Chapter 5

Log-Linear Models for Contingency Tables

In this chapter we study the application of Poisson regression models to the analysis of contingency tables. This is perhaps one of the most popular applications of log-linear models, and is based on the existence of a very close relationship between the multinomial and Poisson distributions.

5.1 Models for Two-dimensional Tables

We start by considering the simplest possible contingency table: a two-by-two table. However, the concepts to be introduced apply equally well to more general two-way tables where we study the joint distribution of two categorical variables.

5.1.1 The Heart Disease Data

Table 5.1 was taken from the Framingham longitudinal study of coronary heart disease (Cornfield, 1962; see also Fienberg, 1977). It shows 1329 patients cross-classified by the level or their serum cholesterol (below or above 260) and the presence or absence of heart disease.

There are various sampling schemes that could have led to these data, with consequences for the probability model one would use, the types of questions one would ask, and the analytic techniques that would be employed. Yet, all schemes lead to equivalent analyses. We now explore several approaches to the analysis of these data.

TABLE 5.1: Serum Cholesterol and Heart Disease

Serum Cholesterol	Heart Disease		Total
	Present	Absent	
< 260	51	992	1043
260+	41	245	286
Total	92	1237	1329

5.1.2 The Multinomial Model

Our first approach will assume that the data were collected by sampling 1329 patients who were then classified according to cholesterol and heart disease. We view these variables as two responses, and we are interested in their joint distribution. In this approach the total sample size is assumed fixed, and all other quantities are considered random.

We will develop the random structure of the data in terms of the row and column variables, and then note what this implies for the counts themselves. Let C denote serum cholesterol and D denote heart disease, both discrete factors with two levels. More generally, we can imagine a row factor with I levels indexed by i and a column factor with J levels indexed by j, forming an $I \times J$ table. In our example $I = J = 2$.

To describe the joint distribution of these two variables we let π_{ij} denote the probability that an observation falls in row i and column j of the table. In our example words, π_{ij} is the probability that serum cholesterol C takes the value i and heart disease D takes the value j. In symbols,

$$\pi_{ij} = \Pr\{C = i, D = j\}, \tag{5.1}$$

for $i = 1, 2, \ldots, I$ and $j = 1, 2, \ldots, J$. These probabilities completely describe the *joint* distribution of the two variables.

We can also consider the *marginal* distribution of each variable. Let $\pi_{i.}$ denote the probability that the row variable takes the value i, and let $\pi_{.j}$ denote the probability that the column variable takes the value j. In our example $\pi_{i.}$ and $\pi_{.j}$ represent the marginal distributions of serum cholesterol and heart disease. In symbols,

$$\pi_{i.} = \Pr\{C = i\} \qquad \text{and} \qquad \pi_{.j} = \Pr\{D = j\}. \tag{5.2}$$

Note that we use a dot as a placeholder for the omitted subscript.

The main hypothesis of interest with two responses is whether they are *independent*. By definition, two variables are independent if (and only if)

their joint distribution is the product of the marginals. Thus, we can write the hypothesis of independence as

$$H_0 : \pi_{ij} = \pi_{i.}\pi_{.j} \tag{5.3}$$

for all $i = 1, \ldots, I$ and $j = 1, \ldots, J$. The question now is how to estimate the parameters and how to test the hypothesis of independence.

The traditional approach to testing this hypothesis calculates expected counts under independence and compares observed and expected counts using Pearson's chi-squared statistic. We adopt a more formal approach that relies on maximum likelihood estimation and likelihood ratio tests. In order to implement this approach we consider the distribution of the counts in the table.

Suppose each of n observations is classified independently in one of the IJ cells in the table, and suppose the probability that an observation falls in the (i, j)-th cell is π_{ij}. Let Y_{ij} denote a random variable representing the number of observations in row i and column j of the table, and let y_{ij} denote its observed value. The joint distribution of the counts is then the *multinomial* distribution, with

$$\Pr\{\mathbf{Y} = \mathbf{y}\} = \frac{n!}{y_{11}!y_{12}!y_{21}!y_{22}!}\pi_{11}^{y_{11}}\pi_{12}^{y_{12}}\pi_{21}^{y_{21}}\pi_{22}^{y_{22}}, \tag{5.4}$$

where $\mathbf{Y}$ is a random vector collecting all four counts and $\mathbf{y}$ is a vector of observed values. The term to the right of the fraction represents the probability of obtaining y_{11} observations in cell (1,1), y_{12} in cell (1,2), and so on. The fraction itself is a combinatorial term representing the number of ways of obtaining y_{11} observations in cell (1,1), y_{12} in cell (1,2), and so on, out of a total of n. The multinomial distribution is a direct extension of the binomial distribution to more than two response categories. In the present case we have four categories, which happen to represent a two-by-two structure. In the special case of only two categories the multinomial distribution reduces to the familiar binomial.

