The numerical/computational methods underlying these models are illustrated using the GAUSS programming language. The custom programming capabilities of Stata, R, TDA, and LIMDEP are also illustrated.[12] SAS, Stata, and R have procedures to estimate generalized linear models, which would include all the models discussed in this chapter and most of the models discussed in Chapter 4.

3.6. Summary

This chapter provides an overview of the most common techniques for modeling binary data. We first considered the transformational approach, and showed that when data are in the form of contingency tables (or fixed grids), FGLS estimation techniques can be used on the transformed empirical probabilities. With grouped or ungrouped data, the IRLS technique can be used to obtain ML estimates from logit, probit, and complementary log-log models as generalized linear models. The IRLS technique thus provides a natural extension of the FGLS technique to individual-level data. Logit and probit models can be justified by invoking the concept of utility maximization. The latent variable formulation is a natural extension of this viewpoint.

12. We are continuously updating our website to include examples of special purpose software such as ℓEM, TDA, and aML.

Chapter 4

Loglinear Models for Contingency Tables

4.1. Contingency Tables

Contingency tables are joint frequency distributions of two or more categorical variables. More formally, a contingency table can be thought of as a cross-classification of possible values (or categories) of two or more variables, together with the number of observations in each cross-classified cell reported. When two variables are involved, the resulting contingency table is called a *two-way table*. When three variables are involved, it is called a *three-way table*. A three-way or higher-way table is also referred to simply as a *multiway table*.

Contingency tables, or "crosstabs," are among the oldest and the most widely used statistical tools available to social scientists. One major reason for their popularity is simplicity. Another reason is that contingency tables are nonparametric or require very weak parametric (or distributional) assumptions. More often than not, the researcher directly interprets the descriptive statistics revealed by a contingency table and reaches substantive conclusions without resort to explicit modeling. This eye-balling method, however, is very imprecise when the researcher explores complicated relationships or analyzes multiway tables. In this chapter, we will learn how to model contingency tables according to theoretically derived hypotheses and, in so doing, smooth out apparent irregularities due to sampling variability.

4.1.1. Types of Contingency Tables

Goodman (1981a) lists three ideal types of two-way contingency tables. They are:

1. The joint distribution of two explanatory variables (e.g., height and weight).
2. The causal relationship of an outcome variable depending on an explanatory variable (e.g., smoking and lung cancer).
3. The association between two outcome variables (e.g., attitude toward abortion and attitude toward premarital sex).

Note that the distinction among the three types of contingency tables is conceptual, for they appear in the same form. In fact, statistical models for contingency tables discussed in this chapter are estimated with frequencies rather than with the outcome variable as the expressed dependent variable in a generalized

Table 4.1: Education and attitude toward premarital sex.

Education	**Attitude toward premarital sex**		
	Disapproval	**Approval**	**Total**
High school or less	873	1190	2063
College and above	533	1208	1741
Total	1406	2398	3804

linear model. As in the case of correlations, cross-tabulations are inherently symmetric. For a simple regression involving only one independent variable, the slope coefficient can be recovered from the symmetric correlation coefficient between the dependent variable and the independent variable plus the scale parameters for the two variables. Likewise, statistical models for the analysis of contingency tables are also symmetric from the standpoint of estimation, although conceptual distinctions can be drawn between an outcome variable and an explanatory variable.

4.1.2. *An Example and Notation*

To go beyond the simplistic method of eye-balling, one needs to model contingency tables. Loglinear models are designed for this purpose. For the three types of contingency tables discussed earlier, the same statistical models are applicable. They are called loglinear models and are formally defined later in this chapter. For now, let us illustrate the setup and the notation with a concrete example.

Table 4.1 shows a cross-tabulation between level of education and attitude toward premarital sex.[1] The data are drawn from the 1987–1991 pooled General Social Surveys (GSS). For our illustration, let us assume the table to be of type 2, with education being the explanatory variable and attitude toward premarital sex the outcome variable.

Note that we use the symbol $+$ to denote summation. Subscript $i+$ stands for the row marginal total:

$$f_{i+} = \sum_{j=1}^{J} f_{ij} \text{ and } F_{i+} = \sum_{j=1}^{J} F_{ij}$$

1. The original GSS question was: "If a man and woman have sexual relations before marriage, do you think it is always wrong, almost always wrong, wrong only sometimes, or not wrong at all?" We collapsed the first two responses into "disapproval" and the last two responses into "approval."

Table 4.2: Observed (expected) frequencies.

Education	**Attitude toward premarital sex**		
	Disapproval	**Approval**	**Total**
High school or less	f_{11} (F_{11})	f_{12} (F_{12})	f_{1+} (F_{1+})
College and above	f_{21} (F_{21})	f_{22} (F_{22})	f_{2+} (F_{2+})
Total	f_{+1} (F_{+1})	f_{+2} (F_{+2})	f_{++} (F_{++})

Similarly, subscript $+j$ stands for the column marginal total:

$$f_{+j} = \sum_{i=1}^{I} f_{ij} \text{ and } F_{+j} = \sum_{i=1}^{I} F_{ij}$$

and subscript $++$ represents the grand total:

$$f_{++} = \sum_{j=1}^{J} \sum_{i=1}^{I} f_{ij} \text{ and } F_{++} = \sum_{j=1}^{J} \sum_{i=1}^{I} F_{ij}$$

Obviously, $f_{++} = n$, the sample size. In practice, almost all models have the property $F_{++} = f_{++}$. As will be shown later in this chapter, most applications in social science actually maintain equality in marginal totals between the expected and observed frequencies.

In general, we denote a two-way contingency table as consisting of two variables: a row variable (R) and a column variable (C). Let R vary by a row index i, where $i = 1, \ldots, I$, and let C vary by a column index j, where $j = 1, \ldots, J$. When one of the two variables is an outcome variable and the other is an explanatory variable (i.e., type 2), it is customary to let R denote the explanatory variable and C the outcome variable. We denote the cell frequency of the ith row and the jth column by f_{ij} and the expected frequency under a model by F_{ij}. The distinction between the observed frequency (f) and the expected frequency (F) disappears in the special case of a saturated model, in which $f = F$ for all the cells in the table. In this example, the observed frequencies and expected frequencies (in parentheses) are denoted by the notation in Table 4.2.

4.1.3. *Independence and the Pearson χ^2 Statistic*

As with any observed data from a sample, we should treat frequencies in a contingency table as realizations of an underlying process. Because of sampling variation, observed frequencies may appear much less regular than an underlying pattern. One possible, and often interesting, pattern is the independence of the row and column variables. That is, we sometimes want to know whether or not observed

Table 4.3: Expected probabilities.

Education	Attitude toward premarital sex		
	Disapproval	Approval	Total
High school or less	π_{11}	π_{12}	π_{1+}
College and above	π_{21}	π_{22}	π_{2+}
Total	π_{+1}	π_{+2}	$\pi_{++}=1$

frequencies fit the null hypothesis of independence (i.e., deviating from independence only within sampling error). For our GSS example, the independence hypothesis means that education is unrelated to attitude toward premarital sex.

To test the independence hypothesis, it is useful to think of it as a special statistical model. In general, let F_{ij} denote the expected value of f_{ij} under some model. Let the expected probability associated with the cell (i,j) be denoted by π_{ij}. By definition,

$$F_{ij} = n\pi_{ij} \tag{4.1}$$

Likewise, we define π_{i+} and π_{+j} as the expected marginal probabilities of the row and column variables. For our example, the notation of expected probabilities is given in Table 4.3.

The independence model means that the joint probability π_{ij} is the product of two associated marginal probabilities:

$$\pi_{ij} = \pi_{i+}\pi_{+j} \tag{4.2}$$

Let the marginal probabilities be fitted as observed:

$$\begin{aligned}\pi_{i+} &= f_{i+}/f_{++}\\ \pi_{+j} &= f_{+j}/f_{++}\end{aligned} \tag{4.3}$$

Combining Eqs. 4.1, 4.2, and 4.3, we have

$$F_{ij} = f_{i+}f_{+j}/f_{++} \tag{4.4}$$

meaning that the expected frequency of any cell is determined by the sizes of its associated marginal totals. That is, the independence hypothesis allows for dissimilarity in the marginal distributions of the row and column variables. This makes intuitive sense. For our example in Table 4.1, we expect the cell (2,1) to be small in the absence of any relationship between education and attitude because, for our GSS sample, less than half (46%) of the respondents had attained college education, and less than half (37%) disapproved of premarital sex. Numerically, we

Table 4.4: Expected frequencies under independence.

Education	Attitude toward premarital sex		
	Disapproval	Approval	Total
High school or less	762.51	1300.49	2063
College and above	643.49	1097.51	1741
Total	1406	2398	3804

Table 4.5: Contribution to Pearson χ^2.

Education	Attitude toward premarital sex		
	Disapproval	Approval	Total
High school or less	16.01	9.39	25.40
College and above	18.97	11.12	30.10
Total	34.98	20.51	55.50

can easily solve for the predicted frequencies under the independence model, as shown in Table 4.4.

A widely used test statistic for testing the independence model is the Pearson χ^2 statistic. It is computed as

$$\chi^2 = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}(F_{ij}-f_{ij})^2}{F_{ij}} \tag{4.5}$$

with degrees of freedom equal to $(I-1)(J-1)$. Since the difference between fitted and observed frequencies, $F_{ij}-f_{ij}$, is called the residual, χ^2 statistics such as Eq. 4.5 measuring the *lack* of fit are also said to be residual-based χ^2 statistics. More will be said about degrees of freedom later in this chapter. For our example, the Pearson χ^2 statistic is 55.50 for 1 degree of freedom, which is significant beyond the 0.001 α-level, meaning that the chance of observing the actual association between education and attitude toward premarital sex in Table 4.1 is very small if the two variables are independent of each other in the population. The contribution of specific cells to χ^2 (as in Eq. 4.5) is given in Table 4.5.

Note that Eq. 4.5 is a generic formula for calculating the residual-based Pearson χ^2 statistic, although it is commonly associated with the independence model with the expected frequencies defined by Eq. 4.4. Statistical programs that compute cross-tabulations routinely report this statistic under the independence hypothesis.

4.2. Measures of Association

4.2.1. *Homogeneous Proportions*

An alternative way to express the independence model is to examine conditional proportions. This is particularly appropriate when one of the two categorical variables in a two-way cross-tabulation is an outcome variable, as in our attitude example. Table 4.6 presents the row-specific proportions for the expected frequencies under independence (i.e., Table 4.4).

We see that the proportion is the same in each row. This should not surprise us because the expected frequencies were derived under the independence model, which implies the homogeneity of proportions. Let the conditional proportion be denoted by $\pi_{j|i}$. It is easy to show that under independence

$$\pi_{j|i} = \frac{F_{ij}}{f_{i+}} = \frac{f_{i+}f_{+j}}{f_{i+}f_{++}} = \frac{f_{+j}}{f_{++}} = \pi_{+j} \tag{4.6}$$

Clearly, the independence model constrains all row-specific proportions to be equal to the marginal proportion and thus to each other. By symmetry, the same property holds true for column-specific proportions. Conversely, if the proportions are not homogeneous across rows or columns, there is dependence between the row and column variables. From an earlier test with the Pearson χ^2 statistic, we infer that the proportions are not homogeneous for our data set. To show this, Table 4.7 presents the row-specific proportions in the observed data.

Table 4.6: Row-specific proportions under independence.

Education	Attitude toward premarital sex		
	Disapproval	Approval	Total
High school or less	0.370	0.630	1.000
College and above	0.370	0.630	1.000
Total	0.370	0.630	1.000

Table 4.7: Row-specific proportions for observed data.

Education	Attitude toward premarital sex		
	Disapproval	Approval	Total
High school or less	0.423	0.577	1.000
College and above	0.306	0.694	1.000
Total	0.370	0.630	1.000

This table reveals that respondents with higher education are more likely to approve of premarital sex than are those with a high-school education or less. The Pearson χ^2 statistic reported earlier shows that this relationship is unlikely to be due to chance alone.

4.2.2. *Relative Risks*

For a dichotomous outcome variable, only one proportion is needed to summarize the information. The other is its complement and thus redundant. In general, for an outcome variable with J categories, only $J-1$ proportions are nonredundant. For our attitude example, we only need to know either the proportion of disapproval or the proportion of approval. After focusing on either of the outcome categories, it is often useful to have a summary measure of the difference by the explanatory variable. When there are only two categories for the explanatory variable, such as in our attitude example, only one summary measure is needed. One convenient measure is to take the ratio of conditional proportions, treating the first category for both row and column as the reference:

$$\frac{\pi_{2|2}}{\pi_{2|1}} \tag{4.7}$$

Formula 4.7 is called the relative risk, as defined in Chapter 3. Note that the relative risk for the first outcome category is a different but constrained number: $\pi_{1|2}/\pi_{1|1} = (1-\pi_{2|2})/(1-\pi_{2|1})$. In general, there are $(I-1)$ nonredundant comparisons for an explanatory variable with I categories. For our attitude example, the relative risk of approval between the respondents with higher education and those without higher education is $0.694/0.577 = 1.203$.

