CHAPTER 7

Building and Applying Logit and Loglinear Models

Chapter 5 presented the logistic regression model, which uses the logit link for a binomial response. Chapter 6 presented the loglinear model that uses the log link for Poisson cell counts in a contingency table. Chapter 4 showed that they are both generalized linear models (GLMs), and Section 6.5 discussed equivalences between them. This chapter discusses further topics relating to building and applying these two types of models.

Section 7.1 introduces graphical representations that portray a model's association and conditional independence patterns. They also provide simple ways of indicating when conditional odds ratios are identical to marginal odds ratios.

The loglinear models of Chapter 6 treat all variables as nominal. Section 7.2 presents a loglinear model of association between ordinal variables. Inferences utilizing the ordering are more powerful than those that ignore it.

Section 7.3 presents tests of the hypothesis of conditional independence in three-way tables for nominal and ordinal variables. One approach compares the fits of two loglinear or logit models. A related approach uses generalized versions of the Cochran–Mantel–Haenszel test for multicategory responses.

Most inferential analyses in this text use large-sample approximations. Section 7.4 discusses the effects on model parameter estimates and goodness-of-fit statistics of small samples or zero cell counts. Finally, Section 7.5 summarizes theory underlying the fitting of logit and loglinear models and their use for large-sample inference.

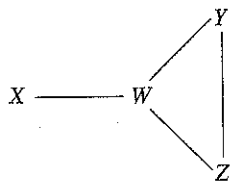
7.1 ASSOCIATION GRAPHS AND COLLAPSIBILITY

We begin by presenting a graphical representation for associations in loglinear models. For a model with a particular set of variables, the graph indicates which pairs of variables are independent and which pairs are associated, given the others. This representation is helpful for revealing implications of models, such as determining when marginal and conditional odds ratios are identical.

7.1.1 Association Graphs

An *association graph* for a loglinear model has a set of vertices, each vertex representing a variable. There are as many vertices as dimensions of the contingency table. An edge connecting two vertices represents a partial association between the corresponding two variables.

We illustrate for a four-way table with variables W, X, Y, Z. The loglinear model (WX, WY, WZ, YZ) lacks X-Y and X-Z association terms. It assumes that X and Y are independent and that X and Z are independent, conditional on the remaining two variables, but permits association between W and X and between each pair of variables in the set $\{W, Y, Z\}$. The association graph



portrays this model. The four variables form the vertices of the graph. The four edges, connecting W and X, W and Y, W and Z, and Y and Z, represent pairwise partial associations. Edges do not connect X and Y or X and Z, since those pairs are conditionally independent, given the remaining variables.

Two loglinear models that have the same pairwise associations have the same association graph. For instance, the association graph just portrayed for model (WX, WY, WZ, YZ) is also the one for model (WX, WYZ) that also contains a three-factor WYZ interaction.

A *path* in an association graph is a sequence of edges leading from one variable to another. Two variables X and Y are said to be *separated* by a subset of variables if all paths connecting X and Y intersect that subset. For instance, in the above graph, W separates X and Y, since any path connecting X to Y goes through W. The subset $\{W, Z\}$ also separates X and Y. A fundamental result states that two variables are conditionally independent given *any* subset of variables that separates them. Thus, not only are X and Y conditionally independent given W and Z, but also given W alone. Similarly, X and Z are conditionally independent given W alone.

For another example, we consider loglinear model (WX, XY, YZ). Its association graph is

$$W \text{——} X \text{——} Y \text{——} Z.$$

Since W and Z are separated by X, by Y, and by X and Y, this graph reveals that W and Z are independent given X alone or given Y alone or given both X and Y. Also, W and Y are independent, given X alone or given X and Z; X and Z are independent, given Y alone or given Y and W.

7.1.2 Collapsibility in Three-Way Tables

Section 3.1.4 showed that associations in partial tables may differ from marginal associations. For instance, if X and Y are conditionally independent, given Z, they are not necessarily marginally independent. We next present conditions under which a model's odds ratios are identical in partial tables as in the marginal table. These *collapsibility conditions* imply that the association is unchanged when we combine the partial tables.

> For three-way tables, X-Y marginal and partial odds ratios are identical if either Z and X are conditionally independent or if Z and Y are conditionally independent.