Taking logs and ignoring the combinatorial term, which does not depend on the parameters, we obtain the multinomial log-likelihood function, which for a general $I \times J$ table has the form

$$\log L = \sum_{i=1}^{I}\sum_{j=1}^{J} y_{ij}\log(\pi_{ij}). \tag{5.5}$$

To estimate the parameters we need to take derivatives of the log-likelihood function with respect to the probabilities, but in doing so we must take into

account the fact that the probabilities add up to one over the entire table. This restriction may be imposed by adding a Lagrange multiplier, or more simply by writing the last probability as the complement of all others. In either case, we find the unrestricted maximum likelihood estimate to be the sample proportion:

$$\hat{\pi}_{ij} = \frac{y_{ij}}{n}.$$

Substituting these estimates into the log-likelihood function gives its unrestricted maximum.

Under the hypothesis of independence in Equation 5.3, the joint probabilities depend on the margins. Taking derivatives with respect to $\pi_{i.}$ and $\pi_{.j}$, and noting that these are also constrained to add up to one over the rows and columns, respectively, we find the m.l.e.'s

$$\hat{\pi}_{i.} = \frac{y_{i.}}{n} \qquad \text{and} \qquad \hat{\pi}_{.j} = \frac{y_{.j}}{n},$$

where $y_{i.} = \sum_j y_{ij}$ denotes the row totals and $y_{.j}$ denotes the column totals. Combining these estimates and multiplying by n to obtain expected counts gives

$$\hat{\mu}_{ij} = \frac{y_{i.}y_{.j}}{n},$$

which is the familiar result from introductory statistics. In our example, the expected frequencies are

$$\hat{\mu}_{ij} = \begin{pmatrix} 72.2 & 970.8 \\ 19.8 & 266.2 \end{pmatrix}.$$

Substituting these estimates into the log-likelihood function gives its maximum under the restrictions implied by the hypothesis of independence. To test this hypothesis, we calculate twice the difference between the unrestricted and restricted maxima of the log-likelihood function, to obtain the deviance or likelihood ratio test statistic

$$D = 2\sum_i \sum_j y_{ij} \log(\frac{y_{ij}}{\hat{\mu}_{ij}}). \tag{5.6}$$

Note that the numerator and denominator inside the log can be written in terms of estimated probabilities or counts, because the sample size n cancels out. Under the hypothesis of independence, this statistic has approximately in large samples a chi-squared distribution with $(I-1)(J-1)$ d.f.

Going through these calculations for our example we obtain a deviance of 26.43 with one d.f. Comparison of observed and fitted counts in terms of

Pearson's chi-squared statistic gives 31.08 with one d.f. Clearly, we reject the hypothesis of independence, concluding that heart disease and serum cholesterol level are associated.

5.1.3 The Poisson Model

An alternative model for the data in Table 5.1 is to treat the four counts as realizations of independent Poisson random variables. A possible physical model is to imagine that there are four groups of people, one for each cell in the table, and that members from each group arrive randomly at a hospital or medical center over a period of time, say for a health check. In this model the total sample size is not fixed in advance, and all counts are therefore random.

Under the assumption that the observations are independent, the joint distribution of the four counts is a product of Poisson distributions

$$\Pr\{\mathbf{Y}=\mathbf{y}\} = \prod_i \prod_j \frac{\mu_{ij}^{y_{ij}} e^{-\mu_{ij}}}{y_{ij}!}. \tag{5.7}$$

Taking logs we obtain the usual Poisson log-likelihood from Chapter 4.

In terms of the systematic structure of the model, we could consider three log-linear models for the expected counts: the null model, the additive model and the saturated model. The null model would assume that all four kinds of patients arrive at the hospital or health center in the same numbers. The additive model would postulate that the arrival rates depend on the level of cholesterol and the presence or absence of heart disease, but not on the combination of the two. The saturated model would say that each group has its own rate or expected number of arrivals.

At this point you may try fitting the Poisson additive model to the four counts in Table 5.1, treating cholesterol and heart disease as factors or discrete predictors. You will discover that the deviance is 26.43 on one d.f. (four observations minus three parameters, the constant and the coefficients of two dummies representing cholesterol and heart disease). If you print the fitted values you will discover that they are exactly the same as in the previous subsection.

This result, of course, is not a coincidence. *Testing the hypothesis of independence in the multinomial model is exactly equivalent to testing the goodness of fit of the Poisson additive model.* A rigorous proof of this result is beyond the scope of these notes, but we can provide enough information to show that the result is intuitively reasonable and to understand when it can be used.