4.2.3. *Odds-Ratios*

Odds-ratios are the basic building blocks of loglinear models, since many loglinear models can be characterized in terms of odds-ratios. Before we define odds-ratios, let us first review odds. As discussed earlier in connection with the logit model, the odds are the ratio of the probability of an event occurring to the probability of the event not occurring. For the first and second rows of a 2×2 table, with $j = 2$ as the positive outcome, the odds are

$$\omega_1 = \pi_{12}/\pi_{11}$$
$$\omega_2 = \pi_{22}/\pi_{21}$$

A monotonic transformation of a probability, odds measure the likelihood that an event occurs. A higher value means that the likelihood of outcome 2 in reference to outcome 1 is higher. To measure the *relative* likelihood, we can take the ratio in odds between two categories of another (often explanatory) categorical variable, called the *odds-ratio*. Formally, the odds-ratio for a 2×2 table is

$$\theta = \frac{\omega_2}{\omega_1} = \frac{\pi_{22}/\pi_{21}}{\pi_{12}/\pi_{11}} = \frac{\pi_{22}\pi_{11}}{\pi_{12}\pi_{21}} = \frac{F_{11}F_{22}}{F_{12}F_{21}} \tag{4.8}$$

Although formula 4.8 refers to expected frequencies (F's), observed frequencies (f's) are also often used in practice. When observed frequencies are used, the resulting odds-ratios are called *observed odds-ratios*. Note that odds-ratios are always positive, varying in the range $(0, \infty)$. Like all relative measures, the interpretation of odds-ratios depends on the choice of reference categories. As defined by Eq. 4.8, an odds-ratio higher than 1 means that the second categories of the row and column variables, or conversely, the first categories of the row and column variables, are positively associated. An odds-ratio of 1 indicates a null relationship between the two variables, corresponding to statistical independence. It is often convenient to take the natural logarithm of an odds-ratio to convert it to a log-odds-ratio (LOR). LOR vary in the range $(-\infty, \infty)$, with 0 corresponding to independence. For our attitude example, the odds-ratio is 1.663, and the LOR $= 0.508$.

For a 2×2 table, there is only one meaningful odds-ratio. Rearranging reference categories yields either the same odds-ratio or its reciprocal. Owing to formula 4.8, odds-ratios are also called cross-product ratios, with the product across the main diagonal as the numerator. For a general two-way table of dimension $I \times J$, there are $(I-1)(J-1)$ nonredundant odds-ratios, from which other odds-ratios can be derived. For convenience, we define as the basic nonredundant odds-ratios those from the $(I-1) \times (J-1)$ 2×2 subtables with adjacent rows and columns. Let θ_{ij} denote these "local odds-ratios," defined as

$$\theta_{ij} = \frac{F_{ij}F_{(i+1)(j+1)}}{F_{i(j+1)}F_{(i+1)j}}, \quad i = 1, \ldots, I-1; \quad j = 1, \ldots, J-1 \tag{4.9}$$

To see how these fundamental local odds-ratios constrain other odds-ratios, first recognize that for an $I \times J$ table, there are many possible odds-ratios, since each odds-ratio involves the combination of two categories of the row variable and two categories of the column variable. For illustration, we use a fuller table for our attitude example (Table 4.8), which is based on four categories of education and four Likert-scale categories of attitude toward premarital sex.

For this 4×4 table, we can easily calculate the nine local odds-ratios, each involving a 2×2 subtable with adjacent rows and columns. They are reported in Table 4.9. Any other odds-ratio can be derived from the local odds-ratios. For example, say that we wish to know the odds-ratio involving rows 2 and 3 and columns 2 and 4. Using the notation of Eq. 4.9, we see that (by multiplying $F_{23}F_{33}$ in both numerator and denominator)

$$\begin{aligned}\frac{F_{22}F_{34}}{F_{32}F_{24}} &= \frac{F_{22}F_{34}F_{23}F_{33}}{F_{32}F_{24}F_{23}F_{33}} \\ &= \left(\frac{F_{22}F_{33}}{F_{23}F_{32}}\right)\left(\frac{F_{23}F_{34}}{F_{33}F_{24}}\right) \\ &= \theta_{22}\theta_{23}\end{aligned} \tag{4.10}$$

As an exercise, verify that Eq. 4.10 is true numerically using the observed frequencies in Table 4.8.

Table 4.8: Full table for the attitude example.

Education	**Premarital sex is**			
	Always wrong (1)	**Almost always wrong (2)**	**Sometimes wrong (3)**	**Not wrong at all (4)**
Less than high school (1)	332	99	141	311
High school (2)	313	129	258	480
Some college (3)	199	87	218	423
College and above (4)	176	71	208	359

Table 4.9: Local odds-ratios based on adjacent rows and columns.

Education	**Attitude toward premarital sex**		
	C: 2 versus 1	**C: 3 versus 2**	**C: 4 versus 3**
R: 2 versus 1	1.382	1.404	0.843
R: 3 versus 2	1.061	1.253	1.043
R: 4 versus 3	0.923	1.169	0.890

4.2.4. *The Invariance Property of Odds-Ratios*

Odds-ratios are invariant to changes in (1) the total sample size, (2) the row marginal distribution, and (3) the column marginal distribution. This can be easily demonstrated with the following example. Say that for the following 2 × 2 table, the observed frequencies are denoted by f's, and the odds-ratio by θ:

$$\begin{matrix} f_{11} & f_{12} \\ f_{21} & f_{22} \end{matrix}$$

$$\theta = \frac{f_{11}f_{22}}{f_{12}f_{21}}$$

If we change the sample size by a factor of c, all the frequencies are changed by the same factor c, but the odds-ratio remains unchanged:

$$\begin{matrix} cf_{11} & cf_{12} \\ cf_{21} & cf_{22} \end{matrix}$$

$$\theta = \frac{cf_{11}cf_{22}}{cf_{12}cf_{21}} = \frac{f_{11}f_{22}}{f_{12}f_{21}}$$

If we alter the distribution of the row marginals so that the first row total is changed by a factor of c, and the second row is changed by a factor of d, the odds-ratio still remains unchanged:

$$\begin{matrix} cf_{11} & cf_{12} \\ df_{21} & df_{22} \end{matrix}$$

$$\theta = \frac{cf_{11}df_{22}}{cf_{12}df_{21}} = \frac{f_{11}f_{22}}{f_{12}f_{21}}$$

Likewise, if we alter the distribution of the column marginals so that the first column total is changed by a factor of c, and the second column by a factor of d, we have the same result:

$$\begin{matrix} cf_{11} & df_{12} \\ cf_{21} & df_{22} \end{matrix}$$

$$\theta = \frac{cf_{11}df_{22}}{df_{12}cf_{21}} = \frac{f_{11}f_{22}}{f_{12}f_{21}}$$

In general, odds-ratios are invariant to changes in marginal distributions, since such changes in marginal distributions are translated to proportional increases or

decreases across rows or columns. This invariance property makes odds-ratios the measure of choice in studies that wish to partial out differences in marginal distributions (e.g., Featherman, Jones, & Hauser, 1975). It is due to this invariance property that maximum likelihood estimation for simple random samples can be directly applied to situations where samples are stratified on either the explanatory variable (i.e., stratified samples) or the outcome variable (i.e., case-control studies), as far as the estimation of odds-ratios is concerned (Xie & Manski, 1989).

Odds-ratios are closely related to the independence model, which allows free marginal distributions. As we commented earlier, the independence model for an $I \times J$ table has $(I-1)(J-1)$ degrees of freedom, or $(I-1)(J-1)$ constraints. We now can be more explicit: the independence model specifies that the $(I-1)(J-1)$ nonredundant odds-ratios are equal to 1. Rejection of the independence model implies that some of these odds-ratios are not equal to 1. This explains why most models for contingency tables begin with independence as the baseline model (i.e., controlling for the marginal distributions) and more complicated models can often be expressed in terms of odds-ratios.

4.3. Estimation and Goodness-of-Fit

For any given data, there are always many potential models. Which model the researcher believes to be true depends to a large extent on the researcher's theoretical assumptions and beliefs. However, more often than not, several models are equally appealing on theoretical grounds. To aid in such situations, the researcher may wish to estimate the different models and use empirical tests to assess their relative plausibility. Thus, it is important that we discuss estimation and measures for evaluating a model's goodness-of-fit.

4.3.1. Simple Models and the Pearson χ^2 Statistic

Equation 4.5 provides the general formula for computing the Pearson χ^2 statistic. Although it is most commonly used for computing the Pearson χ^2 statistic under the independence model, it can be used in connection with any model. Let us use the earlier attitude example as an illustration, focusing on model constraints and degrees of freedom. From our earlier discussion, we already know that the independence model does not fit the data. In this subsection, we will discuss some more naive but simpler models. Let us first consider the model of "equal probability" (Model A). The equal probability model says that the distribution of frequencies is equal for all cells:

$$F_{ij} = F$$

Table 4.10: Pearson χ^2 components under model A.

Education	Attitude toward premarital sex		
	Disapproval	Approval	Total
High school or less	6.40	60.06	66.46
College and above	183.73	69.45	253.18
Total	190.12	129.52	319.64

The constraints of the model are such that the expected cell frequencies do not vary either with i or with j. The estimation is simple:

$$\pi_{ij} = \frac{1}{4} \text{ or } \widehat{F}_{ij} = \frac{n}{4} = 951$$

The model consumes 1 degree of freedom for estimating the mean frequency. Thus, $df = 4 - 1 = 3$. Applying formula 4.5 for computing the Pearson χ^2 statistic for this model, we obtain the results in Table 4.10.

Obviously, the model provides a very poor fit to the data. The model is "naive" in the sense that it does not recognize the unequal distributions across categories, either for the explanatory variable (education) or the outcome variable (attitude).

We now consider the model of "equal probability conditional on column" (Model B). The model specifies that the distribution of frequencies is equal for both cells within each column:

$$\pi_{i|j} = \frac{1}{2}$$

From this, it is easy to derive the estimates of $\widehat{F}_{ij}$, which is invariant with i:

$$\pi_{ij} = \pi_{+j}\pi_{i|j} = \frac{\pi_{+j}}{2}; \quad \widehat{F}_{ij} = \frac{\widehat{F}_{+j}}{2} = \frac{f_{+j}}{2}$$

The last equality holds because we fit the column marginals exactly. Thus, in this model, 2 degrees of freedom are used for two column marginal totals: $df = 4 - 2 = 2$. The calculated Pearson χ^2 statistic from the model is 82.35. Model B is an improvement over Model A in that it takes into account the lower proportion of respondents in the first category of the outcome variable than in the second category. The marginal distribution of the education variable is left unaccounted for.

Similarly, we may also fit a naive model of "equal probability conditional on row" (Model C). Model C would fit the row marginals but not the column marginals (with Pearson χ^2 statistic being 310.41 with $df = 2$). If we fit both row and column marginal distributions, the model becomes the "independence" model (Model D). The Pearson χ^2 statistic under Model D is 55.50 with 1 degree of freedom, as shown before. Only 1 degree of freedom remains because 3 degrees of freedom are consumed for the independence model: total sample size, the column proportion (either π_{+1} or π_{+2}), and the row proportion (either π_{1+} or π_{2+}): $df = 4 - 3 = 1$.

From the preceding examples, we see that the Pearson χ^2 statistic is very easy to compute. The major step is to obtain expected frequencies ($\widehat{F}$'s). Once $\widehat{F}$'s are known, it is straightforward to apply formula 4.5. Under the hypothesis that the model is true, the Pearson χ^2 statistic is asymptotically distributed as χ^2. The researcher can compare the calculated χ^2 with the corresponding critical value from a χ^2 table.

4.3.2. *Sampling Models and Maximum Likelihood Estimation*

The class of models that we wish to introduce in this chapter are called *loglinear models*, with $\log(F)$ as the dependent variable. As discussed in an earlier chapter, such nonlinear models are best estimated by using maximum likelihood or, equivalently, by using iterative reweighted least squares for generalized linear models. The key requirement for ML estimation, however, is the prior knowledge of (or assumptions about) the sampling distribution for the stochastic component of a model. For the problem of contingency tables, three sampling models are usually invoked.

4.3.2.1. Poisson The Poisson model is the most natural sampling model for observed counts within a fixed space and time. It is one of the simplest distributions with a single parameter, λ. If variable f follows a Poisson distribution with $\lambda = F$, its probability mass function is

$$p(f|F) = \frac{\exp(-F)F^f}{f!} \quad \text{for } f = 0, 1, 2, \ldots \tag{4.11}$$

Equation 4.11 ensures that $E(f) = \text{var}(f) = F$. Note that we are omitting subscripts for f and F, respectively, denoting observed and expected frequencies, since they may contain two or more subscripts for multiple dimensions of a contingency table. What is important here is the assumption that the frequency count defined within each cross-classified cell follows an independent Poisson distribution. Examples are accidents, arrests, scientific discoveries, and births. In the context of contingency tables, we assume that these counts are cross-classified by certain characteristics defining either the outcome or the explanatory variables (such as age and education). The sum of independent Poisson-distributed variables, the total sample size (n), is a random variable that is also distributed as Poisson.