The conditions state that the variable treated as the control (Z) is conditionally independent of X or Y, or both. These conditions correspond to loglinear models (XY, YZ) and (XY, XZ). That is, the X-Y association is identical in the partial tables and the marginal table for models with association graphs

$$X \text{——} Y \text{——} Z \quad \text{and} \quad Y \text{——} X \text{——} Z$$

or even simpler models, but not for the model with graph X——Z——Y in which an edge connects Z to both X and Y.

We illustrate for the drug use data (Table 6.3) from Section 6.2.3, denoting A = alcohol use, C = cigarette use, and M = marijuana use. Consider (AM, CM), the model of conditional independence of A and C, given M, which has association graph

$$A \text{——} M \text{——} C.$$

Consider the A-M association, controlling for C; that is, we identify C with Z in the collapsibility conditions. In this model, since C is conditionally independent of A, the A-M partial odds ratios are the same as the A-M marginal odds ratio collapsed over C. In fact, Table 6.5 showed that both the fitted marginal and partial A-M odds ratios equal 61.9. Similarly, the C-M association is collapsible. The A-C association is not, however. The collapsibility conditions are not satisfied, because M is conditionally dependent with both A and C in model (AM, CM). Thus, A and C may be marginally dependent, even though they are conditionally independent in this model. In fact, Table 6.5 showed that the fitted A-C marginal odds ratio for this model equals 2.7, not 1.0.

The model (AC, AM, CM) of homogeneous association has association terms for each pair of variables, so no pair is conditionally independent. No collapsibility conditions are fulfilled. In fact, Table 6.5 showed that each pair of variables has quite different fitted marginal and partial associations for this model. When a model contains all two-factor effects, collapsing over any variable may cause effects to change.

7.1.3 Collapsibility and Logit Models

The collapsibility conditions apply also to corresponding logit models. For instance, suppose a clinical trial studies the association between a treatment variable X and a binary response Y, using data from several centers regarded as levels of a control variable Z. The logit model

$$\text{logit}(\pi) = \alpha + \beta_i^X + \beta_k^Z \tag{7.1.1}$$

for the probability π that Y is a "success" assumes that treatment effects are the same for each center. Since this logit model corresponds to loglinear model (XY, XZ, YZ), the estimated treatment effects may differ if we collapse the table over the center factor. That is, the estimated X-Y odds ratio for this model, $\exp(\hat{\beta}_1^X - \hat{\beta}_2^X)$, differs from the sample odds ratio in the marginal 2×2 table relating X and Y.

Next, consider the simpler model for this three-way table that lacks the center effects,

$$\text{logit}(\pi) = \alpha + \beta_i^X.$$

For each treatment, this model states that the success probability π is identical for each center. The partial and marginal treatment effects are identical for this model. It satisfies a collapsibility condition, because the model states that Z is conditionally independent of Y. This logit model is equivalent to loglinear model (XY, XZ) with association graph Y——X——Z, for which the X-Y association is collapsible. In practice, this suggests that when center effects seem negligible and the simpler model does not fit poorly compared to the full model (i.e., when center does not seem to be a "confounding" variable), one can collapse the table and estimate the treatment effect using the marginal odds ratio.

7.1.4 Collapsibility and Association Graphs for Multiway Tables

The next result provides collapsibility conditions for models for multiway tables.

> Suppose that variables in a model for a multiway table partition into three mutually exclusive subsets, A, B, C, such that B separates A and C; thus, the model does not contain parameters linking variables from A with variables from C. When one collapses the table over the variables in C, model parameters relating variables in A and model parameters relating variables in A with variables in B are unchanged.

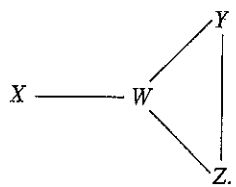
In other words, suppose that every path between a variable in A and a variable in C involves at least one variable in B. That is, the subsets of variables have the form

$$A \text{——} B \text{——} C.$$

When one collapses over the variables in C, the same parameter values relate the variables in A, and the same parameter values relate variables in A to variables in B.

It follows that the corresponding associations are unchanged, as described by odds ratios based on those parameters.

We illustrate using model (WX, WY, WZ, YZ) for a four-way table, which has association graph



Let $A = \{X\}, B = \{W\}$, and $C = \{Y, Z\}$. Since the X-Y and X-Z association parameters do not appear in this model, all parameters linking set A with set C equal zero, and B separates A and C. If we collapse over Y and Z, the W-X association is unchanged. Next, identify $A = \{Y, Z\}, B = \{W\}, C = \{X\}$. Then, partial associations among W, Y, and Z remain the same when the table is collapsed over X.