First, note that if the four counts have independent Poisson distributions, their sum is distributed Poisson with mean equal to the sum of the means. In symbols, if $Y_{ij} \sim P(\mu_{ij})$ then the total $Y_{\cdot\cdot} = \sum_i \sum_j Y_{ij}$ is distributed Poisson with mean $\mu_{\cdot\cdot} = \sum_i \sum_j \mu_{ij}$. Further, the conditional distribution of the four counts *given* their total is multinomial with probabilities

$$\pi_{ij} = \mu_{ij}/n,$$

where we have used n for the observed total $y_{\cdot\cdot} = \sum_{i,j} y_{ij}$. This result follows directly from the fact that the conditional distribution of the counts $\mathbf{Y}$ given their total $Y_{\cdot\cdot}$ can be obtained as the ratio of the joint distribution of the counts and the total (which is the same as the joint distribution of the counts, which imply the total) to the marginal distribution of the total. Dividing the joint distribution given in Equation 5.7 by the marginal, which is Poisson with mean $\mu_{\cdot\cdot}$, leads directly to the multinomial distribution in Equation 5.4.

Second, note that the systematic structure of the two models is the same. In the model of independence the joint probability is the product of the marginals, so taking logs we obtain

$$\log \pi_{ij} = \log \pi_{i\cdot} + \log \pi_{\cdot j},$$

which is exactly the structure of the additive Poisson model

$$\log \mu_{ij} = \eta + \alpha_i + \beta_j.$$

In both cases the log of the expected count depends on the row and the column but not the combination of the two. In fact, it is only the constraints that differ between the two models. The multinomial model restricts the joint and marginal probabilities to add to one. The Poisson model uses the reference cell method and sets $\alpha_1 = \beta_1 = 0$.

If the systematic and random structure of the two models are the same, then it should come as no surprise that they produce the same fitted values and lead to the same tests of hypotheses. There is only one aspect that we glossed over: the equivalence of the two distributions holds conditional on n, but in the Poisson analysis the total n is random and we have not conditioned on its value. Recall, however, that the Poisson model, by including the constant, reproduces exactly the sample total. It turns out that we don't need to condition on n because the model reproduces its exact value anyway.

The morale of this long-winded story is that we do not need to bother with multinomial models and can always resort to the equivalent Poisson

model. While the gain is trivial in the case of a two-by-two table, it can be very significant as we move to cross-classifications involving three or more variables, particularly as we don't have to worry about maximizing the multinomial likelihood under constraints. The only trick we need to learn is how to translate the questions of independence that arise in the multinomial context into the corresponding log-linear models in the Poisson context.

5.1.4 The Product Binomial*

(On first reading you may wish to skip this subsection and the next and proceed directly to the discussion of three-dimensional tables in Section 5.2.)

There is a third sampling scheme that may lead to data such as Table 5.1. Suppose that a decision had been made to draw a sample of 1043 patients with low serum cholesterol and an independent sample of 286 patients with high serum cholesterol, and then examine the presence or absence of heart disease in each group.

Interest would then focus on the *conditional* distribution of heart disease given serum cholesterol level. Let π_i denote the probability of heart disease at level i of serum cholesterol. In the notation of the previous subsections,

$$\pi_i = \Pr\{D = 1|C = i\} = \frac{\pi_{i1}}{\pi_{i\cdot}},$$

where we have used the fact that the conditional probability of falling in column one given that you are in row i is the ratio of the joint probability π_{i1} of being in cell (i,1) to the marginal probability $\pi_{i\cdot}$ of being in row i.

Under this scheme the row margin would be fixed in advance, so we would have n_1 observations with low cholesterol and n_2 with high. The number of cases with heart disease in category y of cholesterol, denoted Y_{i1}, would then have a binomial distribution with parameters π_i and n_i independently for $i = 1, 2$. The likelihood function would then be a product of two binomials:

$$\Pr\{\mathbf{Y} = \mathbf{y}\} = \frac{n_1!}{y_{11}!y_{12}!}\pi_1^{y_{11}}(1-\pi_1)^{y_{12}}\frac{n_2!}{y_{21}!y_{22}!}\pi_2^{y_{21}}(1-\pi_2)^{y_{22}}, \qquad (5.8)$$

where we have retained double subscripts and written y_{i1} and y_{i2} instead of the more familiar y_i and $n_i - y_i$ to facilitate comparison with Equations 5.4 and 5.7.

The main hypothesis of interest would be the hypothesis of *homogeneity*, where the probability of heart disease is the same in the two groups:

$$H_o : \pi_1 = \pi_2.$$

To test this hypothesis you might consider fitting logistic regression models to the data, treating heart disease as the response and serum cholesterol as the predictor, and working with two observations representing the two groups. If you try this, you will discover that the deviance for the null model, which can be interpreted as a likelihood ratio test of the hypothesis of homogeneity, is 26.43 with one d.f., and coincides with the multinomial and Poisson deviances of the previous two subsections.

Again, this is no coincidence, because the random and systematic components of the models are equivalent. The product binomial distribution in Equation 5.8 can be obtained starting from the assumption that the four counts Y_{ij} are independent Poisson with means μ_{ij}, and then conditioning on the totals $Y_{i.} = \sum_j Y_{ij}$, which are Poisson with means $\mu_{i.} = \sum_j \mu_{ij}$, for $i = 1, 2$. Taking the ratio of the joint distribution of the counts to the marginal distribution of the two totals leads to the product binomial in Equation 5.8 with $\pi_i = \mu_{i1}/\mu_{i.}$.