4.3.2.2. Multinomial The multinomial distribution is a generalization of the binomial distribution. If the total sample size n is fixed, the distribution of

n observations into multiple categories can be thought of as following a multinomial distribution. When we use notation for a two-way table, the probability mass function of a multinomial distribution is

$$p(f_{11},\ldots,f_{IJ}) = \frac{n!}{\prod_{i=1}^{I}\prod_{j=1}^{J} f_{ij}!}\prod_{i=1}^{I}\prod_{j=1}^{J}\pi_{ij}^{f_{ij}} \tag{4.12}$$

Equation 4.12 is an extension of the more familiar binomial distribution. The uncertainty here does not pertain to sample size, as in the case of the Poisson model, but to the assignment of elements in a sample of fixed size into a classification. For example, the researcher may create a grid for classifying marital status by gender, with marital status categorized as single, married, divorced/separated, and widowed. Every sampled individual in a survey has to fall into one of the eight mutually exclusive categories.

4.3.2.3. Product-Multinomial Fixing the total sample size is very often insufficient. In many studies, particularly in experiments or in stratified samples, the marginal totals of the different categories of the explanatory variable are fixed by design. In other situations, marginal totals of the outcome categories are also fixed either by study design or by research needs. For example, situations like this occur when a sample is drawn using disproportionate stratification. Thus, it is sometimes more natural to condition a contingency table on the marginal totals of either the row or the column variable. When this is the case, the sampling distribution reduces to an independent multinomial distribution within a broader class, which is the same as the product-multinomial sampling for the entire dataset. When we condition on row totals, for example, the probability mass function is

$$p(f_{i1},\ldots,f_{iJ}) = \frac{n_{i+}!}{\prod_{j=1}^{J} f_{ij}!}\prod_{j=1}^{J}\pi_{j|i}^{f_{ij}} \tag{4.13}$$

Fortunately, maximum likelihood estimation under the three sampling models is identical (see e.g., Fienberg, 1980, pp. 167–170). Appendix B outlines the procedures involved under Poisson sampling. The main difference among the three sampling models is the treatment of the grand total and marginal totals. In practice, this distinction is inconsequential, since researchers usually include parameters to fit the grand total and marginal totals exactly. Thus, it is not necessary to choose a particular sampling model, so long as the marginal totals are fitted.

4.3.3. *The Likelihood-Ratio χ^2 Statistic*

As reviewed in Appendix B, maximum likelihood estimation yields parameter estimates that maximize the joint probability of all observed events occurring. Since it is common that only a part of the likelihood function involves unknown parameters, we can focus on just this part, called the *kernel*, and ignore the rest. For example, under multinomial sampling (Eq. 4.12), the likelihood function is proportional to

$$\prod_{i=1}^{I}\prod_{j=1}^{J}\pi_{ij}^{f_{ij}}, \quad \text{where all } \pi_{ij} \geq 0 \text{ and } \sum_{i=1}^{I}\sum_{j=1}^{J}\pi_{ij} = 1 \tag{4.14}$$

Formula 4.14 is the kernel. The part of the likelihood function containing a ratio in factorials does not involve unknown parameters and thus can be left out of the expression for the kernel. When the kernel is maximized, the likelihood is maximized as well. In practice, we maximize the logarithm of the likelihood function and thus the logarithm of the kernel for ease of computation. Let M_r denote a restricted model and M_u the saturated model. Under M_r, let us denote the ML estimate of π_{ij} by $\widehat{\pi}^r_{ij}$, and the ML estimate of F_{ij} by $\widehat{F}^r_{ij}$, or simply $\widehat{F}_{ij}$ when there is no confusion. $\widehat{F}^r_{ij} = \widehat{\pi}^r_{ij}n$. Under M_u, $\widehat{\pi}^u_{ij} = f_{ij}/n$ and $\widehat{F}^u_{ij} = f_{ij}$. The ratio in the kernel between M_r and M_u is

$$Q = \frac{\prod_{i=1}^{I}\prod_{j=1}^{J}(\widehat{\pi}^r_{ij})^{f_{ij}}}{\prod_{i=1}^{I}\prod_{j=1}^{J}(\widehat{\pi}^u_{ij})^{f_{ij}}} = \frac{\prod_{i=1}^{I}\prod_{j=1}^{J}(\widehat{F}^r_{ij}/n)^{f_{ij}}}{\prod_{i=1}^{I}\prod_{j=1}^{J}(\widehat{F}^u_{ij}/n)^{f_{ij}}} = \prod_{i=1}^{I}\prod_{j=1}^{J}(\widehat{F}_{ij}/f_{ij})^{f_{ij}} \tag{4.15}$$

$Q \leq 1$, as a restriction can only deteriorate the goodness-of-fit. Now let us define the test statistic G^2 (sometimes also denoted as L^2) as

$$\begin{aligned} G^2 = -2\log Q &= -2\sum_{i=1}^{I}\sum_{j=1}^{J} f_{ij}\log(\widehat{F}_{ij}/f_{ij}) \\ &= 2\sum_{i=1}^{I}\sum_{j=1}^{J} f_{ij}\log(f_{ij}/\widehat{F}_{ij}) \end{aligned} \tag{4.16}$$

The statistic G^2 is called the likelihood-ratio χ^2 statistic. Always nonnegative, G^2 is asymptotically distributed as χ^2 under the assumption that the restricted model is true. The degrees of freedom can be calculated as the difference between the number of cells to begin with (i.e., IJ) and the number of parameters fitted.

In general, G^2 can be seen as the difference in $-2\times$ log-likelihood:

$$G^2 = -2(\log L_r - \log L_u) \tag{4.17}$$

where $\log L_r$ and $\log L_u$ respectively denote the log-likelihood for the restricted and unrestricted models. When G^2 is reported for a model without an explicit unrestricted model, the implicit reference is the saturated model. In this case, G^2 is also called the *scaled deviance* in the output of several computer packages. For the saturated model, $G^2 = 0$. When the researcher is interested in testing the statistical significance of the difference between two nested models, it does not matter whether he/she works directly from the general formula Eq. 4.17 or indirectly from the difference in G^2 between the two models, since the two formulas lead to the same χ^2 test.

For the independence hypothesis ($\pi_{ij} = \pi_{i+}\pi_{+j}$ for all i and j), for example, the likelihood is maximized when $\widehat{\pi}_{i+} = f_{i+}/n$ and $\widehat{\pi}_{+j} = f_{+j}/n$. We can then calculate G^2 according to Eq. 4.16. G^2 is distributed as χ^2 with $(I-1)(J-1)$ degrees of freedom if the null hypothesis is true. For our example of attitude toward premarital sex (2×2 version in Table 4.1), $G^2 = 55.89$ for 1 degree of freedom. The Pearson χ^2 statistic (χ^2) and the likelihood-ratio χ^2 statistic (G^2) are asymptotically equivalent.

4.3.4. *Bayesian Information Criterion*

The use of the G^2 statistic as a goodness-of-fit measure has been criticized by Raftery (1986, 1995) as an unsatisfactory procedure for rejecting one model in favor of another in large samples. The essence of the argument is that, when the sample size is large, it is much easier to accept (or at least harder to reject) more complex models because the likelihood-ratio test (G^2) is designed to detect *any* departure between a model and observed data. Adding more terms to a model will always improve the fit, but with large samples it becomes harder to distinguish a "real" improvement in fit from a trivial one. In this sense, the likelihood-ratio test often rejects acceptable models. One solution to this problem is to use the BIC (Bayesian Information Criterion) statistic in searching for parsimonious models that provide an "adequate" fit to the data.

The BIC index provides an approximation to a $-2\times$ log-transformed *Bayes factor*, which may be viewed as the ratio in likelihood between one model (M_0) and another model (M_1). The basic idea is to compare the relative plausibility of two models rather than to find the absolute deviation of observed data from a particular model. In practice, the researcher often chooses the saturated model to be M_1 as the reference in assessing the adequacy of M_0. The statistical methods for calculating the Bayes factor are complicated and beyond the scope of this book. Many applied researchers have found the BIC statistic popularized by Raftery (1986, 1995) to be useful. It is defined as

$$\text{BIC} = G^2 - df \log n \tag{4.18}$$

This expression shows that BIC penalizes G^2 more, per degree of freedom, for a larger sample than for a smaller sample, at the rate of $\log n$. A smaller value of BIC for model M_0 means that model M_0 is more likely than model M_1 so that the researcher should choose M_0 over M_1. When comparing multiple models, a lower value of BIC means a better-fitting model. However, the researcher should not blindly rely on BIC as the sole criterion for model selection, since it is based on an approximation. In practice, we recommend that the researcher consider a variety of goodness-of-fit criteria, including but not restricted to G^2 and BIC. Numerical examples will be given later.

It should be noted that for models estimated using maximum likelihood on *individual-level* data, such as the binary response models discussed in Chapter 3, implicit comparison is made against the null model, not the saturated model. In this case, BIC is defined as

$$\text{BIC} = -2\log L + df \log n \tag{4.19}$$

4.4. Models for Two-Way Tables

4.4.1. *The General Setup*

There are two ways to express the loglinear model for a contingency table. Let R denote row, C denote column, and f_{ij} ($i = 1, \ldots, I$; $j = 1, \ldots, J$) denote the observed frequency for the ith row and the jth column. We begin with the multiplicative version of the model, with the expected frequency (F_{ij}) specified as a function of multiplicative terms:

$$F_{ij} = \tau\tau_i^R\tau_j^C\tau_{ij}^{RC} \tag{4.20}$$

where the τ parameters are subject to normalization constraints to be discussed in depth later. With the ANOVA-like normalization constraints (i.e., τ's multiply to one along all appropriate dimensions), τ represents the (unweighted) grand mean; τ^R and τ^C represent respectively the marginal effects of R and C; and τ^{RC} represents the two-way interaction between R and C. As will be shown later, the interaction τ^{RC} parameters essentially measure the odds-ratios between R and C. When $\tau_{ij}^{RC} = 1$ for all i and j, the model is the familiar independence model.

Given that frequency is always positive, we further restrict the τ parameters to be positive. A τ parameter greater than 1 raises the expected frequency, and a τ parameter less than 1 lowers it. A τ of 1 does not affect the expected frequency at all. Since multiplication is harder to work with than addition, we can transform the multiplicative version of the model in Eq. 4.20 into the log-additive form:

$$\begin{aligned}\log F_{ij} &= \log(\tau) + \log(\tau_i^R) + \log(\tau_j^C) + \log(\tau_{ij}^{RC}) \\ &= \mu + \mu_i^R + \mu_j^C + \mu_{ij}^{RC}\end{aligned} \tag{4.21}$$

Note the one-to-one correspondence between the τ parameters in the multiplicative form (Eq. 4.20) and the μ parameters in the log-additive form of Eq. 4.21. Since loglinear is a more familiar and more general term than log-additive, we will refer to the second version as the loglinear form.

4.4.2. *Normalization*

Not all parameters in Eqs. 4.20 and 4.21 are uniquely identified. This is no different from situations in linear regressions where the researcher can use only up to $J-1$ dummy variables for a nominal independent variable with J categories. For an $I \times J$ table, the upper limit of the number of parameters identifiable is given in Table 4.11.

This upper limit is achieved when a saturated model is fitted. There are many different ways to normalize the parameters to achieve identification. Some of them are not widely used by researchers but can be useful in certain situations. For example, the parameter for the grand total sometimes can be conveniently deleted in order to identify uniquely I row marginal totals or J column marginal totals.

For most applications, two conventions are used. One is ANOVA-type coding (also referred to as effect coding) that preserves the meaning of the grand total:

$$\prod_i \tau_i^R = \prod_j \tau_j^C = \prod_i \tau_{ij}^{RC} = \prod_j \tau_{ij}^{RC} = 1 \quad \text{or}$$
$$\sum_i \mu_i^R = \sum_j \mu_j^C = \sum_i \mu_{ij}^{RC} = \sum_j \mu_{ij}^{RC} = 0 \tag{4.22}$$

The second convention is dummy-variable coding, which is equivalent to blocking out one category for R and C. Let us block out the first category for both:

$$\tau_1^R = \tau_1^C = \tau_{1j}^{RC} = \tau_{i1}^{RC} = 1 \quad \text{or}$$
$$\mu_1^R = \mu_1^C = \mu_{1j}^{RC} = \mu_{i1}^{RC} = 0 \tag{4.23}$$

Table 4.11: Identifiable parameters.

Type of parameters	Notation in τ	Notation in μ	Number of parameters
Grand total	τ	μ	1
Row marginals	τ_i^R	μ_i^R	$I-1$
Column marginals	τ_j^C	μ_j^C	$J-1$
Interactions	τ_{ij}^{RC}	μ_{ij}^{RC}	$(I-1)(J-1)$
Sum			IJ

Which normalization system to use often depends on the computer program. ECTA, for example, uses ANOVA coding. Most other computer programs, such as GLIM, Stata, R, S-Plus, and SAS (`proc genmod`) use dummy-variable coding. For empirical examples in this chapter, we will use the dummy-variable normalization with the first category as reference, unless otherwise stated. However, it is important to realize that the difference between the two normalization systems is an arbitrary one. The reader should be able to go back and forth between the two. Let us take a simple example to illustrate this point in the linear regression context. Say that we have a dichotomous variable sex (male versus female). We can create two dummy variables with the following design matrix:

Sex	x_1	x_2
Male	1	0
Female	0	1

However, in general, we cannot use both x_1 and x_2 because they are redundant when an intercept is included in the model. For any data set coded in this manner, $x_1 + x_2 = x_0$, where x_0 is a vector of ones. Thus, using x_1 and x_2, together with an intercept term (β_0), introduces perfect multicollinearity in the model. There are a number of ways to remedy this problem. A common dummy-variable normalization sets $\beta_1 = 0$ so that $\beta_0 + \beta_1 x_1 + \beta_2 x_2 = \beta_0 + \beta_2 x_2$. In contrast, the ANOVA-type normalization uses both x_1 and x_2 with the constraint that $\beta_1 + \beta_2 = 0$ so that $\beta_1 = -\beta_2$.

