do not differ in their attitudes on sex roles in model  $\{PCRT\}$  for Table 2.10, despite the overall differences between men and women found above. Apparently, these sex differences only apply to the children, not to the parents. Second, there was a substantial main effect of  $R$  on  $T$  in model  $\{PCRT\}$ . It is approximately a linear effect (as, by the way, all direct effects on  $T$  in model  $\{PCRT\}$  are). The extreme odds ratio for the relationship  $R - T$  in model  $\{PCRT\}$ , i.e., for the differences between parents and children regarding the odds nontraditional ( $T = 1$ ) versus traditional ( $T = 3$ ) is .157 and its inverse is 6.366. This effect is just a little bit stronger than the corresponding overall effect  $G - T$  discussed above (that was .171 with its inverse 5.836). But there is also a significant and non-negligible three-variable interaction term  $CRT$ . Because of this interaction term, when sons ( $C = 1$ ) are being interviewed, the extreme conditional odds ratio for the differences in traditionalism between parents and children becomes .264 and its inverse is 3.790, while for daughters ( $C = 2$ ), the corresponding extreme conditional odds ratio equals .094 and its inverse is 10.693. In sum, parents are generally more traditional than their children, but sons depart substantially less from their parents than daughters do.

Summing up, there are no differences whatsoever with regard to traditional attitudes towards sex roles between fathers and mothers. Boys and girls, on the other hand differ; boys being more traditional than girls. The largest attitude differences were found for *Generation*: children are much less traditional than their parents, and this is especially true for daughters, but less so for sons.

So far, all marginal models for the NKPS data concerned the comparison of the distributions of the one characteristic attitude towards sex role. However, in NKPS, the respondents were not only asked about their attitudes regarding sex roles but also regarding marriage: to what extent did the respondents feel that marriage is a sacred institute? This attitude was measured by items such as 'having sex before marriage is forbidden', and 'marriage among homosexuals is not allowed'. Traditionalism concerning marriage, which was also originally expressed in terms of mean scale scores,

**Table 2.11.** *Traditional sex role and Marriage attitudes*; marginal distributions for Parents and Children; source NKPS, see text

<i>I.</i> Item <i>R.</i> Respondent's Status	1. Sex Role		2. Marriage	
	1. Parent	2. Child	1. Parent	2. Child
<i>T. Traditional attitude</i>				
1. Nontraditional	622	1053	783	1310
2. Moderately traditional	886	722	896	501
3. Traditional	376	109	205	73
Total	1884	1884	1884	1884

was coded in three categories in the same way as the categorization of the variable sex role attitude.

A relevant research question then might be whether or not parents and children differ in the same way regarding both characteristics, viz. traditionalism regarding sex roles and marriage. The marginal one-variable distributions of both characteristics for parents and children were formed from the full table (not presented here, but see our website) and shown in Table 2.11.

The observations in the different columns in Table 2.11 are not only dependent because of the partial matching between particular parents and children, but also because the data for marriage traditionalism have been obtained from exactly the same respondents as the data on sex role traditionalism. Estimation and testing procedures have to take these complex patterns of dependencies into account. Table 2.11 will be treated as an *IRT* table with variable *T* now representing *Traditionalism* (regarding sex roles and marriage). The loglinear models considered for the data in Table 2.11 will be logit models for the effects on *T*, conditioning on the observed frequencies for marginal table *IR*.

The hypothesis that there are no parent-child and item differences, in other words, that all column distributions in Table 2.11 are homogeneous is represented by model  $\{IR, T\}$  for Table 2.11. In the less restrictive model  $\{IR, IT\}$ , the distributions of *T* are allowed to be different for the two items, but not between parents and children. Assuming homogeneous distributions for the two items but different distributions for parents and children leads to model  $\{IR, RT\}$ . Finally, the hypothesis that there are both item and parent-child differences regarding *T*, but no special three-variable interactions, is represented by model  $\{IR, IT, RT\}$ .

Note that these models bear strong resemblances to traditional MANOVA models or ANOVA models for repeated measures, as were the models for the body items data in Table 2.7; see Chapters 3 and 5 for more MANOVA-like analyses.

However, all these models for Table 2.11 have to be rejected; all *p*-values are  $p = .000$ . That leaves us with the saturated model  $\{IRT\}$  for Table 2.11. However, in this saturated model, the three variable interaction effect is very small (although statistically significant) and does not lead to really different conclusions from the no three-variable interaction model  $\{IR, IT, RT\}$ . The outcomes in model  $\{IR, IT, RT\}$

(not presented here) indicate that parents are substantially more traditional than their children regarding both characteristics, and that both parents and children have a bit more traditional views on sex roles than on marriage.

The complex patterns of dependencies in the observations in Table 2.11 as a consequence of the matched parent-child relation, and of the repeated items, led to somewhat unexpected outcomes for standard errors and test statistics when the above correct marginal procedures are compared with the inappropriate analyses ignoring the dependencies. For example, the  $G^2$  statistics for models  $\{IR, T\}$  and  $\{IR, IT\}$  in Table 2.11 are smaller when calculated in the correct manner than when inappropriately applied assuming independent observations. For models  $\{IR, RT\}$  and  $\{IR, IT, RT\}$  the opposite is true. In the same vein, the standard errors for some of the parameter estimates in the saturated model are smaller but also larger for some parameters when estimated in the correct way than when estimated incorrectly assuming independent observations. This is different from the patterns found so far. It turns out that predictions about the consequences of ignoring the (complex) dependencies between the observations may not be of a simple nature: it is not guaranteed at all that one gains statistical power by taking the dependencies into account; one might as well lose power. As was seen in many complex analyses, even when expecting 'positive' dependencies, ignoring the dependencies in the observations may lead to smaller standard errors and larger test statistics, but also to the opposite situation: larger standard errors and smaller test statistics.

### Associations Among Variables

So far, only one-way marginals have been considered, either for the sample as a whole or for subgroups, and either for one or even more characteristics. But multiway marginals may be at least as interesting for researchers. Comparisons of more complex multiway marginals will be discussed in Chapters 4, 5, and 6. Now, the basic approach will be outlined by means of a simple example showing how to conduct analyses of associations using two-way marginals. As a first example, the data in Table 2.8 will be used in which the relationship was shown between Parents' attitudes ( $A$ ) and Children's attitudes towards sex roles ( $B$ ) for different subgroups. One may wonder whether the associations  $AB$  for homogeneous pairs (father-son and mother-daughter) are stronger than for nonhomogeneous pairs (father-daughter and mother-son). If the data in Table 2.8 had come from a design different from NKPS, in which the information for the particular combinations of the four relevant respondents, viz. father, mother, son, and daughter had all been collected within the same family, marginal-modeling methods would have been needed to test such a hypothesis. However, given the module of NKPS that is used here, in which one respondent (child) is coupled with one parent, the four subgroups in Table 2.8 have been independently observed and regular loglinear modeling can be used. The hypothesis that the relationship  $AB$  expressed in terms of odds ratios is the same in all four subgroups is identical to applying model  $\{PCA, PCB, AB\}$  directly and in the standard way to Table 2.8. This hypothesis need not be rejected:  $G^2 = 12.26$ ,  $df = 12$  ( $p = .414$ ,  $X^2 = 12.26$ ). Consequently, the idea that the association  $AB$  might be stronger in

**Table 2.12.** Association between *Sex role* and *Marriage attitudes* for Parents and Children; source NKPS, see text. Response categories for Marriage and Sex Role: 1 = nontraditional, 2 = moderately traditional, 3 = traditional

<i>R. Respondent's Status</i>		Parent			Child		
<i>B. Marriage</i>		1	2	3	1	2	3
<i>A. Sex Role</i>	1	459	152	11	923	120	10
	2	251	542	93	345	339	38
	3	73	202	101	42	42	25

the homogeneous subgroups (mother-daughter and father-son combinations) than in the nonhomogeneous subgroups (father-daughter and mother-son combinations) is not accepted. There is the same strong, positive, and statistically significant relation between *A* and *B* within each subtable; the relationship is also approximately linear and the deviations from linearity are not significant (precise results not reported here). The test results for a conditional test of the linear restrictions for *AB*, assuming a linear  $\times$  linear (or uniform) association against model  $\{PCA, PCB, AB\}$  without the linear restrictions for *AB* are  $G^2 = 1.85$ ,  $df = 3$ , ( $p = .605$ ).

A research question that does require marginal modeling with the NKPS data would be whether or not the association between the sex role and the marriage attitudes are different for parents and children. One may postulate, for example, that different specific attitudes are more crystalized into a consistent attitude system among parents than among children. The full data set is presented on the book's webpage; the relevant marginal data are displayed in Table 2.12, which will be treated as an *RBA* table.