Similarly, the hypothesis of homogeneity turns out to be equivalent to the hypothesis of independence and hence the additive log-linear model. To see this point note that if two variables are independent, then the conditional distribution of one given the other is the same as its marginal distribution. In symbols, if $\pi_{ij} = \pi_{i.}\pi_{.j}$ then the conditional probability, which in general is $\pi_{j|i} = \pi_{ij}/\pi i.$, simplifies to $\pi_{j|i} = \pi_{.j}$, which does not depend on i. In terms of our example, under independence or homogeneity the conditional probability of heart disease is the same for the two cholesterol groups.

Again, note that the binomial and Poisson models are equivalent conditioning on the row margin, but in fitting the additive log-linear model we did not impose any conditions. Recall, however, that the Poisson model, by treating serum cholesterol as a factor, reproduces exactly the row margin of the table. Thus, it does not matter that we do not condition on the margin because the model reproduces its exact value anyway.

The importance of this result is that the results of our analyses are in fact independent of the sampling scheme.

- If the row margin is fixed in advance we can treat the row factor as a predictor and the column factor as a response and fit a model of homogeneity using the product binomial likelihood.

- If the total is fixed in advance we can treat both the row and column factors as responses and test the hypothesis of independence using the multinomial likelihood.

- Or we can treat all counts as random and fit an additive log-linear

model using the Poisson likelihood.

Reassuringly, the results will be identical in all three cases, both in terms of fitted counts and in terms of the likelihood ratio statistic.

Note that if the total is fixed and the sampling scheme is multinomial we can always condition on a margin and use binomial models, the choice being up to the investigator. This choice will usually depend on whether one wishes to treat the two variables symmetrically, assuming they are both responses and studying their correlation, or asymmetrically, treating one as a predictor and the other as a response in a regression framework.

If the row margin is fixed and the sampling scheme is binomial then we must use the product binomial model, because we can not estimate the joint distribution of the two variables without further information.

5.1.5 The Hypergeometric Distribution*

There is a fourth distribution that could apply to the data in Table 5.1, namely the hypergeometric distribution. This distribution arises from treating both the row and column margins as fixed. I find it hard to imagine a sampling scheme that would lead to fixed margins, but one could use the following conditioning argument.

Suppose that the central purpose of the enquiry is the possible association between cholesterol and heart disease, as measured, for example, by the odds ratio. Clearly, the total sample size has no information about the odds ratio, so it would make sense to condition on it. Perhaps less obviously, the row and column margins carry very little information about the association between cholesterol and heart disease as measured by the odds ratio. It can therefore be argued that it makes good statistical sense to condition on both margins.

If we start from the assumption that the four counts are independent Poisson with means μ_{ij}, and then condition on the margins $Y_{i\cdot}$ and $Y_{\cdot j}$ as well as the total $Y_{\cdot\cdot}$ (being careful to use $Y_{1\cdot}, Y_{\cdot 1}$ and $Y_{\cdot\cdot}$ to maintain independence) we obtain the hypergeometric distribution, where

$$\Pr\{\mathbf{Y}=\mathbf{y}\} = \frac{y_{\cdot 1}!}{y_{11}!y_{21}!}\frac{y_{\cdot 2}!}{y_{21}!y_{22}!}\Big/\frac{n!}{y_{1\cdot}!y_{2\cdot}!}.$$

In small samples this distribution is the basis of the so-called Fisher's exact test for the two-by-two table. McCullagh and Nelder (1989, Sections 7.3–7.4) discuss a conditional likelihood ratio test that differs from the unconditional one. The question of whether one should use conditional or unconditional tests is still a matter of controversy, see for example Yates (1934, 1984). We will not consider the hypergeometric distribution further.

5.2 Models for Three-Dimensional Tables

We now consider in more detail linear models for three-way contingency tables, focusing on testing various forms of complete and partial independence using the equivalent Poisson models.

5.2.1 Educational Aspirations in Wisconsin

Table 5.2 classifies 4991 Wisconsin male high school seniors according to socio-economic status (low, lower middle, upper middle, and high), the degree of parental encouragement they receive (low and high) and whether or not they have plans to attend college (no, yes). This is part of a larger table found in Fienberg (1977, p. 101).

TABLE 5.2: Socio-economic Status, Parental Encouragement and Educational Aspirations of High School Seniors

Social Stratum	Parental Encouragement	College Plans		Total
		No	Yes	
Lower	Low	749	35	784
	High	233	133	366
Lower Middle	Low	627	38	665
	High	330	303	633
Upper Middle	Low	420	37	457
	High	374	467	841
Higher	Low	153	26	179
	High	266	800	1066
Total		3152	1938	4991

In our analysis of these data we will view all three variables as responses, and we will study the extent to which they are associated. In this process we will test various hypotheses of complete and partial independence.