In general, a normalization takes the form

$$\sum_{k=1}^{K} w_k \beta_k = 0$$

where K is the number of categories, and w_k the category-specific weight. For example, the dummy-variable normalization is achieved by setting $w_1 = 1$; $w_k = 0$ for $k \neq 1$. The usual (unweighted) ANOVA-type normalization is achieved by setting $w_k = 1$, for all k. Sometimes, w_k is set to the sample proportion in the kth category in the marginal distribution of the variable to enable the interpretation of the intercept as the weighted grand mean.

In short, both the dummy-variable and ANOVA-type normalizations identify parameters of a loglinear model. Depending on the particular normalization, resulting parameters will be different. However, there should not be any substantive difference due to the choice of normalization *per se*.

4.4.3. *Interpretation of Parameters*

Loglinear models are different from linear regressions in that the "dependent" variable in loglinear models is the frequency rather than the outcome variable.

That is, the outcome variable and the explanatory variable appear in a loglinear model symmetrically. It is up to the researcher to infer causal association between them, if it is present, from model parameters. This fact has certain implications for appropriate interpretations of loglinear parameters. Loglinear models often contain many parameters. The researcher should mentally separate "nuisance" (or uninteresting) parameters from substantively meaningful parameters. In most applications, the substantively meaningful parameters are interaction parameters. This is true because the "main" effect parameters serve to saturate, or fit exactly, the marginal distributions of the row and column variables.

One way to interpret loglinear parameters is to consider conditional odds. For example, assume that the row variable is the explanatory variable and that the column variable is the outcome variable. Let j and j' denote two arbitrary categories for the column (outcome) variable. We have

$$\begin{aligned}\log\left(\frac{\pi_{j|i}}{\pi_{j'|i}}\right) &= \log\left(\frac{F_{ji}}{F_{j'i}}\right) = \log(F_{ji}) - \log(F_{j'i})\\ &= \mu + \mu_i^R + \mu_j^C + \mu_{ij}^{RC} - (\mu + \mu_i^R + \mu_{j'}^C + \mu_{ij'}^{RC})\\ &= \mu_j^C - \mu_{j'}^C + \mu_{ij}^{RC} - \mu_{ij'}^{RC}\end{aligned} \tag{4.24}$$

If the researcher uses the dummy-variable normalization with j' as the reference category, Eq. 4.24 simplifies to $\mu_j^C + \mu_{ij}^{RC}$. Under the independence model, $\mu_{ij}^{RC} = 0$, and the marginal parameters for the column variable (μ_j^C's) define conditional odds or log-odds, as well as marginal odds or log-odds due to the homogeneity of proportions property.

If the independence model does not hold, then the conditional odds vary by row, with the amount of variation determined by the interaction parameters (μ_{ij}^{RC}'s). In general, marginal parameters absorb marginal distributions, and two-way interaction parameters measure two-way associations. In fact, two-way interaction parameters correspond directly to LOR measures. For $\log\theta$ involving the four cells in the subtable of rows (i,i') and columns (j,j'):

$$\begin{aligned}\log\theta &= \log\frac{F_{ij}F_{i'j'}}{F_{ij'}F_{i'j}} = \log F_{ij} + \log F_{i'j'} - \log F_{ij'} - \log F_{i'j}\\ &= (\mu + \mu_i^R + \mu_j^C + \mu_{ij}^{RC}) + (\mu + \mu_{i'}^R + \mu_{j'}^C + \mu_{i'j'}^{RC})\\ &\quad - (\mu + \mu_i^R + \mu_{j'}^C + \mu_{ij'}^{RC}) - (\mu + \mu_{i'}^R + \mu_j^C + \mu_{i'j}^{RC})\\ &= \mu_{ij}^{RC} + \mu_{i'j'}^{RC} - \mu_{ij'}^{RC} - \mu_{i'j}^{RC}\end{aligned} \tag{4.25}$$

If dummy-variable coding is used for normalization, and i' and j' happen to be the reference categories for the row and column variables, Eq. 4.25 simplifies to μ_{ij}^{RC}. That is, with dummy-variable coding, two-way interaction parameters represent the LOR between the current row and column categories and their reference categories. LORs involving any other two pairs can be easily obtained according to Eq. 4.25.

4.4.4. *Topological Model*

With the marginal totals fitted exactly, the substantive interest of a researcher in a two-way table lies in the association between R and C. This interest is represented as two-way interaction parameters (τ_{ij}^{RC} in Eq. 4.20 and μ_{ij}^{RC} in Eq. 4.21). One of the easiest ways to understand this association is to estimate all nonredundant interaction parameters in a saturated model. For an $I \times J$ table, this means that we can estimate the local odds-ratios for the $(I-1)\times(J-1)$ 2×2 subtables with adjacent rows and columns. Let us take a closer look at the following example of an intergenerational social mobility table (Table 4.12) cross-tabulating son's occupation by father's occupation (taken from Hauser, 1979).

Let us now estimate the saturated model for this table using a dummy-variable normalization with the first row and the first column as reference categories. The estimated μ_{ij}^{RC} coefficients are given in Table 4.13 (with asymptotic standard errors in parentheses).

Interpretation of the estimates in terms of odds-ratios (Eq. 4.25) should be straightforward. We observe, for example, the likelihood of son's occupation in farm highly depends on father's occupation in farm. For concreteness, contrasting the last and the first columns and the last and the second rows, we obtain the logged ratio in odds of being a farm worker rather than an upper nonmanual worker between the son of a farm worker and the son of a lower nonmanual worker:

$$\log\frac{F_{55}/F_{51}}{F_{25}/F_{21}} = 5.065 + 0 - 0.852 - 0 = 4.213$$

The saturated model is not very interesting because it is not parsimonious. In searching for parsimonious models, the researcher may group cells with similar

Table 4.12: Hauser's mobility table.

Father's occupation	**Son's occupation**[a]				
	(1)	**(2)**	**(3)**	**(4)**	**(5)**
Upper nonmanual (1)	1414	521	302	643	40
Lower nonmanual (2)	724	524	254	703	48
Upper manual (3)	798	648	856	1676	108
Lower manual (4)	756	914	771	3325	237
Farm (5)	409	357	441	1611	1832

[a]Son's occupation (column) is defined in the same way as father's occupation (row).

Table 4.13: Interaction parameters of the saturated model: intergenerational mobility example.

Father's occupation	Son's occupation				
	(1)	(2)	(3)	(4)	(5)
Upper nonmanual (1)	0	0	0	0	0
	—	—	—	—	—
Lower nonmanual (2)	0	0.675	0.496	0.759	0.852
	—	(0.077)	(0.097)	(0.071)	(0.219)
Upper manual (3)	0	0.790	1.614	1.530	1.565
	—	(0.074)	(0.080)	(0.064)	(0.190)
Lower manual (4)	0	1.188	1.563	2.269	2.405
	—	(0.071)	(0.081)	(0.062)	0.177)
Farm (5)	0	0.862	1.619	2.159	5.065
	—	(0.089)	(0.089)	(0.093)	(0.169)

values of odds-ratios into a type, or level, and thus map out the interaction parameters into a topological pattern or levels. Hauser (1979), for example, designed the following matrix based on the observed odds-ratio pattern in Table 4.12:

$$\begin{matrix} 2 & 4 & 5 & 5 & 5 \\ 3 & 4 & 5 & 5 & 5 \\ 5 & 5 & 5 & 5 & 5 \\ 5 & 5 & 5 & 4 & 4 \\ 5 & 5 & 5 & 4 & 1 \end{matrix}$$

The values in this matrix delineate unique interaction parameters. A model that fits such a levels matrix for two-way interactions with row and column marginals fitted is called a *levels model*, or a *topological model*. If we use dummy-variable coding with category 1 as the reference category, four levels parameters will be estimated. Let us denote them as μ_2^h, μ_3^h, μ_4^h, and μ_5^h. Our estimation yields $G^2 = 66.57$ for 12 degrees of freedom, a huge improvement over the independence model with $G^2 = 6170.1$ for 16 degrees of freedom. Given the large sample size (19,912), Hauser's topological model fits the data well (BIC = -52.22).[2] The estimated coefficients of the interaction parameters are given in Table 4.14.

2. This is not surprising given that the levels were chosen to maximize goodness-of-fit.

Table 4.14: Estimated μ^h parameters.

Parameter	Estimate	(S.E.)
μ_2^h	−1.813	(0.076)
μ_3^h	−2.497	(0.080)
μ_4^h	−2.803	(0.058)
μ_5^h	−3.403	(0.060)

A LOR involving any two rows and two columns can be obtained from these μ^h parameters. For example, consider rows 2 and 3, columns 2 and 3.

$$\log\theta_{22} = \log\frac{F_{22}F_{33}}{F_{23}F_{32}} = \mu_4^h + \mu_5^h - \mu_5^h - \mu_5^h$$
$$= \mu_4^h - \mu_5^h = -2.803 + 3.403 = 0.6$$

Note that Hauser's design matrix covers the whole table, not just the 16 cells with nonzero values in Table 4.13. In general, a design matrix for a topological model assigns levels to all cells in a table. The saturated, or full-interaction, model is a special case of the topological model with the following design matrix:

$$\begin{matrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 5 \\ 1 & 6 & 7 & 8 & 9 \\ 1 & 10 & 11 & 12 & 13 \\ 1 & 14 & 15 & 16 & 17 \end{matrix}$$

In fact, as will be shown later, many special models can be conveniently parameterized as topological models.

4.4.5. *Quasi-Independence Model*

In mobility tables and similar tables where there is a correspondence between row and column variables, diagonal cells tend to be large. That is, there is a tendency for tables to exhibit clustering along the main diagonal. Researchers in social stratification call such a tendency to cluster along diagonal cells *inheritance effects*. These large diagonal cells often contribute significantly to the poor fit of the independence model. One substantively interesting hypothesis is whether the rest of the table satisfies the independence hypothesis net of the diagonal cells. This leads to the quasi-independence model.

A square table satisfies quasi-independence if R and C are independent of each other in off-diagonal cells. That is,

$$\pi_{ij} = \pi_{i+}\pi_{+j}, \quad \text{for } i \neq j$$

Compared to the independence model, the quasi-independence model consumes I additional degrees of freedom, thus with $(I-1)(I-1)-I$ degrees of freedom for residuals. The loss of the I degrees of freedom can be interpreted either in terms of the reduction of the number of data points by I, or in terms of the increase of I additional parameters. In fact, each interpretation corresponds to an estimation method. For the first interpretation, the researcher can block out the diagonal cells (e.g., by using a weight matrix) while estimating the independence model. For the second interpretation, the researcher can add unique parameters to the diagonal cells, effectively estimating a topological model. For a 5×5 table, the design matrix is

2	1	1	1	1
1	3	1	1	1
1	1	4	1	1
1	1	1	5	1
1	1	1	1	6

The two estimation methods yield identical results. The main difference is that the second method yields estimates for diagonal cells, whereas the first does not. The second method can also be used to constrain some diagonal parameters to be equal. For this example, the quasi-independence model is a significant improvement in goodness-of-fit over that of the independence model, with $G^2 = 683.34$ for 11 degrees of freedom. As shown by Goodman (1972), this method can be used effectively to test for partial independence in limited regions or account for a few especially large cells.

4.4.6. *Symmetry and Quasi-Symmetry*

For square $I \times I$ tables, the researcher may be interested in whether or not the row and column variables are symmetric with respect to each other. The symmetry model is

$$\log F_{ij} = \mu + \mu_i + \mu_j + \mu_{ij} \tag{4.26}$$

where $\mu_{ij} = \mu_{ji}$. Here we purposely omit the superscripts R and C for the μ terms because they pertain to both R and C. The symmetry model means that all cells are symmetric to each other across the main diagonal: $F_{ij} = F_{ji}$. Obviously, it is highly

constrained and thus of limited use. The residual degrees of freedom are one-half of the number of off-diagonal cells, or $I(I-1)/2$.