The null hypothesis that the association between *A* and *B* is the same for parents and children corresponds to model  $\{RB, RA, AB\}$  for Table 2.12. This model fits the data excellently:  $G^2 = 3.791$ ,  $df = 4$  ( $p = .435$ ,  $X^2 = 3.802$ ). Assuming independent observations and applying this model directly to the data in Table 2.12 yields  $G^2 = 3.891$ ,  $df = 4$ ,  $p = .421$ ,  $X^2 = 3.904$ . There is no reason to assume that the relationship between the two attitudes is stronger for the parents than for the children. The common relationship between sex role and marriage attitudes is very strong and monotonically increasing. All local log odds ratios in the common  $3 \times 3$  table *AB* are positive. The extreme odds ratio involving cells 11, 13, 31, and 33 equals 52.47. The relationship is not strictly linear since the linear  $\times$  linear or uniform association model has to be rejected.

The above illustrations should have made clear when and how substantive research questions involve the comparison of marginal tables and how they can be translated into the language of loglinear models for marginal tables. Later chapters will exemplify still more and more complicated research questions. But first, attention must be paid to the central question of how the maximum likelihood estimates for loglinear marginal models can be obtained.

## 2.3 Maximum Likelihood Inference for Loglinear Marginal Models

Maximum likelihood inference for marginal models requires that the cell probabilities of the joint table are estimated under the constraints imposed by the marginal model. The main difficulty in fitting and testing marginal models, compared to ordinary loglinear models, is the dependency of the observations. The estimation procedure discussed here is Bergsma's (1997) modification of the method of Lang and Agresti (1994). Fitting marginal models generally requires numerical procedures since closed form expressions are usually not available. Before the details of the proposed estimation procedure are given, several sampling methods that are often used and that are appropriate for maximum likelihood inference are discussed.

### 2.3.1 Sampling Methods

Let  $\pi_i$  represent the probability of observing response pattern  $i$  pertaining to a set of categorical variables for a randomly selected respondent from the population. Each response pattern  $i$  defines a cell in the multidimensional contingency table that contains the observed frequencies  $n_i$ . Let  $N$  be the total number of respondents and  $I$  the total number of cells in the table. It is commonly assumed in social and behavioral science research that the observed frequency distribution is given by the multinomial distribution

$$\Pr(n_1, \dots, n_i, \dots, n_I) = \prod_i \binom{N}{n_i} \pi_i^{n_i}.$$

Most researchers take it for granted that the multinomial distribution is the appropriate sampling distribution of the frequencies in a contingency table. However, one should keep in mind that this is only true if several, not always trivial, conditions are satisfied. For example, the sample size  $N$  should be fixed in advance, and should not depend on other aspects of the sampling process. Further, the respondents are supposed to be sampled independently from the population with replacement implying that the theoretical cell probabilities  $\pi_i$  remain constant during the sampling process. If these conditions are not satisfied, the multinomial assumption is not valid and other sampling distributions apply. If sampling is without replacement, the observed frequencies follow a hypergeometric distribution. If the sample size is not fixed in advance, but depends on the number of times a certain event (called a 'success') occurs, the negative binomial distribution is more appropriate. The full information ML estimation procedure for fitting marginal models and the associated statistical procedures discussed in this book are appropriate for multinomially distributed frequencies. They are, in general, not appropriate for the hypergeometric and negative binomial distributions. These less well-known theoretical probability distributions for observed frequencies will not be discussed further in this book. However, unless stated otherwise, the estimation and test procedures developed here remain valid for two other sampling distributions: the product multinomial and the Poisson distribution.

Sometimes the entire population is stratified before sampling, using a stratifying variable  $S$ . If parameters of different subgroups are of interest, the researcher may consider taking a fixed number of subjects from each subgroup. The numbers  $N_k$  of observations in each stratum are fixed in advance. Let  $n_{ik}$  be the number of observations from stratum  $k$  in cell  $i$ , assuming  $K$  strata. Then, the joint distribution of all cell frequencies is often assumed to follow the product multinomial distribution

$$\Pr(n_{11}, \dots, n_{ik}, \dots, n_{IK}) = \prod_{k=1}^K \prod_{i=1}^I \binom{N_k}{n_{ik}} \pi_{ik}^{n_{ik}}.$$

The product multinomial distribution is applicable when the sampling within each stratum satisfies the conditions stated above for the multinomial distribution, and additionally when the sampling from different strata occurs independently.

In other applications, it is not the number of observations that is fixed in advance but some other aspect of the sampling process such as the total observation time. In that case, the observed frequencies may follow a Poisson distribution

$$\Pr(n_1, \dots, n_i, \dots, n_I) = \prod_{i=1}^I \frac{\mu_i^{n_i} e^{-\mu_i}}{n_i!}$$

with expected frequencies  $\mu_i$  for  $i = 1, \dots, I$ . The Poisson distribution can be used when the events that are counted occur randomly over time or space with outcomes in disjoint periods or regions independent of each other. If the rate of occurrence of an event is the parameter of interest, such as the number of pedestrians passing a shopping street per hour, of the above schemes only Poisson sampling would be appropriate.

### 2.3.2 Specifying Loglinear Marginal Models by Constraining the Cell Probabilities

In order to describe the ML estimation procedure, it is useful to specify marginal models in matrix notation and define the loglinear models in the form of restrictions on the cell probabilities rather than in terms of loglinear parameters. It will be shown below why this is useful and how this can be done. The notation that will be used is an adaptation of the notation proposed by Grizzle et al. (1969).

A vector of loglinear marginal parameters  $\phi$  can generally be written in the form

$$\phi(\pi) = C' \log A' \pi,$$

where  $\pi$  is a vector of cell probabilities, and  $A$  and  $C$  are matrices of constants. The basic principle of this representation will be illustrated by means of a few simple examples. But first, it will be made clear what it means when a scalar function  $f(x)$  is applied to a vector of values. Let  $f(x)$  be a function of a scalar variable  $x$  such as, for example,  $f(x) = \exp(x)$  or  $f(x) = \log(x)$ . If this function is applied to a vector of values like

$$x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix},$$

the result is another vector with the values of  $f(x_i)$  as its elements:

$$f(x) = \begin{pmatrix} f(x_1) \\ f(x_2) \\ f(x_3) \end{pmatrix}.$$

This definition can easily be extended to vectors containing arbitrary numbers of elements.

*Example 1: Marginal Homogeneity in a  $3 \times 3$  Table*

Suppose a categorical variable with three categories is observed at two time points. The theoretical cell probabilities for the data from this simple panel study can be written in a  $3 \times 3$  matrix

$$\Pi = \begin{pmatrix} \pi_{11} & \pi_{12} & \pi_{13} \\ \pi_{21} & \pi_{22} & \pi_{23} \\ \pi_{31} & \pi_{32} & \pi_{33} \end{pmatrix}.$$

The rows of this matrix correspond to the first measurement and its columns correspond to the second one. For further use, this matrix will have to be written as vector. In this book we decided to vectorize a matrix row-wise so that when the elements of a matrix are written in vector form, the last index changes the fastest. Hence, for matrix  $\Pi$  above,

$$\pi = \text{vec}(\Pi) = \begin{pmatrix} \pi_{11} \\ \pi_{12} \\ \pi_{13} \\ \pi_{21} \\ \pi_{22} \\ \pi_{23} \\ \pi_{31} \\ \pi_{32} \\ \pi_{33} \end{pmatrix}.$$

Note that our definition of the vectorization operations differs from the  $\text{vec}$  operation defined in textbooks on linear algebra as Searle (2006) and Schott (1997) where vectorization is carried out column-wise. This vectorization operation can also be applied to general multidimensional arrays, not just to two-dimensional matrices. Whenever such a multidimensional array is vectorized, its last dimension changes the fastest and its first dimension changes the slowest. So, for a  $2 \times 2 \times 2$  array  $F$  with entries  $f_{ijk}$ , one has

$$\text{vec}(F) = \begin{pmatrix} f_{111} \\ f_{112} \\ f_{121} \\ f_{122} \\ f_{211} \\ f_{212} \\ f_{221} \\ f_{222} \end{pmatrix}.$$

Going back to our example, the univariate marginals of matrix  $\Pi$  are

$$\pi_{i+} = \sum_j \pi_{ij}$$

and

$$\pi_{+j} = \sum_i \pi_{ij}.$$

These marginals can be written in the vector

$$\begin{pmatrix} \pi_{1+} \\ \pi_{2+} \\ \pi_{3+} \\ \pi_{+1} \\ \pi_{+2} \\ \pi_{+3} \end{pmatrix}.$$