Let us first introduce some notation. We will use three subscripts to identify the cells in an $I \times J \times K$ table, with i indexing the I rows, j indexing the J columns and k indexing the K layers. In our example $I = 4, J = 2$, and $K = 2$ for a total of 16 cells.

Let π_{ijk} denote the probability that an observation falls in cell (i, j, k). In our example, this cell represents category i of socio-economic status (S), category j of parental encouragement (E) and category k of college plans (P). These probabilities define the joint distribution of the three variables.

We also let y_{ijk} denote the observed count in cell (i, j, k), which we treat as a realization of a random variable Y_{ijk} having a multinomial or Poisson distribution.

We will also use the dot convention to indicate summing over a subscript, so $\pi_{i..}$ is the marginal probability that an observation falls in row i and $y_{i..}$ is the number of observations in row i. The notation extends to two dimensions, so $\pi_{ij.}$ is the marginal probability that an observation falls in row i and column$j$ and $y_{ij.}$ is the corresponding count.

5.2.2 Deviances for Poisson Models

In practice we will treat the Y_{ijk} as independent Poisson random variables with means $\mu_{ijk} = n\pi_{ijk}$, and we will fit log-linear models to the expected counts.

Table 5.3 lists all possible models of interest in the Poisson context that include all three variables, starting with the three-factor additive model $S + E + P$ on status, encouragement and plans, and moving up towards the saturated model SEP. For each model we list the abbreviated model formula, the deviance and the degrees of freedom.

TABLE 5.3: Deviances for Log-linear Models
Fitted to Educational Aspirations Data

Model	Deviance	d.f.
$S + E + P$	2714.0	10
$SE + P$	1877.4	7
$SP + E$	1920.4	7
$S + EP$	1092.0	9
$SE + SP$	1083.8	4
$SE + EP$	255.5	6
$SP + EP$	298.5	6
$SE + SP + EP$	1.575	3

We now switch to a multinomial context, where we focus on the joint distribution of the three variables S, E and P. We consider four different types of models that may be of interest in this case, and discuss their equivalence to one of the above Poisson models.

5.2.3 Complete Independence

The simplest possible model of interest in the multinomial context is the model of complete independence, where the joint distribution of the three variables is the product of the marginals. The corresponding hypothesis is

$$H_0 : \pi_{ijk} = \pi_{i..}\pi_{.j.}\pi_{..k}, \tag{5.9}$$

where $\pi_{i..}$ is the marginal probability that an observation falls in row i, and $\pi_{.j.}$ and $\pi_{..k}$ are the corresponding column and layer margins.

Under this model the logarithms of the expected cell counts are given by

$$\log \mu_{ijk} = \log n + \log \pi_{i..} + \log \pi_{.j.} + \log \pi_{..k},$$

and can be seen to depend only on quantities indexed by i, j and k but none of the combinations (such as ij, jk or ik). The notation is reminiscent of the Poisson additive model, where

$$\log \mu_{ijk} = \eta + \alpha_i + \beta_j + \gamma_k,$$

and in fact the two formulations can be shown to be equivalent, differing only on the choice of constraints: the marginal probabilities add up to one, whereas the main effects in the log-linear model satisfy the reference cell restrictions.

The m.l.e.'s of the probabilities under the model of complete independence turn out to be, as you might expect, the products of the marginal proportions. Therefore, the m.l.e.'s of the expected counts under complete independence are

$$\hat{\mu}_{ijk} = y_{i..}y_{.j.}y_{..k}/n^2.$$

Note that the estimates depend only on row, column and layer totals, as one would expect from considerations of marginal sufficiency.

To test the hypothesis of complete independence we compare the maximized multinomial log-likelihoods under the model of independence and under the saturated model. Because of the equivalence between multinomial and Poisson models, however, the resulting likelihood ratio statistic is exactly the same as the deviance for the Poisson additive model.

In our example the deviance of the additive model is 2714 with 10 d.f., and is highly significant. We therefore conclude that the hypothesis that social status, parental encouragement and college plans are completely independent is clearly untenable.

5.2.4 Block Independence

The next three log-linear models in Table 5.3 involve one of the two-factor interaction terms. As you might expect from our analysis of a two-by-two table, the presence of an interaction term indicates the existence of association between those two variables.

For example the model $SE + P$ indicates that S and E are associated, but are jointly independent of P. In terms of our example this hypothesis would state that social status and parental encouragement are associated with each other, and are jointly independent of college plans.