We can decompose the symmetry of Eq. 4.26 into two components: marginal homogeneity and symmetric interactions. If we replace marginal homogeneity with marginal heterogeneity while retaining symmetric interactions, we have the model of quasi-symmetry:

$$\log F_{ij} = \mu + \mu_i^R + \mu_j^C + \mu_{ij}^{RC} \tag{4.27}$$

where $\mu_{ij}^{RC} = \mu_{ji}^{RC}$. That is, the quasi-symmetry model allows for marginal heterogeneity but restricts the interaction parameters to be symmetric across the main diagonal. Many researchers find the quasi-symmetry model to be more useful because it conditions on differences in marginal distribution that should be left unconstrained. Sobel, Hout, and Duncan (1985), for example, use the quasi-symmetric model to describe structural mobility with parameters measuring the difference in row and column marginal distributions. Because the quasi-symmetry model adds $I-1$ additional parameters compared to the symmetry model, the residual degrees of freedom for the quasi-symmetry model are $I(I-1)/2-(I-1)=(I-1)(I-2)/2$.

The quasi-symmetry model can be easily estimated using a topological coding. For the 5×5 case, for example, the design matrix for the interactions can be expressed as

2	1	1	1	1
1	3	7	8	9
1	7	4	10	11
1	8	10	5	12
1	9	11	12	6

For our mobility example, $G^2 = 27.45$ with 6 degrees of freedom; BIC $= -31.95$. Quasi-symmetry is more general than quasi-independence.

4.4.7. *Crossings Model*

Not all models of potential interest can be expressed in terms of topological models with a single design matrix. One such example is the crossings model (Goodman, 1972). The hypothesis implied by the crossings model is that different categories of a nominal variable present varying degrees of difficulty for crossing. The further apart two categories are for the row variable, the smaller the interaction parameter between two categories for the column variable. Formally, the crossings model simplifies Eq. 4.20 to

$$F_{ij} = \tau\tau_i^R\tau_j^C\nu_{ij}^{RC} \tag{4.28}$$

with

$$v_{ij}^{RC} = \begin{cases} \prod_{u=j}^{i-1} v_u & \text{for } i>j \\ \prod_{u=i}^{j-1} v_u & \text{for } i<j \\ \xi_i & \text{for } i=j \end{cases}$$

For a 5 × 5 table (such as the mobility example of Table 4.12), for example, the v_{ij}^{RC} interaction parameters can be displayed as

ξ_1	v_1	v_1v_2	$v_1v_2v_3$	$v_1v_2v_3v_4$
v_1	ξ_2	v_2	v_2v_3	$v_2v_3v_4$
v_1v_2	v_2	ξ_3	v_3	v_3v_4
$v_1v_2v_3$	v_2v_3	v_3	ξ_4	v_4
$v_1v_2v_3v_4$	$v_2v_3v_4$	v_3v_4	v_4	ξ_5

Note that, in this formulation, we follow Goodman (1972) in fitting the diagonals exactly, as in the quasi-independence and quasi-symmetry models. Researchers may not want to fit the diagonal cells exactly for parsimony reasons (e.g., Mare, 1991). In the loglinear form of Eq. 4.28, for cells $i \neq j$, the v_{ij}^{RC} interaction parameters can be parameterized as the sum of the coefficients of the following four sets of design matrices:

$$\begin{array}{ccccc} 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{array} \quad \begin{array}{ccccc} 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \end{array} \quad \begin{array}{ccccc} 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \end{array} \quad \begin{array}{ccccc} 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \end{array}$$

with the four matrices respectively corresponding to v_1, v_2, v_3, and v_4. With the diagonal cells blocked out by the ξ parameters, only $(I-3)$ of the $(I-1)$ v parameters in Eq. 4.28 are identified. For our 5 × 5 example, only two v parameters are identified. Goodman (1972) recommends normalizing the first and the last v: $v_1 = v_{I-1} = 1$. Without the diagonals blocked, all $(I-1)$ v's are identifiable.

One interesting feature of the crossings model is that the local odds-ratios for adjacent rows and columns not involving diagonal cells satisfy local independence. For our 5 × 5 example, let us consider the odds-ratio involving cells in rows 4, 5 and columns 1, 2:

$$\theta_{41} = \frac{F_{41}F_{52}}{F_{42}F_{51}} = \frac{(v_1v_2v_3)(v_2v_3v_4)}{(v_2v_3)(v_1v_2v_3v_4)} = 1$$

For Hauser's example of intergenerational mobility, the crossings model fits the observed data rather well. The G^2 statistic is 64.24 for 9 degrees of freedom (BIC = −24.85). The version not blocking out the diagonal cells has a G^2 statistic of 89.91 for 12 degrees of freedom (BIC = −28.88). Although the crossings model (in either version) is short of Hauser's topological model in pure goodness-of-fit, the crossings model yields estimates of parameters that may be easier to interpret. For the second version, for example, the crossings estimates in the log scale are (from the second occupational category to the last one) (−0.4256, −0.3675, −0.2935, −1.403). Thus, of all the barriers separating adjacent categories, the last barrier separating lower manual occupations and farming is the most difficult to cross, and the next most difficult barrier is the one separating upper nonmanual and lower nonmanual occupations.

4.5. Models for Ordinal Variables

So far, we have treated the row and column variables as nominal variables (i.e., discrete variables with unordered categories). In substantive applications, it is often reasonable to assume that categories are ordinal, meaning that they are ranked on either an observed or latent scale. This additional information of ordering can be used to obtain parsimonious model specifications.

Typically, researchers use ordering information to specify the interaction terms only (i.e., μ_{ij}^{RC} of Eq. 4.21), leaving the marginal distributions fitted exactly. This is a conservative approach, for the ordinal information is used only for the association between the row and column variables. As shown before, for an $I \times J$ table, there are $(I-1)(J-1)$ degrees of freedom for interactions after the marginal totals are fitted. If ordering information is used, it may take few (sometimes just 1) degrees of freedom to describe the association. Note that the payoff to ordering information goes up as the numbers of categories increase.

4.5.1. Linear-by-Linear Association

Let x_i and y_j denote respectively the measured attributes (or indexes) of the row and column variables. They can be used in the specification of a linear-by-linear association, as

$$\log F_{ij} = \mu + \mu_i^R + \mu_j^C + \beta x_i y_j \tag{4.29}$$

Compared to Eq. 4.21, the linear-by-linear association model replaces the μ_{ij}^{RC} term with a more parsimonious form $\beta x_i y_j$, where β can be seen as the association coefficient between x and y. For an odds-ratio involving any pair of rows (i and i') and any pair of columns (j and j'), Eq. 4.25 simplifies to

$$\log \theta = \log \frac{F_{ij}F_{i'j'}}{F_{ij'}F_{i'j}} = \beta(x_i - x_{i'})(y_j - y_{j'}) \tag{4.30}$$

That is, the LOR is proportional to the product of the distances of the row and column variables in index scores. Multiple linear-by-linear terms can be used, so long as they are fewer than $(I-1)(J-1)$, changing Eq. 4.29 to

$$\log F_{ij} = \mu + \mu_i^R + \mu_j^C + \sum_m \beta_m x_{im} y_{jm} \tag{4.31}$$

where x_m and y_m are the row and column attributes for the mth linear-by-linear association. The same attribute for row (or column) may be used in combination with different attributes for column (or row). Examples using this approach are found in work by Hout (1984) and Lin and Xie (1998).

For example, in Lin and Xie's (1998) model of interstate migration, state-level economic growth rates (denoted as g's) are used to capture the "push" and "pull" forces of migration. The push force of origin is measured by $1/g_i$, the pull force of destination by g_j. The model is similar to Eq. 4.29 with $(1/g_i)g_j$ as an interaction term between origin and destination. Lin and Xie find the push-and-pull interaction to be highly significant in explaining interstate migration streams in the United States.

4.5.2. *Uniform Association*

The previous discussion presumes the existence of attribute (or index) variables for the row and column variables. In the absence of such index variables, what can the researcher do? There are two answers to this question. One is to impose an interval-score structure on the categories. The second is to estimate the latent score associated with the categories. We discuss the first approach in this subsection and leave the second approach to the Section 4.5.4, "Goodman's *RC* Model."

The easiest way to impose an interval structure is to assign consecutive integers to categories, if the categories form an ordinal scale and are correctly ordered. For the example of attitude toward premarital sex (full version, Table 4.8), we may assign the scores as follows. For the outcome variable, assign Always Wrong=1, Almost Always Wrong=2, Sometimes Wrong=3, and Not Wrong at All=4. For the explanatory variable of education, assign Less than H.S.=1, High School=2, Some College =3, and College and Above=4. This method of assigning scores basically assumes that the distance between any two adjacent categories is uniform across all possible values. We call this scoring method integer-scoring, and the resulting model the uniform association model. The particular values assigned are inconsequential, so long as they are uniformly spaced. That is, (1,2,3,4) yields the same model as $(-10,-8,-6,-4)$. As a convention, however, we use consecutive integers beginning with 1. That is, set $x_i = i$ and $y_j = j$. Substituting these imposed uniform scores to Eq. 4.29 yields

$$\log F_{ij} = \mu + \mu_i^R + \mu_j^C + \beta ij \tag{4.32}$$

Consuming 1 degree of freedom for interactions, the uniform-association model has $(I-1)(J-1)-1$ degrees of freedom for residuals. A special feature of the uniform association model is that its odds-ratios involving two adjacent rows and columns are invariant. Using the constraint of Eq. 4.32 and solving for Eq. 4.9, we see that

$$\theta_{ij} = \frac{F_{ij}F_{(i+1)(j+1)}}{F_{i(j+1)}F_{(i+1)j}} = \exp(\beta) \tag{4.33}$$

and

$$\log \theta_{ij} = \beta$$

In fact, this important property of Eq. 4.32 can be used to define the uniform association model (Goodman, 1979). For an odds-ratio involving arbitrary pairs (i and i' for row, j and j' for column), the LOR is simply

$$\beta(i-i')(j-j')$$

For our example of attitude toward premarital sex, the uniform association model yields a G^2 of 31.33 with 8 degrees of freedom. The estimate of β is 0.097 with a standard error of 0.013.

The uniform association model is a special case of the linear-by-linear model in which integer-scoring is used. More generally, other scoring methods may also be reasonable. For example, midpoints or weighted means may be used to "linearize" categories that were originally interval. For the education variable in the attitude example, one may wish to assign 10 to Less Than High School, 12 to High School, 14 to Some College, and 17 to College and Above.

4.5.3. *Row-Effect and Column-Effect Models*

The uniform association model imposes integer-scoring to both the row and column variables. A less restrictive approach is to assume integer-scoring for either the row or the column variable, but not for both. When integer-scoring is used for the column variable, the resulting model is called the row-effect model. Conversely, when integer-scoring is used for the row variable, the model is called the column-effect model. These models were developed by Goodman (1979). For an innovative application, see Duncan's (1979) study of an 8×8 intergenerational mobility table.

For the row-effect model, Eq. 4.21 is simplified to

$$\log F_{ij} = \mu + \mu_i^R + \mu_j^C + j\phi_i \tag{4.34}$$

where ϕ_i can be seen as the row effect (or row score) *estimated* from the model. Assuming that the column categories are correctly ordered and approximately follow the integer-scoring scale, the row-effect model is a generalization of the uniform association model. Comparing Eq. 4.34 to Eq. 4.32 for the uniform association model, however, we see that there is no β for the row-effect model. This is due to the fact that the row effect (ϕ_i) is latent and needs to be normalized. In other words, it is not possible to separate β from $\beta\phi_i$ when ϕ_i is latent. So, we set $\beta = 1$ to normalize the scale of ϕ_i. In addition, we also need to normalize the location of ϕ_i. One convenient normalization is to use dummy-variable coding with the first category as the reference category so that $\phi_1 = 0$. With $(I-1)$ parameters for interactions between row and column, the row-effect model has $(I-1)(J-2)$ degrees of freedom. With these specifications for the *R–C* interactions, it is easy to see that, as a special variant of Eq. 4.30,

$$\log \frac{F_{ij}F_{i'j'}}{F_{ij'}F_{i'j}} = (\phi_i - \phi_{i'})(j - j'). \tag{4.35}$$

For the fundamental local LORs,

$$\log \theta_{ij} = \log\left(\frac{F_{ij}F_{(i+1)(j+1)}}{F_{i(j+1)}F_{(i+1)j}}\right) = \phi_{i+1} - \phi_i \tag{4.36}$$

Likewise, we can define the column-effect model in a similar manner, which changes Eq. 4.21 to

$$\log F_{ij} = \mu + \mu_i^R + \mu_j^C + i\varphi_j \tag{4.37}$$

where φ_j is called the column-effect and requires a normalization. The column-effect model has $(I-2)(J-1)$ degrees of freedom. Note that the column-effect model presumes that the row categories are correctly ordered and approximately follow the integer-scoring scale. For the column-effect model, the LOR structure for any two pair of categories is

$$\log \frac{F_{ij}F_{i'j'}}{F_{ij'}F_{i'j}} = (\varphi_j - \varphi_{j'})(i - i') \tag{4.38}$$

and the LOR for local subtables is

$$\log \theta_{ij} = \log\left(\frac{F_{ij}F_{(i+1)(j+1)}}{F_{i(j+1)}F_{(i+1)j}}\right) = \varphi_{j+1} - \varphi_j \tag{4.39}$$

Hence, we see similarities and differences among the uniform association, the row effect and the column effect models. The uniform association model can be viewed as a special case either of the row effect model or of the column effect model. They are far more parsimonious than the saturated model. The gain in parsimony increases rapidly with the dimension of a table. Like many parsimonious models, these three types of models can also be used in combination with the key feature of the quasi-independence model (i.e., blocking out diagonal cells (or any subsets of a table)).