When the set B contains more than one variable, the collapsibility conditions require a slight qualifier. Though the true parameter values are unchanged when one collapses over set C, the ML estimates of those parameters may differ slightly. (The estimates are also identical if the model contains the highest-order term relating variables in B to each other.)

7.1.5 Model Building for the Dayton Drug-Use Example

Sections 6.2 and 6.3 analyzed data on usage of alcohol (A), cigarettes (C), and marijuana (M) by a sample of high school seniors. When the students are also classified by the demographic factors gender (G) and race (R), the five-dimensional contingency table shown in Table 7.1 results. In selecting a model for these data, we treat A, C, and M as response variables and G and R as explanatory variables.

Table 7.1 Alcohol, Cigarette, and Marijuana Use for High School Seniors by Gender and Race

Race		White				Other			
Gender		Female		Male		Female		Male	
		Marijuana Use							
Alcohol Use	Cigarette Use	Yes	No	Yes	No	Yes	No	Yes	No
Yes	Yes	405	268	453	228	23	23	30	19
	No	13	218	28	201	2	19	1	18
No	Yes	1	17	1	17	0	1	1	8
	No	1	117	1	133	0	12	0	17

Source: Prof. Harry Khamis, Wright State University.

Table 7.2 Goodness-of-Fit Tests for Loglinear Models Relating Alcohol (A), Cigarette (C), and Marijuana (M) Use, by Gender (G) and Race (R)

Model	G^2	df
1. Mutual Independence + GR	1325.1	25
2. Homogeneous Association	15.3	16
3. All Three-Factor Terms	5.3	6
4a. (2) − AC	201.2	17
4b. (2) − AM	107.0	17
4c. (2) − CM	513.5	17
4d. (2) − AG	18.7	17
4e. (2) − AR	20.3	17
4f. (2) − CG	16.3	17
4g. (2) − CR	15.8	17
4h. (2) − GM	25.2	17
4i. (2) − MR	18.9	17
5. $(AC, AM, CM, AG, AR, GM, GR, MR)$	16.7	18
6. $(AC, AM, CM, AG, AR, GM, GR)$	19.9	19
7. (AC, AM, CM, AG, AR, GR)	28.8	20

Since G and R are explanatory, it does not make sense to estimate association or assume conditional independence for that pair. It follows from remarks near the end of Section 6.5.4 that a model should contain the G-R term. Including this term forces the G-R fitted marginal totals to be the same as the corresponding sample marginal totals. The sample contained 1040 white females, for instance, so the model's fitted total of white females then equals 1040.

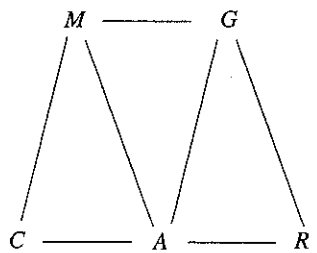
Table 7.2 displays results of goodness-of-fit tests for several loglinear models. Because many cell counts are small, the chi-squared approximation for G^2 may be poor. It is best not to take the G^2 values too seriously for a particular model, but this index is useful for comparing models. The first model listed in Table 7.2 contains only the G-R association and assumes conditional independence for the other nine pairs of associations. It fits horribly, which is no surprise. The homogeneous association model, on the other hand, seems to fit well. The only large adjusted residual results from a fitted value of 3.1 in the cell having a count of 8.

The model containing all the three-factor interaction terms also fits well, but the improvement in fit is not great (difference in G^2 of $15.3 - 5.3 = 10.0$ based on $df = 16 - 6 = 10$). Thus, we consider models without three-factor terms. Beginning with the homogeneous association model as the baseline, we eliminate two-factor associations that do not make significant contributions. We use a backward elimination process, sequentially taking out terms for which the resulting increase in G^2 is smallest, when refitting the model. However, we do not delete the G-R association term relating the explanatory variables.

Table 7.2 shows the start of this process. Nine pairwise associations are candidates for removal from model (2), shown in models numbered (4a)–(4i) in the table. The smallest increase in G^2, compared to model (2), occurs in removing the C-R term. The increase is $15.8 - 15.3 = 0.5$, based on $df = 17 - 16 = 1$, so this elimination

seems reasonable. After removing the C-R term (model 4g), the smallest additional increase results from removing the C-G term (model 5), resulting in $G^2 = 16.7$ with $df = 18$, and a change in G^2 of 0.9 based on $df = 1$. Removing next the M-R term (model 6) yields $G^2 = 19.9$ with $df = 19$, a change in G^2 of 3.2 based on $df = 1$.