Under this hypothesis the joint distribution of the three variables factors into the product of two blocks, representing S and E on one hand and P on the other. Specifically, the hypothesis of block independence is

$$H_0 : \pi_{ijk} = \pi_{ij.}\pi_{..k}. \tag{5.10}$$

The m.l.e.'s of the cell probabilities turn out to be the product of the SE and P marginal probabilities and can be calculated directly. The m.l.e.'s of the expected counts under block independence are then

$$\hat{\mu}_{ijk} = y_{ij.}y_{..k}/n.$$

Note the similarity between the structure of the probabilities and that of the estimates, depending on the combination of levels of S and E on the one hand, and levels of P on the other.

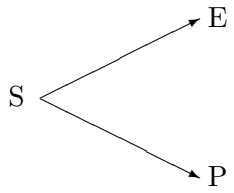
To test the hypothesis of block independence we compare the maximized multinomial log-likelihood under the restrictions imposed by Equation 5.10 with the maximized log-likelihood for the saturated model. Because of the equivalence between multinomial and Poisson models, however, the test statistic would be exactly the same as the deviance for the model $SE + P$.

In our example the deviance for the model with the SE interaction and a main effect of P is 1877.4 on 7 d.f., and is highly significant. We therefore reject the hypothesis that college plans are independent of social status and parental encouragement.

There are two other models with one interaction term. The model $SP+E$ has a deviance of 1920.4 on 7 d.f., so we reject the hypothesis that parental encouragement is independent of social status and college plans. The model $EP + S$ is the best fitting of this lot, but the deviance of 1092.0 on 9 d.f. is highly significant, so we reject the hypothesis that parental encouragement and college plans are associated but are jointly independent of social status.

5.2.5 Partial Independence

The next three log-linear models in Table 5.3 involve two of the three possible two-factor interactions, and thus correspond to cases where two pairs of categorical variables are associated. For example the log-linear model $SE + SP$ corresponds to the case where S and E are associated and so are S and P. In terms of our example we would assume that social status affects both parental encouragement and college plans. The figure below shows this model in path diagram form.



Note that we have assumed no direct link between E and P, that is, the model assumes that parental encouragement has no direct effect on college plans. In a two-way crosstabulation these two variables would appear to be associated because of their common dependency on social status S. However, *conditional* on social status S, parental encouragement E and college plans P would be independent.

Thus, the model assumes a form of partial or conditional independence, where the joint conditional distribution of EP given S is the product of the marginal conditional distributions of E given S and P given S. In symbols,

$$\Pr\{E = j, P = k|S = i\} = \Pr\{E = j|S = i\}\Pr\{P = k|S = i\}.$$

To translate this statement into unconditional probabilities we write the conditional distributions as the product of the joint and marginal distributions, so that the above equation becomes

$$\frac{\Pr\{E = j, P = k, S = i\}}{\Pr\{S = i\}} = \frac{\Pr\{E = j, S = i\}}{\Pr\{S = i\}} \frac{\Pr\{P = k, S = i\}}{\Pr\{S = i\}},$$

from which we see that

$$\Pr\{S = i, E = j, P = k\} = \frac{\Pr\{S = i, E = j\}\Pr\{S = i, P = k\}}{\Pr\{S = i\}},$$

or, in our usual notation,

$$\pi_{ijk} = \frac{\pi_{ij.}\pi_{i.k}}{\pi_{i..}}. \tag{5.11}$$

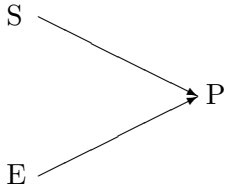
The m.l.e.'s of the expected cell counts have a similar structure and depend only on the SE and SP margins:

$$\hat{\mu}_{ijk} = \frac{y_{ij.}y_{i.k}}{y_{i..}}.$$

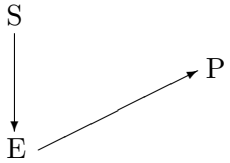
To test the hypothesis of partial independence we need to compare the multinomial log-likelihood maximized under the constraints implied by Equation 5.11 with the unconstrained maximum. Because of the equivalence between multinomial and Poisson models, however, the resulting likelihood ratio test statistic is the same as the deviance of the model $SE + SP$.

In terms of our example, the deviance of the model with SE and SP interactions is 1083.8 on 4 d.f., and is highly significant. We therefore reject the hypothesis that parental encouragement and college plans are independent within each social stratum.

There are two other models with two interaction terms. Although both of them have smaller deviances than any of the models considered so far, they still show significant lack of fit. The model $SP + EP$ has a deviance of 298.5 on 6 d.f., so we reject the hypothesis that given college plans P social status S and parental encouragement E are mutually independent. The best way to view this model in causal terms is by assuming that S and E are unrelated and both have effects on P, as shown in the path diagram below.



The model $SE + EP$ has a deviance of 255.5 on 6 d.f., and leads us to reject the hypothesis that given parental encouragement E, social class S and college plans P are independent. In causal terms one might interpret this model as postulating that social class affects parental encouragement which in turn affects college plans, with no direct effect of social class on college plans.