Table 4.15: Goodness-of-fit statistics for mobility models.

Model specification	G^2	*df*	**BIC**	Δ^a
Independence	6170.13	16	6011.74	20.07
Row effects	2080.18	12	1961.39	12.32
Uniform association (UA)	2280.69	15	2132.21	11.98
Quasi-independence	683.34	11	574.45	5.52
Row effect, diagonals deleted	34.91	7	−34.39	1.10
UA, diagonals deleted	73.01	10	−25.98	1.95
Hauser's topological model	66.57	12	−52.22	1.77
Quasi-symmetry	27.45	6	−31.95	1.13
Crossings (diagonals kept)	89.91	12	−28.88	2.12
Crossings (diagonals blocked)	64.24	9	−24.85	1.63

[a]Δ is the index of dissimilarity between observed and predicted frequencies (in %).

We now apply the uniform association model and the row effect model to Hauser's mobility data, first by themselves and then after blocking out the diagonal cells. The row effect model is borrowed from Duncan (1979). To compare these models against alternative specifications, we also present the goodness-of-fit statistics from other models we have discussed. The results are provided in Table 4.15. The column denoted by G^2 is the likelihood-ratio χ^2 for residuals (e.g., Eq. 4.17), with degrees of freedom reported in the column labeled *df*. With $n = 19{,}912$, BIC is calculated according to Eq. 4.18. As a purely descriptive measure of goodness-of-fit, we also use the Index of Dissimilarity (Shryock & Siegel, 1976, p. 131), denoted as Δ. The Index of Dissimilarity here can be interpreted as the proportion of misclassified counts according to the expected frequencies under a model.

As shown in Table 4.15, several models other than Hauser's topological model fit the data reasonably well. For example, the row-effect and the uniform-association models fit the data reasonably well (BIC = −34.39, −25.98) with diagonal cells blocked. The other two models that fit the data well are the quasi-symmetry (BIC = −31.95) and the crossings (BIC = −28.88) models.

4.5.4. *Goodman's RC Model*

If we further follow the path of generalization from the uniform association model to the row effect and column effect models, we may want to know what happens if we treat both the row and the column scores as unknown. In an influential paper published in the *Journal of the American Statistical Association* in 1979, Goodman addresses this question. Goodman's initial solution consists of two types of models, row-and-column-effects association model I and row-and-column-effects association model II (which was renamed as the *RC* model by Goodman (1981b) and is now commonly referred to by that name).

Goodman's association model I simplifies Eq. 4.21 into

$$\log F_{ij} = \mu + \mu_i^R + \mu_j^C + j\phi_i + i\varphi_j \tag{4.40}$$

where ϕ_i and φ_j are respectively row and column scores as in the row- and column-effect models. That is, model I can be seen as specifying the interaction term μ^{RC} of Eq. 4.21 as the sum of the interaction terms of the row- and column-effect models $(j\phi_i + i\varphi_j)$. However, it is necessary to add a scale normalization to either ϕ_i or φ_j, in addition to normalizing their locations. For example, one possible normalization is $\phi_1 = \varphi_1 = \varphi_J = 0$. The degrees of freedom for residuals equal $(I-2)(J-2)$. The general formula for LOR is

$$\log\left(\frac{F_{ij}F_{i'j'}}{F_{ij'}F_{i'j}}\right) = (\phi_i - \phi_{i'})(j - j') + (\varphi_j - \varphi_{j'})(i - i') \tag{4.41}$$

which is the sum of the weighted distances between the row scores and between the column scores. The reader should compare Eq. 4.41 to Eqs. 4.35 and 4.38. Similar to the row- and column-effect models, model I also assumes that the rows and columns are correctly ordered. This property means that the model is not invariant to positional changes in the categories of the row and column variables. If the researcher has no knowledge that the categories are correctly ordered, or in fact needs to determine the correct ordering of the categories, model I is of limited use. For this reason, Goodman's model II has received the most attention. It is of the form

$$\log F_{ij} = \mu + \mu_i^R + \mu_j^C + \phi_i\varphi_j \tag{4.42}$$

where ϕ_i and φ_j are respectively row and column scores collectively requiring three normalization constraints. One possible normalization is to set the location of both ϕ_i and φ_j (e.g., $\sum\phi_i = 0$ and $\sum\varphi_j = 0$) and the scale of either ϕ_i or φ_j (say $\sum\phi_i^2 = 1$). Model II has the same degrees of freedom as model I, $(I-2)(J-2)$, for only $I+J-3$ parameters are used to describe the row–column association. The model does not require the correct ordering of either the row or the column categories. The estimation of the scores (ϕ_i's and φ_j's) reveals the ordering of the categories implicit in the model. The LOR for any two pair of categories is

$$\log\left(\frac{F_{ij}F_{i'j'}}{F_{ij'}F_{i'j}}\right) = (\phi_i - \phi_{i'})(\varphi_j - \varphi_{j'}) \tag{4.43}$$

which is the product of the distances between the row scores and between the column scores. Goodman's association model II is also called the *log-multiplicative model* (Clogg, 1982), since two-way interaction is characterized by a multiplicative term involving two unknown parameters in Eq. 4.42. This creates some difficulty for estimation, which requires an iterative procedure, since ϕ_i and φ_j cannot be separated in a single estimation. The iterative procedure alternately treats one set of estimates (or initial values), say ϕ_i's, as known in updating the other set of estimates (or initial values), say φ_j's, until they stabilize.[3] Besides the GLIM macros available from this book's website, special-purpose computer packages for this type of model are also available (such as ASSOC and *ℓEM*).

The association models proposed by Goodman are parsimonious because the number of parameters for interactions increases by $(I+J-3)$ instead of $(I-1)(J-1)$ as in the saturated case. Obviously, the parsimony of these models can be achieved only with tables of a sufficiently large dimension. As a rule, the number of categories should be at least three for such models to be applicable.

The dimensionality requirement is even clearer in light of interpretations of the estimated scores (ϕ_i's and φ_j's). As Clogg (1982) shows, the real meaning of the estimated scores lies in differences in intervals between two adjacent categories. Such differences in intervals are not meaningful for variables with less than three categories.

The log-multiplicative model of Eq. 4.42 can further be generalized to multiple dimensions, mimicking the case with multiple dimensions of observed attributes. This is called the *RC*(*m*) model, extensively discussed by Goodman (1986) and Becker and Clogg (1989). In the multiple dimension case, it is convenient to reparameterize the unknown parameters differently by adding an unknown coefficient β and renormalizing

$$\log F_{ij} = \mu + \mu_i^R + \mu_j^C + \sum_m \beta_m\phi_{im}\varphi_{jm} \tag{4.44}$$

with

$$\sum\phi_{im} = 0; \quad \sum\phi_{im}^2 = 1$$
$$\sum\varphi_{jm} = 0; \quad \sum\varphi_{im}^2 = 1$$

where β_m measures the strength of association for the *m*th dimension. For some applications, the researcher may wish to save degrees of freedom by constraining ϕ_{im} and φ_{jm} across *m*, and even between ϕ_{im} and φ_{jm} for square tables.

To illustrate the usefulness of the *RC* association model, let us take a look at one of Clogg's (1982) examples from the 1977 GSS. The tabular data are reproduced in Table 4.16. The row variable consists of patterns conforming to a Guttman scale measuring attitudes toward abortion. In parentheses are responses to questions

3. In standard packages such as GLIM, this procedure does not produce correct standard errors.

Table 4.16: Attitudes toward abortion and premarital sex.

Attitude toward abortion	Attitude toward premarital sex[a]			
	(1)	(2)	(3)	(4)
(1) Error	44	11	38	62
(2) (yes,yes,yes)	59	41	147	293
(3) (yes,yes,no)	23	11	13	27
(4) (yes,no,no)	27	8	16	27
(5) (no,no,no)	258	57	105	110

[a]For the column variable, (1) = Always Wrong; (2) = Almost Always Wrong; (3) = Sometimes Wrong; (4) = Not Wrong at All.

asking whether legal abortion should be available to a woman under three different situations: (1) if she is not married and does not want to marry the man; (2) if the family has a very low income and cannot afford any more children; and (3) if a woman is married and does not want any more children. Given the varying severity of the three situations, most respondents fall into the patterns of approval of abortion under a more severe situation if they approve of abortion under a less severe situation. The first category of "Error" consists of respondents who do not neatly fall into the Guttman scale.

As is well known, Guttman scales yield only ordinal variables. That is, for our example, we only know that respondents in category (5) disapprove of abortion more strongly than those in category (4), and those in category (4) in turn disapprove of abortion more strongly than those in category (3), and so on. We do not know the relative distances separating the various categories. In addition, we do not know where the nonconforming respondents in category (1) belong.

Clogg (1982) chose measured attitude toward premarital sex as an instrument in scaling the ordinal measure of the abortion attitude. To do this, Clogg applied the *RC* model to these data. We replicated Clogg's results using an iterative ML estimation procedure implemented as a GLIM macro, which is available from this book's website. We normalize the model using the convention in Eq. 4.44, thereby restricting both the location and scale of the row and column scores and freeing up an association parameter β. The estimated model fits the data very well ($G^2 = 5.55$ for 6 degrees of freedom; $\text{BIC} = -37.81$). The estimated scores are given in Table 4.17, with β estimated to be 1.308, which means a strong positive association between attitude toward abortion and attitude toward premarital sex.

These estimated parameters are essentially the same as those reported by Clogg, although they appear to be different due to different normalizations. The estimated scores should be interpreted in terms of relative distances. For example, the respondents in the first row category ("Error") are estimated to approve of abortion less strongly than those in category (2) but more strongly than those in other categories. It is important to emphasize that a shift of categories would not affect the estimation of the *RC* model. That is, although the *RC* model presumes the ordinal

Table 4.17: Estimated scale scores.

	Row	Column
	(Abortion)	(Premarital sex)
(1)	0.075	−0.743
(2)	0.776	−0.127
(3)	−0.098	0.271
(4)	−0.155	0.598
(5)	−0.598	—

scale of the row and column variables, it does not require the correct ordering of the categories. Estimation reveals such ordering. For our data in Table 4.16, the column categories are correctly ordered. The row categories are not.

4.6. Models for Multiway Tables

Most studies in social science are concerned with relationships among variables, for such relationships often reveal underlying social processes. Two-way tables are the most basic form of representation relating observed variables to each other. In the last two decades, social researchers have fruitfully applied the loglinear models presented in the preceding sections in analyzing associations in two-way tables.

However, two-way tables are inherently limited because they contain little information. For example, two variables may be associated due to their common association with a third variable. When the third variable is controlled, the partial association between the two variables may be nil. To test for such "omitted-variable bias," it is necessary to bring other dimensions into a multivariate study.

Another common situation in which the researcher analyzes three- or higher-way tables is when the key research interest lies in the variation of a two-way association along one or more dimensions. Examples include trend analysis and comparative analysis. We will review some examples in the sociological literature later in this section.

In the next section, we introduce loglinear models for the analysis of three- and higher-way contingency tables. We lump tables of three or higher dimensions under the general label of "multiway" tables. Although our discussion focuses only on models for three-way tables, generalization to higher-way models should be straightforward. It is also important to realize that the models for multiway tables are generalizations of the models presented earlier for two-way tables.

4.6.1. Three-Way Tables

Let *R*, *C*, and *L* respectively denote the row, column, and layer variables, with layer being the additional third variable. The three-way table of $R \times C \times L$ gives the detailed association among *R*, *C*, and *L*. In this three-way table, the researcher can obtain

Table 4.18: Graduate admission data from UC-Berkeley.

Major	Men		Women	
	Number of applicants	Percent admitted	Number of applicants	Percent admitted
A	825	62	108	82
B	560	63	25	68
C	352	37	593	34
D	417	33	375	35
E	191	28	393	24
F	373	6	341	7

partial tables between any two variables (say R and C) while holding the third variable (L) constant at a given level. The R–C association in $R \times C$ partial tables is called the partial association. When the R–C partial association varies across different categories of L, it is said that there is three-way interaction involving R, C, and L. The researcher could also ignore the third variable (say L) and collapse the three-way table ($R \times C \times L$) into a two-way table, called the *marginal* table ($R \times C$), containing the marginal association between the two variables (R and C). In general, partial associations are different from marginal associations. Otherwise, the researcher would opt for simpler tables and model them statistically. In the next subsection, we will discuss conditions under which a partial association equals a marginal association.

In Table 4.18, we present a table pertaining to data on graduate admissions at the University of California-Berkeley. Table 4.18 involves three variables: sex of applicants (men versus women), admission outcome (admitted versus rejected), and major (A through F). The data came from a study looking into the allegation that graduate admission at the University of California-Berkeley was biased in favor of men against women (Bickel, Hammel, & O'Connell, 1975; Freedman, Pisani, & Purves, 1978). For convenience, let us label sex R, admission outcome C, and major L. Although Table 4.18 presents the data in the form of proportions and counts for sex by major combinations, converting the table into frequencies by $R \times C \times L$ is easy.[4]

Table 4.18 shows clearly that the admission rate of women applicants is *not* appreciably lower than men applicants in any major. If there is a notable difference by sex, it is that women have a higher admission rate (at 82%) than men (at 62%) for major A. However, the relationship between sex and admission outcome looks very different if we collapse the data into a two-way marginal table over major. Table 4.19 is the resulting table.