Now, consider the following  $6 \times 9$  matrix:

$$A' = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \end{pmatrix}.$$

Then it is easy to see that

$$\begin{pmatrix} \pi_{1+} \\ \pi_{2+} \\ \pi_{3+} \\ \pi_{+1} \\ \pi_{+2} \\ \pi_{+3} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} \pi_{11} \\ \pi_{12} \\ \pi_{13} \\ \pi_{21} \\ \pi_{22} \\ \pi_{23} \\ \pi_{31} \\ \pi_{32} \\ \pi_{33} \end{pmatrix} = A' \pi.$$



Premultiplication of  $\pi$  by matrix  $A'$  yields the appropriate marginal distributions, and  $\log(A'\pi)$  is then the vector containing the logarithms of those marginal probabilities:

$$\log(A'\pi) = \begin{pmatrix} \log(\pi_{1+}) \\ \log(\pi_{2+}) \\ \log(\pi_{3+}) \\ \log(\pi_{+1}) \\ \log(\pi_{+2}) \\ \log(\pi_{+3}) \end{pmatrix}.$$

Under the marginal homogeneity hypothesis, it is assumed that  $\pi_{i+} = \pi_{+i}$  for all categories of the response variable. It then also follows that  $\log(\pi_{i+}) = \log(\pi_{+i})$ , implying that the six entries of the vector  $\log(A'\pi)$  are functions of only three parameters (or: two independent ones, see below). Then, the matrix  $X$  is defined as

$$X = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

where the design matrix and the effects are arbitrarily expressed in terms of a dummy-variable-like notation rather than effect coding. The hypothesis of marginal homogeneity can now be represented as

$$\begin{pmatrix} \log(\pi_{1+}) \\ \log(\pi_{2+}) \\ \log(\pi_{3+}) \\ \log(\pi_{+1}) \\ \log(\pi_{+2}) \\ \log(\pi_{+3}) \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix},$$

or more concisely as

$$\log(A'\pi) = X\beta,$$

with the vector  $\beta$  containing the three unknown parameters  $\beta_j$  for  $j = 1, 2, 3$ . This equation provides a parametric representation of the marginal homogeneity hypothesis, with  $\beta_j$  being the (unknown) logarithm of the sum of the cell probabilities in the  $j$ -th row and in the  $j$ -th column of matrix  $\Pi$ .

A parameter-free representation of the marginal homogeneity model is obtained by noting that it implies the following constraints on the marginal probabilities

$$\begin{aligned} \log(\pi_{1+}) &= \log(\pi_{+1}) \\ \log(\pi_{2+}) &= \log(\pi_{+2}) \\ \log(\pi_{3+}) &= \log(\pi_{+3}). \end{aligned}$$

Since the cell probabilities in  $\Pi$  sum to 1, only two (e.g., the first two) constraints of the three given here have to be considered. Now, define the  $2 \times 6$  matrix  $B'$  as

$$B' = \begin{pmatrix} 1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & 0 \end{pmatrix},$$

then

$$\begin{pmatrix} \log(\pi_{1+}) - \log(\pi_{+1}) \\ \log(\pi_{2+}) - \log(\pi_{+2}) \end{pmatrix} = B' \log(A' \pi).$$

Marginal homogeneity is now equivalent to  $B' \log(A' \pi) = 0$ . Here we have a representation of the marginal homogeneity model that does not contain any parameter, but is completely formulated in terms of constraints on the cell probabilities in matrix  $\Pi$ . The hypothesis of marginal homogeneity can now be tested by first determining the maximum likelihood estimates of the cell probabilities under the constraints implied by the model, and then testing whether or not this restricted model provides a significantly worse fit than the unconstrained model. In this testing procedure, no unknown parameters (apart from the cell probabilities) will be estimated.

*Example 2: Independence in a  $3 \times 3$  Table*

In this second example, it will be illustrated how the simple independence model in a two-dimensional table can also be cast in the form of a parameter-free model (switching to a nonmarginal model for our explanations). We return to the  $3 \times 3$  table  $\Pi$  introduced in the previous example. The loglinear model representing independence of the row and column variable is given by

$$\log(\pi_{ij}) = \lambda_{**}^{RC} + \lambda_{i*}^{RC} + \lambda_{*j}^{RC}.$$

Now, take matrix  $C$  as the  $9 \times 9$  identity matrix. Further, define

$$X = \begin{pmatrix} 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 \end{pmatrix},$$

which is the design matrix of the model (arbitrarily using dummy rather than effect coding). Then, our loglinear model can be written as

$$C' \log(\pi) = \log(\pi) = X \lambda,$$

with vector  $\lambda$  containing the loglinear parameters. As is customary in the discussion of loglinear models, the symbol  $\lambda$  is used (rather than  $\beta$ ) to represent the unknown loglinear parameters.

In order to derive a parameter-free representation of the same model, the concept of the null space of a matrix has to be introduced. Take the matrix  $X$  as defined above. This matrix is of the order  $9 \times 7$  and its columns define a vector space whose elements are the linear combinations of the columns of  $X$

$$\mathcal{V} = \{y : y = Xw\}$$

for all weight vectors  $w$  consisting here of seven arbitrary weights  $w_k$ . The vectors  $y$  in  $\mathcal{V}$  contain nine elements. Since the vector space  $\mathcal{V}$  is generated by the columns of matrix  $X$ , it is often called the column space of  $X$ .

Do we really need all columns of  $X$  to generate this vector space? The answer to this question depends on the column rank of  $X$ . The matrix  $X$  is of full column rank if the zero weight vector  $w = 0$  is the only weight vector for which  $Xw = 0$ . When we can find a nonzero vector  $w$ , i.e., a vector with a least one element different from zero, for which  $Xw = 0$ , the matrix is of deficient column rank. In that case, there exists a linear dependency among the columns of matrix  $X$ . In the present example, twice the first column of  $X$  is the sum of the other six columns, implying that for the nonzero weight vector  $w'_1 = (-2, 1, 1, 1, 1, 1, 1)$ , we have  $Xw_1 = 0$ . Moreover, the sum of columns 2 to 4 is equal to the sum of columns 5 to 7. Hence, we have also  $Xw_2 = 0$  for  $w'_2 = (0, 1, 1, 1, -1, -1, -1)$ . For the present matrix  $X$ , one can prove that only two different linear dependencies exist among its columns. These linear dependencies allow us to express two columns of  $X$  as linear combinations of the other columns. For example, for the first ( $x_1$ ) and second column ( $x_2$ ), we can write

$$x_1 = x_5 + x_6 + x_7$$

and

$$x_2 = -x_3 - x_4 + x_5 + x_6 + x_7 ,$$

showing that arbitrary elements of  $\mathcal{V}$  can be generated by a particular selection of five columns of  $X$ . A set of linearly independent vectors that generate a vector space is called a ‘basis’ of the vector space. The vectors in a vector space can be written as linear combinations of the elements in its basis. Bases of vector spaces are not uniquely defined, but the number of generating vectors in them is: the number of vectors in a basis is called the dimension of the vector space. The vectors defining a basis for a vector space will be written column-wise in the matrix  $X_B$ . Removing any generating vector from a basis transforms the given vector space into a different one of lower dimensions. If matrix  $X$  is of full column rank, there exist no linear dependencies among its columns and we need all its columns to generate vector space  $\mathcal{V}$  with its dimension equal to the number of columns of  $X$ . If there exist linear dependencies among the columns of  $X$ , the dimensionality of  $\mathcal{V}$  is equal to the number of linearly independent columns of  $X$ , and its dimension is smaller than the number of columns of  $X$ . See Schott (1997) for an overview of the concepts of linear algebra that are relevant for statistics.

To show that a model can be defined in terms of (restrictions on) its parameters but also (as is true in our marginal-modeling approach) in terms of restrictions on the

cell probabilities, the concept of a null space is needed. The null space  $\mathcal{N}$  of  $X$  is a different vector space that can be associated with matrix  $X$ . It is defined as

$$\mathcal{N} = \{y : y'X = 0\} .$$

The vector space  $\mathcal{N}$  is the set of all vectors that are orthogonal to the columns of  $X$ . It is also often called the orthocomplement of  $X$ . Being itself a vector space, it can be generated by a set of vectors in one of its bases. One can prove that the dimensionality of the null space and the column space of a matrix sum to the number of rows of  $X$ . Let  $X_N$  be the matrix containing a basis of the null space of  $X$ , and  $X_B$  the matrix containing a basis of the column space of  $X$ . Then,

$$X_N'X_B = 0 .$$

An important point to realize is that a vector space can be characterized either by specifying a basis  $X_B$ , or by specifying a basis  $X_N$  of its null space:  $y \in \mathcal{V}$  if and only if  $y = X_B w$  for some  $w$ , if and only if  $y'X_N = 0$  (or  $X_N'y = 0$ ). We will discuss in a later paragraph how to construct bases for both vector spaces.