Note that all models consider so far have had explicit formulas for the m.l.e.'s, so no iteration has been necessary and we could have calculated all test

statistics using the multinomial likelihood directly. An interesting property of the iterative proportional fitting algorithm mentioned earlier, and which is used by software specializing in contingency tables, is that it converges in one cycle in all these cases. The same is not true of the iteratively re-weighted least squares algorithm used in Poisson regression, which will usually require a few iterations.

5.2.6 Uniform Association

The only log-linear model remaining in Table 5.3 short of the saturated model is the model involving all three two-factor interactions. In this model we have a form of association between all pairs of variables, S and E, S and P, as well as E and P. Thus, social class is associated with parental encouragement and with college plans, and in addition parental encouragement has a direct effect on college plans.

How do we interpret the lack of a three-factor interaction? To answer this question we start from what we know about interaction effects in general and adapt it to the present context, where interaction terms in models for counts represent association between the underlying classification criteria. The conclusion is that in this model the association between any two of the variables is the same at all levels of the third.

This model has no simple interpretation in terms of independence, and as a result we cannot write the structure of the joint probabilities in terms of the two-way margins. In particular

$$\pi_{ijk} \qquad \textit{is not} \qquad \frac{\pi_{ij.}\pi_{i.k}\pi_{.jk}}{\pi_{i..}\pi_{.j.}\pi_{..k}},$$

nor any other simple function of the marginal probabilities.

A consequence of this fact is that the m.l.e.'s cannot be written in closed form and must be calculated using an iterative procedure. They do, however, depend only on the three two-way margins SE, SP and EP.

In terms of our example, the model $SP + SE + EP$ has a deviance of 1.6 on three d.f., and therefore fits the data quite well. We conclude that we have no evidence against the hypothesis that all three variables are associated, but the association between any two is the same at all levels of the third. In particular, we may conclude that the association between parental encouragement E and college plans P is the same in all social strata.

To further appreciate the nature of this model, we give the fitted values in Table 5.4. Comparison of the estimated expected counts in this table with the observed counts in Table 5.2 highlights the goodness of fit of the model.

TABLE 5.4: Fitted Values for Educational Aspirations Data Based on Model of Uniform Association $SE + SP + EP$

Social Stratum	Parental Encouragement	College Plans	
		No	Yes
Lower	Low	753.1	30.9
	High	228.9	137.1
Lower Middle	Low	626.0	39.0
	High	331.0	302.0
Upper Middle	Low	420.9	36.1
	High	373.1	467.9
Higher	Low	149.0	30.0
	High	270.0	796.0

We can also use the fitted values to calculate measures of association between parental encouragement E and college plans P for each social stratum. For the lowest group, the odds of making college plans are barely one to 24.4 with low parental encouragement, but increase to one to 1.67 with high encouragement, giving an odds ratio of 14.6. If you repeat the calculation for any of the other three social classes you will find exactly the same ratio of 14.6.

We can verify that this result follows directly from the lack of a three-factor interaction in the model. The logs of the expected counts in this model are

$$\log \mu_{ijk} = \eta + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk}.$$

The log-odds of making college plans in social stratum i with parental encouragement j are obtained by calculating the difference in expected counts between $k = 2$ and $k = 1$, which is

$$\log(\mu_{ij2}/\mu_{ij1}) = \gamma_2 - \gamma_1 + (\alpha\gamma)_{i2} - (\alpha\gamma)_{i1} + (\beta\gamma)_{j2} - (\beta\gamma)_{j1},$$

because all terms involving only i, j or ij cancel out. Consider now the difference in log-odds between high and low encouragement, i.e. when $j = 2$ and $j = 1$:

$$\log(\frac{\mu_{i22}/\mu_{i21}}{\mu_{i12}/\mu_{i11}}) = (\beta\gamma)_{22} - (\beta\gamma)_{21} - (\beta\gamma)_{12} + (\beta\gamma)_{11},$$

which does not depend on i. Thus, we see that the log of the odds ratio is the same at all levels of S. Furthermore, under the reference cell restrictions

all interaction terms involving level one of any of the factors would be set to zero, so the log of the odds ratio in question is simply $(\beta\gamma)_{22}$. For the model with no three-factor interaction the estimate of this parameter is 2.683 and exponentiating this value gives 14.6.

5.2.7 Binomial Logits Revisited

Our analysis so far has treated the three classification criteria as responses, and has focused on their correlation structure. An alternative approach would treat one of the variables as a response and the other two as predictors in a regression framework. We now compare these two approaches in terms of our example on educational aspirations, treating college plans as a dichotomous response and socio-economic status and parental encouragement as discrete predictors.

To this end, we treat each of the 8 rows in Table 5.2 as a group. Let Y_{ij} denote the number of high school seniors who plan to attend college out of the n_{ij} seniors in category i of socio-economic status and category j of parental encouragement. We assume that these counts are independent and have binomial distributions with $Y_{ij} \sim B(n_{ij}, \pi_{ij})$, where π_{ij} is the probability of making college plans We can then fit logistic regression models to study how the probabilities depend on social stratum and parental encouragement.