Table 4.19 suggests that women have a much lower rate of admission (30%) than that of men (45%). Why do the two tables based on the same data tell us two different stories? Understanding this puzzle is essential to a multivariate analysis of tabular data.

4. The converted frequency table is available from this book's website.

Table 4.19: Collapsed graduate admission data.

Sex	No. of applicants	Percent admitted
Men	2691	45
Women	1835	30

4.6.2. *The Saturated Model for Three-Way Tables*

For the three-way table $R \times C \times L$, let f_{ijk} denote the observed frequency, and F_{ijk} the expected frequency for the cell indexed by the ith row, jth column, and kth layer. Analogous to Eq. 4.20, the saturated model for the three-way table can be written as

$$F_{ijk} = \tau\tau_i^R\tau_j^C\tau_k^L\tau_{ij}^{RC}\tau_{ik}^{RL}\tau_{jk}^{CL}\tau_{ijk}^{RCL} \tag{4.45}$$

where the τ parameters are subject to usual normalization constraints. The loglinear form of the model is

$$\log F_{ijk} = \mu + \mu_i^R + \mu_j^C + \mu_k^L + \mu_{ij}^{RC} + \mu_{ik}^{RL} + \mu_{jk}^{CL} + \mu_{ijk}^{RCL} \tag{4.46}$$

where the μ parameters are simply the logarithms of the τ parameters and are thus subject to the same normalization constraints. For the ANOVA-like normalization,

$$\begin{aligned}\prod_i \tau_i^R = \prod_j \tau_j^C = \prod_k \tau_k^L = \prod_i \tau_{ij}^{RC} = \prod_j \tau_{ij}^{RC} \\ = \prod_i \tau_{ik}^{RL} = \prod_k \tau_{ik}^{RL} = \prod_j \tau_{jk}^{CL} = \prod_k \tau_{jk}^{CL} \\ = \prod_i \tau_{ijk}^{RCL} = \prod_j \tau_{ijk}^{RCL} = \prod_k \tau_{ijk}^{RCL} = 1\end{aligned} \tag{4.47}$$

Or, in terms of the μ parameters,

$$\begin{aligned}\sum_i \mu_i^R = \sum_j \mu_j^C = \sum_k \mu_k^L = \sum_i \mu_{ij}^{RC} = \sum_j \mu_{ij}^{RC} = \sum_i \mu_{ik}^{RL} \\ = \sum_k \mu_{ik}^{RL} = \sum_j \mu_{jk}^{CL} = \sum_k \mu_{jk}^{CL} \\ = \sum_i \mu_{ijk}^{RCL} = \sum_j \mu_{ijk}^{RCL} = \sum_k \mu_{ijk}^{RCL} = 0\end{aligned} \tag{4.48}$$

Alternatively, we can use dummy-variable coding and set the following normalization constraints (with the first category as the reference):

$$\begin{aligned}\tau_1^R &= \tau_1^C = \tau_1^L = \tau_{1j}^{RC} = \tau_{i1}^{RC} = \tau_{1k}^{RL}\\ &= \tau_{i1}^{RL} = \tau_{1k}^{CL} = \tau_{j1}^{CL} = \tau_{1jk}^{RCL}\\ &= \tau_{i1k}^{RCL} = \tau_{ij1}^{RCL} = 1\end{aligned} \tag{4.49}$$

Or, in terms of the μ parameters,

$$\begin{aligned}\mu_1^R &= \mu_1^C = \mu_1^L = \mu_{1j}^{RC} = \mu_{i1}^{RC} = \mu_{1k}^{RL}\\ &= \mu_{i1}^{RL} = \mu_{1k}^{CL} = \mu_{j1}^{CL} = \mu_{1jk}^{RCL}\\ &= \mu_{i1k}^{RCL} = \mu_{ij1}^{RCL} = 0\end{aligned} \tag{4.50}$$

The τ^R, τ^C, and τ^L parameters in Eq. 4.45 (or μ^R, μ^C, and μ^L in Eq. 4.46) are called marginal parameters, τ^{RC}, τ^{RL}, and τ^{CL} (or μ^{RC}, μ^{RL}, and μ^{CL} in Eq. 4.46) two-way interactions, and τ^{RCL} (or μ^{RCL} in Eq. 4.46) three-way interactions. Since additive terms are easier to work with than multiplicative terms, the loglinear form of Eq. 4.46 is commonly used.

4.6.3. Collapsibility

Collapsibility is meaningful when research interest lies in the association between two particular variables. The question is whether the measured association differs before and after a table is collapsed from three-way to two-way. A three-way table is said to be collapsible if the partial association equals the marginal association when the three-way table is collapsed over the variable not involved in the association of primary interest. That is, a table is considered collapsible if marginal and partial relationships are identical.

To be more precise, let us say that we are primarily interested in the R–C association in the table of $R \times C \times L$. The table can be collapsed over L to the marginal two-way table of $R \times C$, if the marginal association in the $R \times C$ table is the same as the partial association between R and C conditional on L.

Conditions of collapsibility for the three-way table of $R \times C \times L$ into the two-way table of $R \times C$ follow:

1. There is no three-way RCL interaction: $\mu_{ijk}^{RCL} = 0$, for all i, j, and k.
2. *Either RL or CL* two-way interaction is nil: either $\mu_{ik}^{RL} = 0$ or $\mu_{jk}^{CL} = 0$, for all i, j, and k.

To appreciate these two conditions, let us review conditions for omitted variable bias in the linear regression context: an omitted variable may cause a bias to the

estimated effect of the primary variable of interest on the dependent variable only if both of the following two conditions are true:

1. The omitted variable is (unconditionally) related to the primary variable of interest.
2. The omitted variable affects the dependent variable.

When one of the two conditions is not met, there cannot be an omitted variable bias. For example, researchers can ignore other relevant explanatory variables in an experimental study because randomization ensures their independence with the primary variable of interest — experimental treatment.

By analogy, we can collapse a three-way table over the control variable if the control variable is unrelated either to the primary explanatory variable or to the outcome variable. Unlike the case for omitted variable bias in linear regressions, the condition of unrelatedness for collapsing contingency tables refers to partial association, not unconditional association. The collapsibility property is important in analysis of multiway contingency tables, for the researcher should always try to simplify the analysis whenever possible.

For our example of the admissions data, we can see that the two conditions of collapsibility do not hold if one attempts to collapse the data over major: Sex is related to major due to sex segregation by majors (i.e., $RL \neq 0$), and the proportion being admitted varies radically across majors (i.e., $CL \neq 0$). Given these conditions, collapsing results in a marginal association (as in Table 4.19) that is different from the partial association controlling for major (as in Table 4.18). As in the case of evaluating an omitted variable bias, we can infer the direction of the difference between the marginal association and the partial association. In our admissions example, RL and CL interactions are such that women applicants are poorly represented in majors where the proportion of being admitted is high (such as major A). This combination leads to a lower proportion of women being admitted if the three-way table is collapsed over the dimension of major.

One use of the collapsibility conditions is to purge rates of confounding effects of a third variable. Assuming no three-way interactions, Clogg (1978) proposed to purge the confounding factor of L in studying the association between R (primary explanatory variable) and C (outcome variable for the calculation of rates) by adjusting the frequencies according to

$$f_{ijk}^* = \frac{f_{ijk}}{\tau_{ik}^{RL}} \tag{4.51}$$

where f^* is the adjusted frequency to be used for calculating purged rates. This adjustment by Eq. 4.51 ensures that there is no partial association between R and L for the adjusted frequencies. Thus, the conditions of collapsibility are satisfied so that the third dimension L can be ignored in the adjusted table. Xie (1989) further proposed an alternative way to purge rates of confounding factors by meeting the conditions of collapsibility in a different way:

$$f^*_{ijk} = \frac{f_{ijk}}{\tau^{CL}_{jk}} \tag{4.52}$$

(i.e., by eliminating the partial association between C and L). Clogg (1978) discussed ways to purge three-way interactions (τ^{RCL}) when present.

4.6.4. Classes of Models for Three-Way Tables

As in the case for two-way tables, the saturated model is seldom of research interest, for it simply parameterizes observed frequencies. Researchers often wish to construct more parsimonious models and test them against observed data. Let us now further simplify the saturated model of Eqs. 4.45 and 4.46 into the following special classes of models. From now on, we will use the loglinear notation of Eq. 4.46, although corresponding notation in terms of the multiplicative form of Eq. 4.45 can be easily obtained. We use a notation for models in which additive terms are separated by a comma, and interactions between variables are not separated. Unless explicitly stated, hierarchical structure of terms is maintained so that a higher-order interaction implicitly assumes the presence of lower-order interactions and marginal parameters. Thus, the saturated model of Eq. 4.46 can be denoted simply as (RCL).

Class 1. Let us first consider the "mutual independence" model, denoted as (R,C,L). The key feature of the model is that there are no interactions. Under this model, all two-way and three-way interactions are nil (i.e., $\mu^{RC} = \mu^{RL} = \mu^{CL} = \mu^{RCL} = 0$ for all i, j, and k). This model assumes that the three variables are independent of each other pairwise:

- R and C are independent,
- R and L are independent, and
- C and L are independent.

Because there is no two-way interaction, the three-way table can be collapsed in all three dimensions. That is,

- marginal association = partial association = nil, for any pair of variables.

If the model holds true, it calls for a *univariate* analysis.

Class 2. Let us now consider the "joint independence" model, denoted as (R, CL), (RC, L), or (RL, C). This class of models allows only one two-way interaction. Hence, two sets of two-way interactions and the three-way interactions are nil. Let us use model (R, CL) as an illustration. In this case, $\mu^{RC} = \mu^{RL} = \mu^{RCL} = 0$ for all i, j, and k, so that R is independent with respect to the other two variables (C and L):

- R and L are independent and
- R and C are independent.

The three-way table is collapsible in all three dimensions. We have,

- marginal association = partial association, for any pair of variables. In addition,
- marginal RL and RC association = partial RL and RC = nil.

Class 3. Let us further consider the model of "conditional independence," denoted as (RL, CL), (RC, CL), or (RC, RL). This class of models contains two two-way interactions. With (RL, CL) as our example, the conditional independence model means that R and C are independent of each other at each level of L:

- R and C are independent, given L.

The table is collapsible along R and C, but not along L. In other words,

- marginal RL and CL association = partial RL and CL association, but
- marginal RC association $\neq$ partial RC association (= nil).

This is an important model. It means that the marginal association (RC) may be spurious if one ignores a relevant variable (L), similar to an omitted-variable bias in linear regressions. As will be shown later, the graduate admission example fits this model rather well.

Class 4. Finally, let us consider the "no three-way interaction" model (RC, RL, CL). This model allows for all three two-way interactions. It does not imply conditional independence. No three-way interaction implies homogeneous associations: partial two-way association does not vary with the third variable.

The table is not collapsible in any direction. That is,

- marginal two-way association $\neq$ partial two-way association, for any pair of variables.

Finally, when the (RC, RL, CL) model does not fit the data, partial two-way associations (RC, RL, and CL) vary as a function of the third variable. This property is called heterogeneous association, which requires modeling of three-way interactions.

We now apply various models to the graduate admissions data earlier presented in Table 4.18. The data are of the dimension $2 \times 2 \times 6$ (for $R \times C \times L$), where R is sex, C admission outcome, and L major. Summary measures of fit for the various models are provided in Table 4.20. Model 1 is the mutual independence model, which does not fit the data ($G^2 = 2092.69$, for 16 degrees of freedom; BIC = 1958.01) but is presented here as a baseline for other models. Model 2 is a joint independence model allowing for the interaction between sex and major. In allowing for the interaction between sex and major, we bracket out the sex segregation of major as a pre-existing condition prior to the admission process. With $G^2 = 872.08$ for 11 degrees of freedom and BIC = 779.48, model 2 significantly improves upon model 1 in goodness-of-fit. In model 3, we further allow for the interaction between major and admission outcome and in effect specify conditional independence. There is no net association between sex and admission outcome conditional on major. Model 3 fits

Table 4.20: Goodness-of-fit statistics of models for admission data.

Model	Parameter terms[a]	G^2	*df*	BIC	Δ
(1)	(R, C, L)	2092.69	16	1958.01	25.98
(2)	(RL, C)	872.08	11	779.48	16.85
(3)	(RL, CL)	21.13	6	−29.37	1.66
(4)	$(RL, CL,$ dummy$)$	2.81	5	−39.28	0.81

[a] R = sex, C = admission outcome, and L = major. Dummy refers to the cell where R = female, C = admitted, and L = major A. Δ is the index of dissimilarity between observed and predicted frequencies (in %).

Table 4.21: Estimates of interaction parameters of model 4.