At this stage, the only large adjusted residual refers to a fitted value of 2.9 in the cell having a count of 8. Additional removals have a more severe effect. For instance, removing next the A-G term increases G^2 by 5.3, based on $df = 1$, for a P-value of .02. One cannot take such P-values too literally, since these tests are suggested by the data, but it seems safest not to drop additional terms. Model (6), denoted by $(AC, AM, CM, AG, AR, GM, GR)$, has association graph



Consider the sets $\{C\}$, $\{A, M\}$, and $\{G, R\}$. For this model, every path between C and $\{G, R\}$ involves a variable in $\{A, M\}$. Given the outcome on alcohol use and marijuana use, the model states that cigarette use is independent of both gender and race. Collapsing over the explanatory variables race and gender, the partial associations between C and A and between C and M are the same as with the model (AC, AM, CM) fitted in Section 6.2.3.

Suppose we remove the G-M term from this model, yielding (AC, AM, CM, AG, AR, GR), model (7) in Table 7.2. Its association graph reveals that $\{G, R\}$ are separated from $\{C, M\}$ by A. It follows that all pairwise partial associations among A, C, and M in model (7) are identical to those in model (AC, AM, CM), collapsing over G and R. In fact, model (7) does not fit all that poorly ($G^2 = 28.8$ with $df = 20$, and only one adjusted residual exceeds 3), especially considering the large sample size. Its sample dissimilarity index equals $D = .036$. For practical purposes, one may be able to collapse over gender and race in studying associations among the drug-use variables. An advantage of the full five-variable model portrayed by the above graph is that one can also study the effects of gender and race on these responses, in particular the effects of race and gender on alcohol use and the effect of gender on marijuana use.

7.2 MODELING ORDINAL ASSOCIATIONS

The loglinear models presented so far have a serious limitation: they treat all classifications as nominal. If we change the order of a variable's categories in any way, we get the same fit. For ordinal data, these models ignore important information.

Table 7.3 Fit of Independence Model and Adjusted Residuals for Opinion about Premarital Sex and Availability of Teenage Birth Control

	Teenage Birth Control			
Premarital Sex	Strongly Disagree	Disagree	Agree	Strongly Agree
Always wrong	81	68	60	38
	(42.4)	(51.2)	(86.4)	(67.0)
	7.6	3.1	−4.1	−4.8
Almost always wrong	24	26	29	14
	(16.0)	(19.3)	(32.5)	(25.2)
	2.3	1.8	−0.8	−2.8
Wrong only sometimes	18	41	74	42
	(30.1)	(36.3)	(61.2)	(47.4)
	−2.7	1.0	2.2	−1.0
Not wrong at all	36	57	161	157
	(70.6)	(85.2)	(143.8)	(111.4)
	−6.1	−4.6	2.4	6.8

Source: 1991 General Social Survey.

Table 7.3, taken from the 1991 General Social Survey, illustrates the inadequacy of ordinary loglinear models for analyzing ordinal data. Subjects were asked their opinion about a man and woman having sex relations before marriage, with possible responses "always wrong," "almost always wrong," "wrong only sometimes," and "not wrong at all." They were also asked if they "strongly disagree," "disagree," "agree," or "strongly agree" that methods of birth control should be made available to teenagers between the ages of 14 and 16. Both classifications have ordered categories.

For these data, the loglinear model of independence (i.e., model (6.1.1)), which we denote by I, has goodness-of-fit statistics $G^2(I) = 127.6$ and $X^2(I) = 128.7$, based on $df = 9$. These tests of fit are simply the tests of independence presented in Section 2.4. The model fits poorly, providing strong evidence of dependence. Yet, adding the ordinary association term makes the model saturated (model (6.1.2)) and of little use.

Table 7.3 also contains fitted values and adjusted residuals (Section 2.4.5) for the independence model. The residuals in the corners of the table are very large. Observed counts are much larger than the independence model predicts in the corners where both responses are the most negative possible ("always wrong" with "strongly disagree") or the most positive possible ("not wrong at all" with "strongly agree"). By contrast, observed counts are much smaller than fitted counts in the other two corners, where one response is the most positive and the other is the most negative. Cross-classifications of ordinal variables often exhibit their greatest deviations from independence in the corner cells. This pattern for Table 7.3 indicates lack of fit in the form of a positive trend. Subjects who feel more favorable to making birth control available to teenagers also tend to feel more tolerant about premarital sex.