With this knowledge in mind we can go back to the parametric representation of the loglinear model for independence:

$$\log(\pi) = \log(A'\pi) = X\lambda .$$

This relation shows that  $\log(\pi)$  is an element in the vector space generated by the columns of  $X$ . Letting  $X_N$  be the matrix containing in its columns a basis of the null space of  $X$ , it follows that the loglinear model can equivalently be specified in terms of a set of constraints on the cell probabilities:

$$X_N'\log(\pi) = 0 .$$

In the present example,  $\mathcal{V}$  has dimension 5 whereas its null space has dimension 4. The columns of the following matrix give a basis for this null space:

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

With this choice for  $X_N$ , the constraints on the cell probabilities are

$$X'_N \log(\pi) = \log \begin{pmatrix} (\pi_{11}\pi_{33})/(\pi_{13}\pi_{31}) \\ (\pi_{11}\pi_{23})/(\pi_{13}\pi_{21}) \\ (\pi_{11}\pi_{32})/(\pi_{12}\pi_{31}) \\ (\pi_{11}\pi_{22})/(\pi_{12}\pi_{21}) \end{pmatrix} = 0 .$$

Independence of the row and column variable in our  $3 \times 3$  table is equivalent to constraining four independent log odds ratios to be equal to zero.

*Example 3: Equality of Local Odds in a  $3 \times 3$  Table*

In a  $3 \times 3$  table with cell probabilities  $\pi_{ij}$ , four nonredundant local log odds can be defined as

$$\omega_{ij} = \log \left( \frac{\pi_{i,j}\pi_{i+1,j+1}}{\pi_{i,j+1}\pi_{i+1,j}} \right)$$

for  $i, j = 1, 2$ . Let  $A$  be the  $9 \times 9$  identity matrix and define matrices  $C$  and  $X$  in the following way:

$$C' = \begin{pmatrix} 1 & -1 & 0 & -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 & -1 & 1 & 0 \\ 0 & 1 & -1 & 0 & -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & -1 & 1 \end{pmatrix} \quad \text{and} \quad X = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} .$$

The hypothesis that the four local log odds are equal can be represented in parametric form as

$$C' \log(A' \pi) = X \beta ,$$

with  $\beta$  the common value of the four local log odds. The columns of the following matrix  $U$  provide a basis for the null space of  $X$ :

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

The same hypothesis can now be represented in parameter-free form as

$$U' C' \log(A' \pi) = 0 .$$

Exactly which constraints are imposed on the cell probabilities can be seen from the product

$$B' = U' C' = \begin{pmatrix} 1 & -1 & 0 & -2 & 2 & 0 & 1 & -1 & 0 \\ 1 & -2 & 1 & -1 & 2 & -1 & 0 & 0 & 0 \\ 1 & -1 & 0 & -1 & 0 & 1 & 0 & 1 & -1 \end{pmatrix} ,$$

and we can represent the hypothesis concisely as

$$B' \log(A' \pi) = 0 .$$

*Example 4: Invariance of Log Odds Ratios Over Time*

Suppose that a particular dichotomous variable has been measured at three different time points, and let  $A$ ,  $B$ , and  $C$  represent the three measurements. The full table  $ABC$  contains the cell frequencies  $n_{ijk}^{ABC}$  corresponding to the theoretical cell probabilities  $\pi_{ijk}^{ABC}$ . Remember that when this three-dimensional array is vectorized, the last subscript changes the fastest and the first subscript the slowest. Suppose now that we want to test whether the (log) odds ratio between consecutive measurements remains constant over time:

$$\frac{\pi_{11+}^{ABC} \pi_{22+}^{ABC}}{\pi_{12+}^{ABC} \pi_{21+}^{ABC}} = \frac{\pi_{+11}^{ABC} \pi_{+22}^{ABC}}{\pi_{+12}^{ABC} \pi_{+21}^{ABC}} .$$

In order to test this hypothesis, loglinear marginal modeling is needed with

$$A' = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{pmatrix}$$

$$C' = \begin{pmatrix} 1 & -1 & -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & -1 & 1 \end{pmatrix}$$

and

$$X = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

so that

$$U = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

and

$$U'C' = \begin{pmatrix} 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \end{pmatrix} .$$

It is easy to see that in this example the matrix product  $A'\pi$  yields the cell probabilities in the marginal tables  $AB$  and  $BC$ , and that  $C'\log(A'\pi)$  defines the appropriate contrasts among the logarithms of these cell probabilities. Finally, the constraint  $U'C'\log(A'\pi) = 0$  corresponds to the hypothesis that the two log odds ratios are equal.

*The General Parameter-free Representation of Loglinear Marginal Models*

The examples above illustrate how both loglinear and loglinear marginal models can be represented by imposing constraints on the cell probabilities. Loglinear marginal models have the parameterized form

$$C' \log A' \pi = X\beta . \quad (2.4)$$

For any matrix  $X$ , a matrix  $U$  can be found whose columns contain a basis of the null space of  $X$ . Every column of  $U$  is orthogonal to all columns of  $X$  (i.e.,  $U'X = 0$ ) and the columns of  $U$  and  $X$  together span the vector space with dimensionality equal to number of cells in the frequency table. For any such matrix  $U$  (which is generally not unique), Eq. 2.4 is equivalent to

$$U'C' \log A' \pi = 0 .$$

Ordinary loglinear models are special cases of loglinear marginal models for which  $A$  and  $C$  are identity matrices of the appropriate order, and matrix  $U$  contains a basis of the null space of the design matrix  $X$ .

Here we briefly sketch (without giving a formal proof) how a basis of the null space of matrix  $X$  can be obtained. Let the  $m \times k$  design matrix  $X$  ( $m > k$ ) be of rank  $r \leq k$ . By means of elementary column operations (Schott, 1997), a  $k \times k$  matrix  $Q$  can be defined such that

$$XQ = (X_1 | 0) ,$$

with  $X_1$  an  $m \times r$  matrix of full column rank  $r$ , and  $0$  an  $m \times (k - r)$  zero matrix. Note that this matrix  $Q$  is not uniquely defined, since it depends on the order in which the elementary column operations are carried out. The columns of matrix  $X_1$  define a basis of the column space of  $X$ . Moreover, it is always possible to select  $r$  linearly independent rows from matrix  $X_1$ , since if this were not the case the rank of  $X_1$  would be smaller than  $r$ . Let the  $r \times r$  matrix  $X_{11}$  contain such a selection of  $r$  linearly independent rows, and let the  $(m - r) \times r$  matrix  $X_{21}$  contain the remaining  $m - r$  rows of  $X_1$ . Furthermore, let  $I_{m-r}$  be the  $(m - r) \times (m - r)$  identity matrix. Then, the columns of matrix

$$U = \begin{pmatrix} -(X_{21}X_{11}^{-1})' \\ I_{m-r} \end{pmatrix} ,$$

after rearranging its rows in the original order, constitute a basis for the null space of  $X$ .

### 2.3.3 Simultaneous Modeling of Joint and Marginal Distributions: Redundancy, Incompatibility and Other Issues

In many applications, it may be necessary or interesting to simultaneously test several loglinear and loglinear marginal models for the same full table. For example, for the

$3 \times 3$  table  $\Pi$  in the first and second example in the previous section, one might be interested in a simultaneous test of marginal homogeneity and independence. Both hypotheses are represented by different matrix constraint equations

$$U'_1 C'_1 \log(A'_1 \pi) = 0$$

and

$$U'_2 C'_2 \log(A'_2 \pi) = 0 ,$$

which can be combined in a single overall equation after defining the appropriate supermatrices:

$$\begin{pmatrix} U'_1 & 0 \\ 0 & U'_2 \end{pmatrix} \begin{pmatrix} B'_1 & 0 \\ 0 & B'_2 \end{pmatrix} \log \left[ \begin{pmatrix} A'_1 \\ A'_2 \end{pmatrix} \pi \right] = \begin{pmatrix} 0 \\ 0 \end{pmatrix} .$$

This allows a simultaneous test of both hypotheses, however some caution is needed in combining the constraints of different models in this straightforward way. Sometimes when imposing simultaneous constraints on several marginal tables, or on marginal and joint distributions, several difficulties may be encountered.

A first one is that particular constraints are redundant, i.e., implied by the other ones in the set of constraints. Sometimes these redundancies are easily detected, e.g., by means of design matrices not being of full rank, but this is certainly not always so. In any case, the algorithm will not work and converge to the ML estimates if redundant constraints are specified.