Class of parameters	Parameter	Estimate	(S.E.)
Gender by major	Female × major A	—	—
	Female × major B	−0.329	(0.311)
	Female × major C	3.382	(0.244)
	Female × major D	2.674	(0.244)
	Female × major E	3.502	(0.250)
	Female × major F	2.691	(0.245)
Admission by major	Admitted × major A	—	—
	Admitted × major B	0.052	(0.112)
	Admitted × major C	−1.106	(0.100)
	Admitted × major D	−1.155	(0.104)
	Admitted × major E	−1.572	(0.119)
	Admitted × major F	−3.159	(0.168)
Dummy	Female, major A, admitted	1.027	(0.261)

the data reasonably well ($G^2 = 21.13$ for 6 degrees of freedom, BIC = −29.37). Based on the BIC statistic, we may conclude that the data support the conditional independence hypothesis.

However, we observed earlier that female applicants seem to have an advantage in major A. Testing this specific three-way interaction, we add to model 4 a dummy variable denoting the cell where R = female, C = admitted, and L = major A. As the goodness-of-fit statistics show, model 4 fits the data extremely well ($G^2 = 2.81$ for 5 degrees of freedom, BIC = −39.28). The final model means that conditional independence holds true for all majors except for major A, where there is a sex difference in admission rates. As shown in the estimated parameters presented in Table 4.21, the sex difference for major A is in favor of women, contrary to the criticism that the admission process at the University of California-Berkeley favors male applicants. From the parameter estimates, we also observe clearly that women

applicants are underrepresented in majors A and B and overrepresented in majors C through F, and that the admission rate is higher in majors A and B than in other majors and is particularly low in major F.

4.6.5. *Analysis of Variation in Association*

The preceding subsection considered some common models for three-way tables. These models are primarily used to test the presence or absence of partial associations. From these tests, we are able to say whether a three-way table is collapsible over a dimension. We did not consider complicated cases beyond the no three-way interaction model.

What should we do if the no three-way interaction model does not fit the data? Fitting the saturated model is usually not a satisfactory answer, since the saturated model is not parsimonious. For the admissions example, we were able to identify a local three-way interaction by carefully examining the table.

We now consider a general situation where the research interest centres on the variation of a two-way association over a third dimension (or more generally a combination of other dimensions). Examples of this kind are plentiful in social science. For example, researchers studying comparative social mobility may be interested in whether the association between father's occupation and son's occupation varies systematically as a function of a nation's characteristics (Grusky & Hauser, 1984). Family sociologists may be interested in whether educational assortative mating has strengthened over time (Mare, 1991).

In this section, we recommend a "conditional" approach generalized from loglinear models for two-way tables for the analysis of variation in association. There are two advantages to this conditional approach. First, the researcher often can achieve parsimony. Second, parameters from such an approach are relatively easy to interpret.

Let us illustrate the approach with a three-way table of $R \times C \times L$ where the primary research interest lies in the analysis of R–C association over the dimension of L. Under the saturated model, the expected frequency is given in Eq. 4.45. Given the objective of analysis of variation in association over a third dimension, the researcher often wishes to begin with the model of conditional independence (RL,CL); that is,

$$F_{ijk} = \tau\tau_i^R\tau_j^C\tau_k^L\tau_{ik}^{RL}\tau_{jk}^{CL} \tag{4.53}$$

implying that there is no association between R and C given L. Compared to the saturated model of Eq. 4.45, we see that the baseline two-way interaction (τ^{RC}) and the three-way interaction (τ^{RCL}) are omitted in Eq. 4.53. The researcher typically focuses on the specification of τ^{RC} and τ^{RCL}. This is what we mean by a conditional approach, for the analysis of variation in association is now conditioned on Eq. 4.53. This point is made even clearer if we write out the local odds-ratio conditional on $L = k$ from the saturated model Eq. 4.45:

$$\theta_{ij|k} = \frac{F_{ijk}F_{(i+1)(j+1)k}}{F_{i(j+1)k}F_{(i+1)jk}} = \frac{\tau_{ij}^{RC}\tau_{(i+1)(j+1)}^{RC}}{\tau_{i(j+1)}^{RC}\tau_{(i+1)j}^{RC}}\frac{\tau_{ijk}^{RCL}\tau_{(i+1)(j+1)k}^{RCL}}{\tau_{i(j+1)k}^{RCL}\tau_{(i+1)jk}^{RCL}} \quad (4.54)$$

That is, the conditional odds-ratio depends only on the two-way interaction τ^{RC} parameters and the three-way interaction τ^{RCL} parameters.

There are many ways to parameterize τ^{RC} and τ^{RCL}. See Goodman (1986), Xie (1992), and Goodman and Hout (1998) for more thorough treatments of the subject. Note that both τ^{RC} and τ^{RCL} contain an R–C two-way association. The most common specification for τ^{RCL} is to interact a two-way R–C association pattern with layer. Let us say that τ^{RC} is modeled to follow a baseline association ω^{RC}, and τ^{RCL} modeled by interacting a two-way cross-layer "deviation" association (ψ^{RC}) and layer (L). ω^{RC} and ψ^{RC} can be the same. With this notation, we will provide some general guidelines and illustrate them with a concrete example.

Recommendation 1. It is desirable to have a simple model to reduce ω and ψ to just ψ. This is tantamount to specifying that the two-way R–C association has the same pattern across layers. When this is the case, we can set τ^{RC} to 1 and τ^{RCL} as interaction between ψ with L. This strategy requires that ψ be a more parsimonious specification than the full-interaction of R and C (i.e., consuming less than $(I-1)(J-1)$ degrees of freedom). Otherwise, the resulting model is saturated. This strategy works because we only specify the same association function for the basic two-way association but allow the parameters for the function to differ across layers. For example, the researcher may specify a general RC association model at each layer and estimate the different RC parameters at different layers (Becker & Clogg, 1989; Clogg, 1982).

Recommendation 2. If it is necessary to give different specifications to the baseline association ω and the deviation association ψ, it is desirable to have a more parsimonious specification for ψ than for ω. This is intuitive because the number of parameters for the RCL three-way interaction (which is the interaction between ψ and L) multiplies quickly as the complexity of ψ increases. Researchers sometimes give a saturated model to ω in order to achieve a better fit. In Mare's (1991) study of trends in educational homogamy, for example, ω is the full interaction, but ψ is a crossings models with only four parameters. Interacting the four crossings parameters with time, Mare was able to show the trends in the strength of educational homogamy.

Recommendation 3. It is desirable to specify the log-multiplicative-layer specification between ψ and L rather than the simple interaction specification. This is particularly powerful if recommendation 1 is taken so that ψ and ω are the same. By the log-multiplicative-layer specification, we mean the following model (Xie, 1992) (with $\omega = 1$):

$$F_{ijk} = \tau\tau_i^R\tau_j^C\tau_k^L\tau_{ik}^{RL}\tau_{jk}^{CL}\exp(\psi_{ij}\phi_k) \quad (4.55)$$

where the ψ parameters describe the R–C two-way deviation association, and ϕ's indicate the layer-specific deviations in the association. With this specification, the conditional local LOR (from Eq. 4.54) is simplified to

$$\log(\theta_{ij|k}) = (\psi_{ij} + \psi_{(i+1)(j+1)} - \psi_{i(j+1)} - \psi_{(i+1)j})\phi_k$$
$$= \phi_k \log(\theta_{ij})$$

where θ_{ij} is a function of the ψ parameters and can be thought of as the baseline odds-ratio. This log-multiplicative-layer model is parsimonious, for it only adds $(K-1)$ degrees of freedom to test for three-way interactions, yielding a 1-degree-of-freedom test for each additional layer. In addition, at each layer, the R–C association follows the same pattern but with different strengths. See Goodman and Hout (1998) for a discussion of situations where $\omega \neq 1$.

We now provide an example where these three recommendations are put to practice. The example was drawn from a study of class mobility in three nations (England, France, and Sweden) conducted by Erikson, Goldthorpe, and Portocarero (1979). There are seven categories for both father's class and son's class, giving rise to a $7 \times 7 \times 3$ table. The same data were also analyzed by Hauser (1984) and Xie (1992). In Table 4.22, we present a series of models for these data. Both the data and the estimated models are available from this book's website.

In the first line of Table 4.22, we present the conditional independence model, which fits the data poorly ($G^2 = 4860.03$ for 108 degrees of freedom), as the null model. The second line is the homogeneous full-interaction model (FI_0), which is the same as the no three-way interaction model. Although model FI_0 fits the data reasonably well by the BIC statistic criterion, its heterogeneous form (simply interaction with layer) is the saturated model (FI_l). In the log-multiplicative-layer model with full-interaction as ψ (FI_x), we use 2 more degrees of freedom to test for systematic variation in the full-interaction across layer, providing a superior goodness-of-fit, by both the reduction in G^2 (29.16 for 2 degrees of freedom) and

Table 4.22: Models for three-nation class mobility data.

Model	Features	G^2	*df*	BIC
CA	Conditional independence	4860.03	108	3812.57
FI_0	Homogeneous full-interaction	121.30	72	−577.01
FI_l	Heterogeneous full-interaction	0	0	0
FI_x	Log-multiplicative full-interaction	92.14	70	−586.77
H_0	Homogeneous levels model	244.34	103	−754.63
H_l	Heterogeneous levels model	208.50	93	−693.48
H_x	Log-multiplicative levels model	216.37	101	−763.20
RCQ_0	Homogeneous quasi-RC	337.86	76	−399.24
RCQ_l	Heterogeneous quasi-RC	271.97	54	−251.76
RCQ_x	Log-multiplicative quasi-RC	332.37	74	−385.34

BIC (-586.77). The three H models are based on a topological pattern (with six levels).[5] With this parsimonious baseline specification for the origin-destination association, the homogeneous levels model provides a reasonable fit to the data (notably BIC $= -754.63$) for 5 additional degrees of freedom beyond the CA model. Since the levels model is parsimonious, we can adopt the first recommendation and interact the levels matrix and layer, resulting in the heterogeneous levels model, H_l. Given the large sample size (16,297), Erikson et al. (1979) and Hauser (1984) choose to prefer model H_0 over model H_l, even though, strictly speaking, the χ^2 statistic between the two nested models (35.84 for 10 degrees of freedom) is significant. According to BIC, model H_0 has a lower negative value and thus fits the data better than model H_l. With the specification that the origin-destination association varies log-multiplicatively cross-nationally, model H_x is between models H_0 and H_l. By the log-likelihood-ratio χ^2 statistic, model H_x (G^2 of 216.37 with 101 degrees of freedom) fits the data significantly better than model H_0 ($\Delta G^2 = 27.97$ for 2 degrees of freedom) and not significantly worse than model H_l ($\Delta G^2 = 7.87$ with 8 degrees of freedom). Additionally, of all the models in Table 4.22, model H_x has the lowest BIC value and thus fits the data the best according to the BIC criterion.

In the last three lines of Table 4.22, we change the two-way association specification (ψ) from a levels model to an RC association model while blocking diagonals, called quasi-RC and denoted as RCQ. In this specification, the $\tau^{RC}\tau^{RCL}$ part of Eq. 4.45 becomes

$$\exp(\beta_k \phi_{ik} \varphi_{jk}) \quad \text{for } i \neq j$$

Model RCQ_0 constrains all three parameters (β, ϕ, and φ) to be invariant across L, while RCQ_l allows them to vary freely with L. The log-multiplicative-layer version (RCQ_x) is intermediate in fixing the scores (ϕ and φ) but allowing the strength parameter β to vary by level. The results in Table 4.22 indicate that model RCQ_0 fits the data better than the other two versions of the RCQ model according to the BIC statistic.

One advantage of the log-multiplicative-layer model is that a simple parameter measuring the strength of association can be obtained for each table, subject to a global normalization. Xie (1992) capitalizes on this property to separate *levels* of mobility from *patterns* of mobility, the latter of which Xie assumes to be the same in all modern societies according to his revised interpretation of a classic hypothesis in the social mobility literature. We report these parameters for the three log-multiplicative models in Table 4.23. From these estimates, we can conclude a similar pattern regardless of model specification: the strength of the association between father's class and son's class is weaker in Sweden than in England and France but similar between England and France. This example shows that fine-tuning of two-way specifications (both baseline and deviation) sometimes has little consequence for the main research objective: detection of the variation in a two-way association along a third dimension.

5. The levels matrix is available from the book's website.

Table 4.23: Nation-specific ϕ parameters.

Model	**ϕ_1 (England)**	**ϕ_2 (France)**	**ϕ_3 (Sweden)**
FI_x	0.617	0.633	0.468
H_x	0.613	0.634	0.472
RCQ_x	0.652	0.575	0.495

4.6.6. *Model Selection*

In Table 4.22 we have selected models based on two criteria: the change in G^2 for nested models and BIC for nested as well as unnested models. We recommend using these and other goodness-of-fit criteria (including Pearson χ^2 statistic and the index of dissimilarity) in the course of model fitting.

The likelihood-ratio χ^2 test in terms of the change in G^2 is the most common method for selecting among competing nested models. The likelihood-ratio test has the advantage of having a familiar proportionate reduction in error interpretation, much like the F-test in OLS regression models and of being applicable to any model estimated with ML. The index of dissimilarity provides a descriptive measure, which is useful in assessing how well a model is able to reproduce the observed frequencies. The BIC statistic helps the researcher trade parsimony for goodness-of-fit in large samples, for which even a "good" model might be rejected by the G^2 statistic.