The independence model is too simple to fit most data well. Models for ordinal variables use association terms that permit negative or positive association trends. The models are more complex than the independence model yet simpler than the saturated model.

7.2.1 Linear-by-Linear Association

This section presents an ordinal loglinear model for two-way tables. It requires assigning scores $\{u_i\}$ to the I rows and $\{v_j\}$ to the J columns. To reflect category orderings, $u_1 \leq u_2 \leq \cdots \leq u_I$ and $v_1 \leq v_2 \leq \cdots \leq v_J$. The model is

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \beta u_i v_j. \tag{7.2.1}$$

The independence model is the special case $\beta = 0$.

Model (7.2.1) has form

$$\log \mu_{ij} = \text{independence} + \beta u_i v_j.$$

The final term represents the deviation of $\log \mu_{ij}$ from independence. The deviation is linear in the Y scores at a fixed level of X and linear in the X scores at a fixed level of Y. In column j, for instance, the deviation is a linear function of X, having form (slope) $\times$ (score for X), with slope βv_j. Because of this property, (7.2.1) is called the *linear-by-linear association model* (abbreviated, $L \times L$). This linear-by-linear deviation implies that the model has its greatest departures from independence in the corners of the table.

The parameter β in model (7.2.1) refers to the direction and strength of association. When $\beta > 0$, there is a tendency for Y to increase as X increases. Expected frequencies are larger than expected (under independence) in cells of the table where X and Y are both high or both low. When $\beta < 0$, there is a tendency for Y to decrease as X increases, and for expected frequencies to be relatively larger in cells where X is high and Y is low or where X is low and Y is high. When the data display a positive or negative trend, this model usually fits much better than the independence model.

We describe associations for this model using odds ratios for pairings of categories. For the 2×2 table using the cells intersecting rows a and c with columns b and d, the model has odds ratio equal to

$$\frac{\mu_{ab}\mu_{cd}}{\mu_{ad}\mu_{cb}} = \exp[\beta(u_c - u_a)(v_d - v_b)]. \tag{7.2.2}$$

The association is stronger as $|\beta|$ increases. For given β, pairs of categories that are farther apart have greater differences between their scores and odds ratios farther from 1.

In practice, the most common choice of scores is $\{u_i = i\}$ and $\{v_j = j\}$, simply the row and column numbers. These scores have equal spacings of 1 between each pair of adjacent row or column scores. The odds ratios formed using adjacent rows and adjacent columns are called *local odds ratios*. For these unit-spaced scores, (7.2.2) simplifies so that e^β is the common value of all the local odds ratios. Any set of equally-spaced row and column scores has the property of uniform local odds ratios. This special case of the model is called *uniform association*. Figure 7.1 portrays some of the local odds ratios that take uniform value in this model.

Fitting the linear-by-linear association model requires iterative methods. The model's fitted values, like those for the independence model, have the same row and column totals as the observed data. In addition, the correlation between the row scores for X and the column scores for Y is the same for the observed counts as it is for the joint distribution given by the fitted counts. Thus, the fitted counts display the same positive or negative trend as the observed data. Unlike the observed data, the fitted counts exactly satisfy the odds ratio pattern (7.2.2) implied by the model. Since the model has one more parameter (β) than the independence model, its residual $df = IJ - I - J$ are 1 less; it is unsaturated whenever $I > 2$ or $J > 2$.

7.2.2 Sex Opinions Example

Table 7.4 reports fitted values for the linear-by-linear ($L \times L$) association model applied to the opinions about premarital sex and availability of teen birth control, using row scores $\{1, 2, 3, 4\}$ and column scores $\{1, 2, 3, 4\}$. The goodness-of-fit statistics for this uniform association version of the model are $G^2(L \times L) = 11.5$ and $X^2(L \times L) = 11.5$, with $df = 8$. Compared to the independence model, for which $G^2(I) = 127.6$ with $df = 9$, the $L \times L$ model provides a dramatic improvement in fit. This is especially noticeable in the corners of the table.