Another class of problems is the specification of incompatible restrictions, i.e., restrictions which contradict each other and cannot be satisfied simultaneously. During the estimation process, such incompatibilities may be resolved by means of ‘degenerate’ solutions in the sense of not-intended estimated zero effects or uniform distributions, and then may lead to redundancies in the restrictions. For example, imagine a model specification for the cell entries of a successive series of turnover tables that has the (unintended) implication that the marginals of these turnover tables remain stable over time. At the same time, a model is specified for the marginals of these tables that imply a linear net change in location over time. These two models can be resolved by assuming that the linear increase or decrease in location of the marginal is zero. But then, of course, the model for the bivariate joints entries and the model for the marginals contain redundant restrictions. Finally, even if there are seemingly no problems regarding redundancy or incompatibility, difficulties may still arise in terms of applicability of standard asymptotic theory, and even with the substantive interpretation of the resulting model. Fortunately, in many cases frequently occurring in practice, and in most examples discussed in this book, no problems of these kinds occur. Below, some further details of these kinds of problems and their solutions will be presented. Although the solutions and results are only partial, they do cover important situations that occur in practice. Due to the complexity of the problems involved, it may be unrealistic to expect that, for example, a definitely conclusive test for compatibility may be attained. Further extensively



discussed examples of incompatible sets of restrictions will be presented at the end of Chapter 4.

An insightful example of a combination of constraints that gives rise to surprising results is the following (Dawid, 1980; Bergsma & Rudas, 2002a, Example 7). For a  $2 \times 2 \times 2$  table  $ABC$ , it is simultaneously assumed that  $A$  and  $B$  are marginally independent of each other ( $A \perp\!\!\!\perp B$ ), but also conditionally independent given  $C$  ( $A \perp\!\!\!\perp B \mid C$ ). Denote this model as model  $M_0$ . Furthermore, let  $M_1$  be the model in which  $A$  and  $C$  are jointly independent of  $B$  ( $AC \perp\!\!\!\perp B$ ), and make  $M_2$  the model where  $B$  and  $C$  are jointly independent of  $A$  ( $BC \perp\!\!\!\perp A$ ). Then, model  $M_0$  is equivalent to either  $M_1$  or  $M_2$  or both. In other words, if model  $M_0$  applies, exactly one of three following situations may occur:

- $M_1$  applies but  $M_2$  does not;
- $M_2$  applies but  $M_1$  does not, and
- $M_1$  and  $M_2$  both apply.

In the last case, the three variables  $A$ ,  $B$ , and  $C$  are mutually independent. On the other hand, if either  $M_1$  or  $M_2$  or both apply,  $M_0$  also applies. Because it is not clear what the exact implications are from the original two restrictions in terms of the choice between  $M_1$  and  $M_2$ , the interpretation of this combined model is not straightforward. Moreover, the dimension of the model (i.e., the number of free parameters) is not constant: if all three variables are independent, there are three free parameters; in other cases there are four. This leads to nonstandard asymptotics if the true number of free parameters is three. Fortunately, in practice it is usually not difficult to verify the absence of such problems, as discussed below.

First, consider the case of combining (compatible) restrictions on certain marginals with a loglinear model for the joint table. The above example is such a case, as restrictions were placed on the marginal table  $AB$  and a loglinear model was assumed for joint table  $ABC$ . A simple test for the absence of problems is that the restricted marginals should be a subset of the set of sufficient configurations (Bishop et al., 1975, p. 66) of the loglinear model. For the above example, in model  $M_0$  ( $A \perp\!\!\!\perp B \mid C$ ) the loglinear model for the full table is  $\{AC, BC\}$ , which means that the marginal tables  $AC$  and  $BC$  can freely be restricted (provided of course that the restrictions on the marginals themselves are compatible). Instead, above  $AB$  was restricted ( $A \perp\!\!\!\perp B$ ), which led to some unexpected results. More generally, precise conditions using matrix formulations can be given. Suppose we restrict the linear combinations of probabilities  $A'\pi$ , where  $A$  is a matrix with nonnegative elements, and the loglinear model has design matrix  $X$ , i.e., we assume  $\log \pi = X\beta$  for a vector of loglinear parameters  $\beta$ . Then, a sufficient condition for the absence of problems is that the columns of  $A$  are a linear combination of the columns of  $X$  (Lang & Agresti, 1994; Bergsma & Rudas, 2002a).

Before discussing the more general case of loglinear restrictions on nested sets of marginals (see below), the case of compatibility of fixed marginals must be discussed, as this gives some insight into the former problem. There are some obvious cases where fixed marginals are incompatible: if the  $AC$  distribution is assumed to

be uniform, i.e.,  $\begin{pmatrix} .25 & .25 \\ .25 & .25 \end{pmatrix}$ , and  $BC$  is assumed to be  $\begin{pmatrix} .4 & .2 \\ .2 & .2 \end{pmatrix}$ , then the  $C$  marginals in the two tables are obviously not compatible: in  $AC$  it is  $(.5, .5)$  and in  $BC$  it is  $(.6, .4)$ . In other words, there is no joint distribution with these two bivariate marginals. The set of marginals  $\{AC, BC\}$  is an example of a *decomposable* set of marginals for which compatibility can be checked easily in this way. Generally, a set of marginals is called decomposable if there is an ordering of the marginals so that, for any  $k$ , the intersection of the  $k$ th marginal with the union of the first  $k - 1$  marginals equals the intersection of the  $k$ th and  $\ell$ th marginals for some  $\ell < k$ . It is less easy to check for the compatibility of *nondecomposable* sets of marginals. An example is the nondecomposable set consisting of  $AB$ ,  $BC$ , and  $AC$ . If each of the tables is restricted to be  $\begin{pmatrix} 0 & .5 \\ .5 & 0 \end{pmatrix}$ , then even though the one-dimensional marginals are compatible, there is no joint distribution with these bivariate marginals since the restrictions imply a perfect negative correlation for all pairs of variables, which is impossible.

The problem of compatibility of several marginal tables is very closely related to the existence of maximum likelihood estimates for loglinear models with zero observed cells, which is the reason decomposability comes in here. In particular, ML estimates of loglinear parameters for a certain loglinear model exist if, and only if, there exists a strictly positive distribution compatible with the sufficient statistics for the model. See Haberman (1973, 1974) for a rigorous treatment of the problem of the existence of ML estimates for loglinear models.

Next, more general (loglinear) constraints on possibly nested marginals will be dealt with. A first marginal (like  $AB$ ) is nested in a second marginal (like  $ABC$ ) if it consists of a selection of variables from the second marginal. In this way, every marginal is nested in the complete set of variables from the joint distribution. A general result was obtained by Bergsma and Rudas (2002a). The main sufficient condition for compatibility they formulated is that linear restrictions on loglinear marginal parameters are compatible if no two restricted parameters with different superscripts have indices belonging to the same variables in the subscript. In the above example of marginal and conditional independence, the constraints were  $\lambda_{ij}^{AB} = 0$ ,  $\lambda_{ij*}^{ABC} = 0$ , and  $\lambda_{ijk}^{ABC} = 0$ . The problem arises from the first two restrictions: two corresponding loglinear parameters with subscript set  $\{i, j\}$  belonging to variables  $A$  and  $B$  are restricted in the two different marginal tables  $AB$  and  $ABC$ , leading to the problems. If the compatibility condition is satisfied, then additionally the model interpretation is straightforward and standard asymptotics apply. In Section 4.5, analyses of real data are discussed where these results are relevant.

For affine restrictions on the loglinear marginal parameters, i.e., restrictions in which linear combinations of the loglinear parameters are set equal to a nonzero value, the situation is more complex. Bergsma and Rudas (2002a) showed that if the set of marginals that is restricted is *ordered decomposable*, then the above condition is sufficient to guarantee the compatibility of constraints. A set of marginals is ordered decomposable if there is an ordering such that, for any  $k$ , the maximal elements of the first  $k$  marginals in the ordering are decomposable. See Bergsma and

Rudas (2002b) and Bergsma and Rudas (2002c) for a more extended discussion and several illustrations of compatibility issues. Extensions of these results to marginal models based on cumulative and other types of logits and higher order parameters are given by Bartolucci, Colombi, and Forcina (2007). An algorithm for checking compatibility of marginal distributions with specific values is given by Qaqish and Ivanova (2006).

Finally, there is the question of the uniqueness of ML estimates. Bergsma (1997) showed that for many marginal homogeneity models, the likelihood has a unique local maximum (which then must be the global maximum). The simplest example is when the marginals are disjoint (like marginals  $AB$  and  $CD$  in contrast to marginals  $AC$  and  $BC$ ): then for any practically relevant marginal homogeneity model, the likelihood has a unique stationary point that is the ML estimate.