TABLE 5.5: Deviances for Logistic Regression Models Fitted to the Educational Aspirations Data

Model	Deviance	d.f.
Null	1877.4	7
S	1083.8	4
E	255.5	6
$S + E$	1.575	3

Table 5.5 shows the results of fitting four possible logit models of interest, ranging from the null model to the additive model on socioeconomic status (S) and parental encouragement (E). It is clear from these results that both social class and encouragement have significant gross and net effects on the probability of making college plans. The best fitting model is the two-factor additive model, with a deviance of 1.6 on three d.f. Table 5.6 shows parameter estimates for the additive model.

Exponentiating the estimates we see that the odds of making college plans increase five-fold as we move from low to high socio-economic status.

TABLE 5.6: Parameter Estimates for Additive Logit Model Fitted to the Educational Aspirations Data

Variable	Category	Estimate	Std. Err.
Constant		−3.195	0.119
Socio-economic	low	–	
status	lower middle	0.420	0.118
	upper middle	0.739	0.114
	high	1.593	0.115
Parental	low	–	
encouragement	high	2.683	0.099

Furthermore, in each social stratum, the odds of making college plans among high school seniors with high parental encouragement are 14.6 times the odds among seniors with low parental encouragement.

The conclusions of this analysis are consistent with those from the previous subsection, except that this time we do not study the association between social stratification and parental encouragement, but focus on their effect on making college plans. In fact it is not just the conclusions, but all estimates and tests of significance, that agree. A comparison of the binomial deviances in Table 5.5 with the Poisson deviances in Table 5.3 shows the following 'coincidences':

log-linear model	*logit model*
$SE + P$	Null
$SE + SP$	S
$SE + EP$	E
$SE + SP + EP$	$S + E$

The models listed as equivalent have similar interpretations if you translate from the language of correlation analysis to the language of regression analysis. Note that all the log-linear models include the SE interaction, so they allow for association between the two predictors. Also, all of them include a main effect of the response P, allowing it to have a non-uniform distribution. The log-linear model with just these two terms assumes no association between P and either S or E, and is thus equivalent to the null logit model.

The log-linear model with an SP interaction allows for an association between S and P, and is therefore equivalent to the logit model where the response depends only on S. A similar remark applies to the log-linear model with an EP interaction. Finally, the log-linear model with all three

two-factor interactions allows for associations between S and P, and between E and P, and assumes that in each case the strength of association does not depend on the other variable. But this is exactly what the additive logit model assumes: the response depends on both S and E, and the effect of each factor is the same at all levels of the other predictor.

In general, log-linear and logit models are equivalent as long as the log-linear model

- is saturated on all factors treated as predictors in the logit model, including all possible main effects and interactions among predictors (in our example SE),
- includes a main effect for the factor treated as response (in our example P), and
- includes a two-factor (or higher order) interaction between a predictor and the response for each main effect (or interaction) included in the logit model (in our example it includes SP for the main effect of S, and son on).

This equivalence extends to parameter estimates as well as tests of significance. For example, multiplying the fitted probabilities based on the additive logit model $S + E$ by the sample sizes in each category of social status and parental encouragement leads to the same expected counts that we obtained earlier from the log-linear model $SE+SP+EP$. An interesting consequence of this fact is that one can use parameter estimates based on a log-linear model to calculate logits, as we did in Section 5.2.6, and obtain the same results as in logistic regression. For example the log of the odds ratio summarizing the effect of parental encouragement on college plans within each social stratum was estimated as 2.683 in the previous subsection, and this value agrees exactly with the estimate on Table 5.6.

In our example the equivalence depends crucially on the fact that the log-linear models include the SE interaction, and therefore reproduce exactly the binomial denominators used in the logistic regression. But what would have happened if the SE interaction had turned out to be not significant? There appear to be two schools of thought on this matter.

Bishop et al. (1975), in a classic book on the multivariate analysis of qualitative data, emphasize log-linear models because they provide a richer analysis of the structure of association among all factors, not just between the predictors and the response. If the SE interaction had turned out to be not significant they would probably leave it out of the model. They would

still be able to translate their parameter estimates into fitted logits, but the results would not coincide exactly with the logistic regression analysis (although they would be rather similar if the omitted interaction is small.)

Cox (1972), in a classic book on the analysis of binary data, emphasizes logit models. He argues that if your main interest is on the effects of two variables, say S and E on a third factor, say P, then you should condition on the SE margin. This means that if you are fitting log-linear models with the intention of understanding effects on P, you would include the SE interaction even if it is not significant. In that case you would get exactly the same results as a logistic regression analysis, which is probably what you should have done in the first place if you wanted to study specifically how the response depends on the predictors.