### 2.3.4 \*\*\*Maximum Likelihood Estimates of Constrained Cell Probabilities

Suppose that the observed frequencies  $n_i$ ,  $i = 1, \dots, I$  are multinomially or Poisson distributed with theoretical cell probabilities  $\pi_i$ . As shown above, when these cells satisfy a loglinear marginal model, there exist matrices  $A$  and  $B$  such that

$$h(A'\pi) = B' \log(A'\pi) = 0.$$

The notation  $h(A'\pi) = 0$  allows extension to nonloglinear marginal models as well, which are discussed in the next chapter. In general, all loglinear marginal models can be specified in such a way that all rows of matrix  $B'$  sum to zero, which means that each row represents a contrast among the logarithms of the cell probabilities. A sufficient but not necessary condition for the rows of  $B' = U'C'$  to sum to zero is that the matrix  $X$  in Eq. 2.4 contains a constant column. If the rows sum to zero, it is immaterial whether we formulate the model in terms of expected cell frequencies or in terms of cell probabilities. More technically, the function  $h$  is such that for any  $c > 0$ ,  $h(cx) = h(x)$ . This is an important condition simplifying maximum likelihood estimation. We say that the function  $h$  is *homogeneous*, and this issue will be discussed in more extensively in Chapter 3. For any homogeneous scalar function  $f$ , Euler's theorem says that

$$\sum_i x_i \frac{\partial f(x)}{\partial x_i} = 0. \quad (2.5)$$

In order to test whether a particular marginal model fits the data well, first the maximum likelihood estimates of the cell probabilities must be obtained under the constraints imposed by the model. Utilizing the Lagrange multiplier method for constrained optimization as described by Aitchison and Silvey (1958) and Aitchison and Silvey (1960), these estimates are obtained by determining the saddle point of the kernel of the Lagrangian log likelihood function

$$L(\pi, \lambda, \mu) = p' \log(\pi) - \mu \left( \sum \pi_i - 1 \right) - \lambda' h(A'\pi), \quad (2.6)$$

in which  $p$  is the vector of observed proportions,  $\lambda$  is a vector of unknown Lagrange multipliers,  $\mu$  a Lagrange multiplier and

$$h(A'\pi) = B' \log(A'\pi) .$$

The term  $\mu (\sum \pi_i - 1)$  is added to incorporate the constraint that the cell probabilities sum to one. The maximum likelihood estimates are denoted as  $\hat{\pi}$ ,  $\hat{\lambda}$ , and  $\hat{\mu}$ .

We will now give a set of equations that has the ML estimate  $\hat{\pi}$  as its solution. In these equations, following the method of Bergsma (1997, Appendix A), the Lagrange multipliers will be expressed as a function of  $\pi$ , i.e., we effectively eliminate them from the equations, thus simplifying the problem of finding ML estimates  $\hat{\pi}$ . We need the Jacobian of the constraint function  $h$ , given as

$$H(x) = \frac{dh(x)'}{dx} = D_x^{-1} B ,$$

where  $D_x$  is the diagonal matrix with vector  $x$  on the main diagonal. Note that the  $(i, k)$ th element of  $H(x)$  is given by

$$\frac{\partial h_k(x)}{\partial x_i} .$$

Now using the shorthand

$$H = H(A'\pi) ,$$

the chain rule for matrix differentiation leads to

$$\frac{dh(A'\pi)'}{d\pi} = AH .$$

The derivative of the Lagrangian function (Eq. 2.6) with respect to  $\pi$  then is

$$l(\pi, \lambda, \mu) = \frac{dL(\pi, \lambda, \mu)}{d\pi} = \frac{p}{\pi} - \mu - AH\lambda .$$

Thus, the ML estimates  $(\hat{\pi}, \hat{\lambda}, \hat{\mu})$  are solutions to the simultaneous equations

$$l(\pi, \lambda, \mu) = 0$$

$$h(A'\pi) = 0 .$$

By homogeneity of  $h$  and Euler's theorem (see Eq. 2.5),  $\pi'AH = 0'$ . Hence,  $\pi'l(\pi, \lambda, \mu) = 1'p - \mu 1'\pi = 1 - \mu = 0$ , and so  $\hat{\mu} = 1$ . Let

$$l(\pi, \lambda) = l(\pi, \lambda, \hat{\mu}) = \frac{p}{\pi} - 1 - AH\lambda ,$$

and we now only need to solve the simplified equations

$$l(\pi, \lambda) = 0 \tag{2.7}$$

$$h(A'\pi) = 0 . \tag{2.8}$$

Bergsma (1997) proposed to write the Lagrange multiplier vector  $\lambda$  in terms of the cell probabilities  $\pi$  as follows:

$$\lambda(\pi) = [H'A'D_{\pi}AH]^{-1} [H'A'(p - \pi) + h(A'\pi)] .$$

It can then be verified that if  $\hat{\pi}$  is a solution of the equation

$$l(\pi, \lambda(\pi)) = 0 , \quad (2.9)$$

then  $(\hat{\pi}, \hat{\lambda})$ , with  $\hat{\lambda} = \lambda(\hat{\pi})$ , is a solution of (2.7) and (2.8), and if  $(\hat{\pi}, \hat{\lambda})$  is a solution of Eqs. 2.7 and 2.8, then  $\hat{\lambda} = \lambda(\hat{\pi})$ . Hence, by writing  $\lambda$  in terms of  $\pi$  and substituting into  $l(\pi, \lambda)$ , we have reduced the dimension of the problem by effectively eliminating the Lagrange multiplier  $\lambda$  as an independent parameter. That is, we only need to solve Eq. 2.9 in terms of  $\pi$ .

In general, this optimization problem cannot be solved analytically but requires appropriate numerical optimization procedures. In the next section, we describe such an algorithm, based on the likelihood equation Eq. 2.9.

If a stratified sampling procedure has been used, the definition of the Lagrangian can easily be extended to take the existence of different strata in the population into account. Here it is required that  $h$  is homogeneous relative to the stratification used (see Lang, 1996b for further details). He also showed that for inference about certain higher order loglinear marginal parameters, usually those of most interest, ignoring the fact that sampling is stratified leads to identical asymptotic inferences.

### 2.3.5 \*\*\*A Numerical Algorithm for ML Estimation

Bergsma (1997), building on previous work by Aitchison and Silvey (1958) and Lang and Agresti (1994), derived the following algorithm for fitting marginal models. The first step of the algorithm is to choose an appropriate starting point  $\pi^{(0)}$ , after which subsequent estimates  $\pi^{(k+1)}$  ( $k = 0, 1, 2, \dots$ ) are calculated iteratively using the formula

$$\log \pi^{(k+1)} = \log \pi^{(k)} + \text{step}^{(k)} l[\pi^{(k)}, \lambda(\pi^{(k)})]$$

for an appropriate step size  $\text{step}^{(k)}$ . The algorithm terminates at an iteration  $k$  if  $l[\pi^{(k)}, \lambda(\pi^{(k)})]$  is sufficiently close to zero. Although the algorithm looks like a linear search, it can be viewed as a form of Fisher scoring since it is based on a weighting by the inverse of the expected value of the second derivative matrix of the Lagrangian likelihood  $L(\pi, \lambda)$  (Bergsma, 1997).

The algorithm of Bergsma used here and the one described by Aitchison and Silvey and Lang and Agresti are both based on the Lagrange multiplier technique, but a salient difference is that the latter searches in the product space of the probability simplex and the space of the Lagrange multipliers, whereas our algorithm searches in the lower dimensional probability simplex. In this sense, Bergsma's algorithm is simpler, and practical experience also indicates that it also performs better numerically. For example, Lang, McDonald, and Smith (1999) needed to impose additional restrictions on certain loglinear marginal models in order to achieve convergence with

the Lang-Agresti algorithm, whereas convergence was easily achieved for the same models without the additional restrictions with Bergsma's simplified algorithm.

In our experience, a good choice of starting point  $\pi^{(0)}$  is simply the observed cell proportions if all of them are strictly positive. If there are zero observed cells, we found that a slight smoothing towards uniformity works well, a good choice, in particular, seems to be to add .01 divided by the number of cells to each cell, and rescale so that the proportions add up to one. Starting values should always be strictly positive. For certain starting points (wildly different from the observed proportions) in certain problems, we could not reach convergence. Although there is no guarantee that any starting values lead to convergence of the algorithm, all of the manifest variable models in this book could be fitted with the default starting values of our programme. For certain latent variable models, this did not always work; further details are given in Chapter 6.

Since  $l(\pi, \lambda(\pi))$  is not the derivative of an unrestricted likelihood function to be maximized, the choice of step size becomes more difficult because we do not know if a new estimate is 'better' than the previous one. However, the step size may be chosen such that an appropriate function that measures the 'distance' of the iterative estimate from the ML estimate decreases. A reasonable function is the quadratic form

$$d(\pi) = l[\pi, \lambda(\pi)]' D_{\pi} l[\pi, \lambda(\pi)]$$

which is zero if and only if  $\pi$  is a stationary point of  $L(\pi, \lambda(\pi))$ , and positive otherwise. In the search for an optimal value of the step size,  $step^{(k)}$  is initially set equal to 1. If this results in an increase of the criterion  $d(\pi)$ , the step size is halved. This process of halving the step size is continued until  $d(\pi)$  decreases. Unfortunately, it is not always possible to obtain a decrease of  $d(\pi)$ , because  $l(\pi, \lambda(\pi))$  is not a gradient of  $d(\pi)$ . If that is the case, a 'jump' may have to be made to a different region, for example, by going back to  $step^{(k)} = 1$ . We could not always get convergence with this step-size halving method: for some problems, we needed to set a maximum to the step size (e.g., 0.3). The maximum permissible step size had to be found by trying out different values, but we always managed to find it fairly quickly. Note that generally speaking the smaller the maximum step size, the higher the likelihood of convergence, but the slower the algorithm potentially becomes. Concluding, the overall procedure for choosing a step size is somewhat ad hoc, but has worked well in practice for us.

A potentially serious problem with the algorithm is the possible singularity or ill-conditioning of the matrix  $H'A'D_{\pi}AH$ , which has to be inverted. If the matrix is singular at every value of  $\pi$ , this means that at least one constraint is redundant and needs to be removed, see also the discussion in Section 2.3.3. Another possibility is that the matrix is singular at the ML estimate  $\hat{\pi}$ , but not at values close to it. Our Mathematica programme then gives warnings about the ill-conditioning during the iterative process. We found that in such cases it still appears to be possible to obtain fairly good convergence of the algorithm, say, to four or five decimal places of the likelihood ratio statistic, but the algorithm will not converge any further. Fortunately,

this seems to be rare for manifest data loglinear marginal models discussed in this chapter, although on the latent level (Chapter 6) we did encounter the problem.

Aside from convergence, another important issue is efficiency. Although the algorithm does involve a matrix inversion, for all of the problems in this book the matrix to be inverted is relatively small and does not form a computational bottleneck. The real computational challenge arises from computations involving the matrix  $A$  in the marginal model specification. For large tables, this matrix is large and may not be storable in computer memory. However, for marginal models the matrix consists of zeroes and ones and can be stored much more efficiently using sparse array techniques, which are implemented in computer packages such as Mathematica or MATLAB. The advantage is that the zeroes do not need to be stored, but the disadvantage is that we still need to store (the locations of) the ones. To overcome even the latter disadvantage, a programme can be written that avoids the storing of matrix  $A$  altogether, and uses its special structure to do computations directly. This approach can also lead to potentially significant speed improvements, but this all depends on the details of the programme implementation.

For either of the approaches outlined above to work, the algorithm needs to be written out in more detail, specifying the order of computations to be done. The computational bottleneck of the algorithm is the computation of  $l[\pi, \lambda(\pi)]$ , which can be written out fully as

$$l[\pi, \lambda(\pi)] = \frac{p}{\pi} - 1 - A \left( H \left[ H' (A' D_{\pi} A) H \right]^{-1} \left[ H' (A' (p - \pi)) + h(A' \pi) \right] \right).$$

Here, extra parentheses have been inserted to indicate the order of evaluation. There are three potentially inefficient computations involving matrix  $A$ : 1) multiplication of  $A$  by a vector, 2) computation of  $A' D_{\pi} A$ , and 3) multiplication of  $A'$  by a vector. These operations can be made (much) more efficient both by the use of sparse array techniques or doing the computations using the special structure of matrix  $A$  without creating the matrix itself. Note that the matrix  $A' D_{\pi} A$  is typically small compared to the size of the table. For example, say we have 10 trichotomous variables, and  $A' \pi$  consists of the univariate marginals, then the full table consists of  $3^{10} = 59049$  cells and  $A' D_{\pi} A$  is a  $30 \times 30$  matrix with only 900 elements.

For comparison purposes, a different order of evaluation is given as follows:

$$l[\pi, \lambda(\pi)] = \frac{p}{\pi} - 1 - (AH) \left[ (H' A') D_{\pi} (AH) \right]^{-1} \left[ (H' A') (p - \pi) + h(A' \pi) \right].$$

In this case, matrix  $AH$  and its transpose have to be computed and stored. In the example with 10 trichotomous variables, its size is  $59049 \times 30$ . Especially for even larger tables, storing and manipulating this matrix could lead to problems in terms of computing time and space.

A final issue with the algorithm is that for loglinear marginal models, estimated marginal probabilities should not be zero, because we need to take their logarithm. However, if there are zero observed marginal probabilities, estimated probabilities may be zero as well. Since we cannot take the logarithm of zero, we advise incorporating a minimum value for the (joint) estimated probabilities in the estimation



procedure. Thus, if at any point during the iterative process an estimated probability obtains a value of, say, less than  $10^{-100}$ , we can replace its value by  $10^{-100}$ .

Concluding this subsection, the algorithm proposed here is not without its problems, however, we have successfully applied it to fit models for tables with more than a million cells.

### 2.3.6 \*\*\*Efficient Computation of ML Estimates for Simultaneous Joint and Marginal Models

If we wish to simultaneously test a loglinear model and one or more loglinear marginal models, the procedure described in the previous subsection can be applied but may be (far) too inefficient, especially for large tables, and we describe a more efficient modified procedure here. This case corresponds to the simultaneous models discussed in Section 2.3.3 with either  $A_1$  or  $A_2$  equal to the identity matrix. The procedure is briefly outlined in Bergsma (1997, page 95) and in detail in Lang et al. (1999). It is especially important for use with the EM algorithm for loglinear latent variable models with marginal constraints (see Chapter 6). The modified algorithm gives the same iterative estimates as the algorithm described in the previous section, but computes these more efficiently.

We assume a loglinear model for  $\pi$  specified as

$$\log \pi = W\gamma$$

and a loglinear marginal model of the form

$$B' \log A' \pi = 0, \quad (2.10)$$

where matrices  $A$  and  $B$  satisfy the regularity condition described in Section 2.3.3 that the columns of  $A$  are a linear combination of the columns of  $W$ . Using this regularity condition, we can show that  $l(\pi, \lambda(\pi))$  based on the simultaneous marginal and loglinear model reduces to

$$\begin{aligned} l_W(\pi, \lambda(\pi)) &= l(\pi, \lambda(\pi)) + W(W'D_\pi W)^{-1}W'(p - \pi) \\ &= \frac{p}{\pi} - 1 - AH\lambda(\pi) + W(W'D_\pi W)^{-1}W'(p - \pi) \end{aligned}$$

where  $l(\pi, \lambda(\pi))$  in the formula is based on the marginal model in Eq. 2.10. The advantage of this formulation is that it is not necessary to compute the orthocomplement of  $W$ , which tends to be large. To illustrate, if we have a loglinear model with only first order interactions for 10 trichotomous variables, then  $W$  has size  $3^{10} \times 201 = 59049 \times 201$ , whereas its orthocomplement has size  $3^{10} \times (3^{10} - 201) = 59049 \times 58848$ , which is almost 300 times larger.

With starting values  $\pi^{(0)}$  satisfying the loglinear model, the algorithm is analogous to what we did previously, namely, for  $k = 0, 1, 2, \dots$ ,

$$\log \pi^{(k+1)} = \log \pi^{(k)} - \text{step}^{(k)} l_W[\pi^{(k)}, \lambda(\pi^{(k)})].$$



For  $\pi^{(0)}$ , the uniform distribution can be chosen. Otherwise, the same recommendations for implementation of the algorithm apply as in the previous section.

In this algorithm, in contrast to the one of the previous subsection, not one but two matrices need to be inverted at every iterative step, namely

$$Q(\pi) = W'D_{\pi}W$$

and

$$R(\pi) = H'A'D_{\pi}AH .$$

From the assumption that the columns of  $A$  are linear combinations of the columns of  $W$ , there exists a matrix  $U$  such that  $A = WU$ , and we can write

$$R(\pi) = H'U'W'D_{\pi}WUH = H'U'Q(\pi)UH .$$

Normally  $U$  has full column rank, so if  $H$  also has full column rank and  $Q$  is nonsingular,  $R$  is nonsingular. In practice, either of the matrices  $Q$  and  $R$  may be singular when evaluated at  $\hat{\pi}$ . For  $Q(\pi)$ , we found that a generalized (Moore-Penrose) inverse can be used instead of the true inverse if it doesn't exist. For  $R(\pi)$ , a reduction is needed in the number of constraints that are imposed, thereby reducing the number of columns of  $H$  to make it full column rank. However, in the problems of this book, we found that although the second matrix could be near singular, giving warnings by our Mathematica programme, it was still sufficiently far from singularity to allow the algorithm to converge fairly well. The main potential computational bottleneck of the present algorithm is the actual computation of  $Q(\pi)$  and of  $R(\pi)$  rather than their inversion. The size of these complexities are increasing with 1) the number of constraints imposed by  $h(A'\pi) = 0$ , which determines the number of columns of  $H$ , and 2) the number of loglinear parameters  $\gamma$  in the model, which determines the number of columns of  $W$ .

### 2.3.7 \*\*\*Large Sample Distribution of ML estimates

In the general model formulation of Eq. 2.4, several parameters may be of interest, in particular, 1) the vector of marginal probabilities  $A'\pi$ , 2) the loglinear marginal parameters  $\phi(\pi) = C' \log A'\pi$ , and 3) the vector of model parameters  $\beta$ . Under conditions usually met in practice, the ML estimates of these parameters are consistent estimators of the population values and have an asymptotic normal distribution ( see Section 2.3.3 for exceptions). In particular, the elements of  $\phi(\pi)$  must have continuous second derivatives (Lang, 1996a). Below, using results by Aitchison and Silvey (1958) (see also Lang, 1996a and Bergsma, 1997), we provide the asymptotic covariance matrices of these parameters, assuming the appropriate regularity conditions are met. We first give the asymptotic covariance matrix of the estimated cell probabilities if model defined by Eq. 2.4 is true. Although this matrix can be large and is often of little interest in itself, we can use it to calculate the asymptotic covariance matrix of the aforementioned parameters of interest using the delta method. For Poisson and

multinomial sampling, the asymptotic covariance matrix of the estimated probabilities is

$$V(\hat{\pi}) = \frac{1}{N} (D_{\pi} - D_{\pi} A H (H' A' D_{\pi} A H)^{-1} H' A' D_{\pi} - \pi \pi') . \quad (2.11)$$

By the delta method, the asymptotic covariance for the marginal probabilities is

$$V(A' \hat{\pi}) = A' V(\hat{\pi}) A ,$$

for the parameter vector  $\hat{\phi} = \phi(\hat{\pi})$  it is

$$V(\hat{\phi}) = C' D_{A' \pi}^{-1} A' V(\hat{\pi}) A D_{A' \pi}^{-1} C ,$$

and since  $\beta = (X' X)^{-1} X' \phi(\pi)$ , for  $\hat{\beta}$  it is

$$V(\hat{\beta}) = (X' X)^{-1} X' V(\hat{\phi}) X (X' X)^{-1} .$$

For the residuals  $p_i - \hat{\pi}_i$ , to be discussed in the next subsection, the asymptotic covariance matrix is

$$V(p_i - \hat{\pi}) = \frac{1}{N} D_{\pi} A H (H' A' D_{\pi} A H)^{-1} H' A' D_{\pi} . \quad (2.12)$$

For the perhaps more interesting residuals  $\phi_{\text{obs}} - \hat{\phi}$ , where  $\phi_{\text{obs}} = \phi(A' p)$ , the delta method yields

$$V(\phi_{\text{obs}} - \hat{\phi}) = C' D_{A' \pi}^{-1} A' V(p_i - \hat{\pi}) A D_{A' \pi}^{-1} C . \quad (2.13)$$

For stratified sampling, let  $\pi^{(k)}$  be the probability vector for stratum  $k$ . Then, the asymptotic covariance matrix of the estimated probabilities is

$$V(\hat{\pi}) = \frac{1}{N} \left( D_{\pi} - D_{\pi} A H (H' A' D_{\pi} A H)^{-1} H' A' D_{\pi} - \oplus_k \pi^{(k)} (\pi^{(k)})' \right) ,$$

where  $\oplus$  is the direct sum, defined as the block-diagonal matrix with the summed matrices as blocks. In the same way as above, the corresponding covariance matrices for other parameters are obtained using the delta method. However, we can show that the covariance matrices of loglinear parameters (i.e., the elements of  $\hat{\phi}$  and  $\hat{\beta}$ ) need not be affected by the stratification. In particular, in a two-way table formed by a stratifying variable and a response variable, the (log) odds ratios and the main loglinear effect pertaining to the response variable are unaffected, while the main loglinear effect pertaining to the stratifying variable is affected. The covariance matrix of  $\hat{\pi}$  under more general types of (stratified) sampling schemes is given in Lang (1996a).

If we simultaneously impose a loglinear model  $\log \pi = W \gamma$  and a loglinear marginal model  $h(A' \pi) = 0$ , we obtain the partitioned covariance matrix

$$V(\hat{\pi}) = V(\hat{\pi}_1) + V(\hat{\pi}_2) - V(\hat{\pi}_0) \quad (2.14)$$

where  $\hat{\pi}_1$  is the ML estimate of  $\pi$  under only the marginal model (see Eq. 2.11 for  $V(\hat{\pi}_1)$ ),  $\hat{\pi}_2$  is the ML estimate of  $\pi$  under only the loglinear model, for which

$$V(\hat{\pi}_2) = \frac{1}{N} (D_\pi W (W' D_\pi W)^{-1} W' D_\pi - \pi \pi') ,$$

and  $\hat{\pi}_0 = p$  is the unrestricted ML estimate of  $\pi$  for which  $V(p) = D_\pi - \pi \pi'$ . The partitioning holds because of the orthogonality, in the sense of asymptotic independence of ML estimates of the loglinear and marginal parameters (Bergsma, 1997, Section 5.4.2, Appendix A.3 and references therein). Again, the delta method is used to obtain expressions for the relevant asymptotic covariance matrices of other parameters.

Classical confidence intervals are easily calculated using standard errors, which we frequently provide in this book. Currently, other confidence intervals, such as those obtained by inverting the score statistic, are being developed and are now starting to become feasible for marginal models as well. Typically, such intervals are more cumbersome to compute however. The interested reader can consult (Agresti, 2002, Sections 1.4.2 and 3.1.8), Agresti and Coull (1998), Brown, Cai, and Dasgupta (1999) and Lang (2008).

### 2.3.8 Model Evaluation

In Section 2.1.1, the goodness-of-fit statistics  $G^2$  and  $X^2$  have been described that can be used to evaluate whether a given model fits the data. If the model does not fit well, insight can be gained into the reasons for this lack of fit by analyzing cell (or other) residuals, which are measures for the deviation of observed from fitted cell values. Even if the model fits well, these residuals can be used to detect certain deviations from the model that are not apparent from the overall goodness of fit.

Various types of residuals are in use. For cell  $i$ , the raw residual  $p_i - \hat{\pi}_i$ , where  $p_i$  is the observed proportion in cell  $i$ , depends strongly on the size of  $\hat{\pi}_i$  and is therefore of limited use. A measure that adjusts for the size of  $\hat{\pi}_i$  is the *standardized residual*, which is defined as

$$e_i = \sqrt{N} \frac{p_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i}}.$$

The  $e_i$  are related to the Pearson statistic by  $\sum e_i^2 = X^2$ . Thus, they show, for every cell, exactly how much it contributes to a large value of  $X^2$ . Pearson residuals may be useful for marginal probabilities as well, although in that case their squares do not add up to the  $X^2$  statistic, so a bit more care has to be taken in their interpretation.

One drawback of standardized residuals is that their variance is smaller than 1, so a comparison with the standard normal distribution is not appropriate. The *adjusted residual* proposed by Haberman (1974) is defined as the raw residual  $p_i - \hat{\pi}_i$  divided by its standard error. Because its mean is 0 and variance is 1, it is better suited for comparison with the standard normal than the standardized residual. Denoting the adjusted residuals by  $r_i$ , the definition is

$$r_i = \frac{p_i - \hat{\pi}_i}{\text{se}(p_i - \hat{\pi}_i)}.$$

The values of the standard errors  $\text{se}(p_i - \hat{\pi}_i)$  are given by the square roots of the diagonal entries of the right hand side of Eq. 2.12.

Other adjusted residuals may be considered as well. Perhaps most interesting are the adjusted residuals of marginal loglinear parameters, defined as

$$\frac{\phi_{\text{obs},j} - \hat{\phi}_j}{\text{se}(\phi_{\text{obs},j} - \hat{\phi}_j)}.$$

The denominator is obtained as the square root of the corresponding diagonal element of the right-hand side of Eq. 2.13.

Marginal Models

For Dependent, Clustered, and Longitudinal Categorical  
Data

Bergsma, W.; Croon, M.A.; Hagenaars, J.A.

2009, XI, 268 p., Hardcover

ISBN: 978-0-387-09609-4