The ML estimate of the association parameter is $\hat{\beta} = 0.286$, with $ASE = 0.028$. The positive estimate suggests that subjects having more favorable attitudes about availability of teen birth control also tend to have more tolerant attitudes about premarital sex. The estimated local odds ratio is $\exp(\hat{\beta}) = \exp(0.286) = 1.33$. The strength of association seems weak. From (7.2.2), however, nonlocal odds ratios are

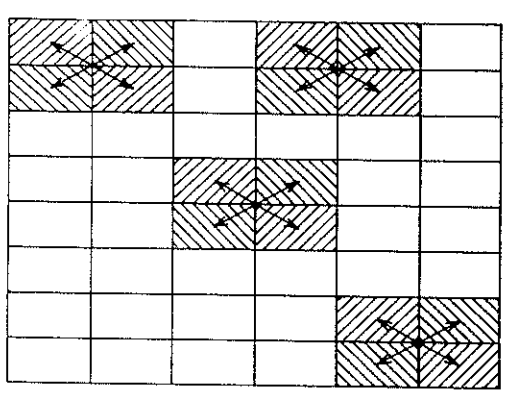


Figure 7.1 Constant local odds ratio implied by uniform association model. (*Note:* β = the constant log odds ratio for adjacent rows and adjacent columns.)

Table 7.4 Fit of Linear-by-Linear Association Model for Table 7.3

	Teenage Birth Control			
Premarital Sex	Strongly Disagree	Disagree	Agree	Strongly Agree
Always wrong	81	68	60	38
	(80.9)	(67.6)	(69.4)	(29.1)
Almost always wrong	24	26	29	14
	(20.8)	(23.1)	(31.5)	(17.6)
Wrong only sometimes	18	41	74	42
	(24.4)	(36.1)	(65.7)	(48.8)
Not wrong at all	36	57	161	157
	(33.0)	(65.1)	(157.4)	(155.5)

stronger. For instance, the estimated odds ratio for the four corner cells equals

$$\exp[\hat{\beta}(u_4 - u_1)(v_4 - v_1)] = \exp[0.286(4-1)(4-1)] = \exp(2.57) = 13.1.$$

One can also obtain this value directly using the fitted values for the corner cells in Table 7.4; that is, $(80.9)(155.5)/(29.1)(33.0) = 13.1$. For those who "strongly agree" with availability of teen birth control, the odds of response "not wrong at all" instead of "always wrong" on premarital sex are estimated to be 13.1 times the odds for those who "strongly disagree" with availability of teen birth control.

Two sets of scores having the same spacings yield the same estimate of β and the same fit. For instance, $\{u_1 = 1, u_2 = 2, u_3 = 3, u_4 = 4\}$ yields the same results as $\{u_1 = -1.5, u_2 = -0.5, u_3 = 0.5, u_4 = 1.5\}$. Any other sets of equally spaced scores yield the same fit but an appropriately rescaled estimate of β, so that the fitted odds ratios (7.2.2) do not change. For instance, the row scores $\{2, 4, 6, 8\}$ with $\{v_j = j\}$ also yield $G^2 = 11.5$, but have $\hat{\beta} = 0.143$ with standard error 0.014 (both half as large).

It is not necessary to use equally-spaced scores in the $L \times L$ model. For the classifications in Table 7.3, one might regard categories 2 and 3 as farther apart than categories 1 and 2 or categories 3 and 4. To recognize this, one could assign scores such as $\{1, 2, 4, 5\}$ to the row and column categories. The $L \times L$ model then has $G^2 = 8.8$. One need not, however, regard the model scores as approximations for distances between categories. They simply imply a certain pattern for the relative sizes of the odds ratios. From (7.2.2), fitted local odds ratios are stronger for pairs of adjacent categories having greater distances between scores. More general models, not discussed here, treat the row and/or column scores as parameters estimated using the data.

7.2.3 Ordinal Tests of Independence

For the linear-by-linear association model, the hypothesis of independence is H_0 : $\beta = 0$. The likelihood-ratio test statistic equals the reduction in G^2 goodness-of-fit statistics between the independence (I) and $L \times L$ models,

$$G^2(I \mid L \times L) = G^2(I) - G^2(L \times L). \tag{7.2.3}$$

This statistic refers to a single parameter (β), and is based on $df = 1$. For Table 7.3, the reduction is $127.6 - 11.5 = 116.1$. This has $P < .0001$, extremely strong evidence of an association.

The Wald statistic $z^2 = (\hat{\beta}/ASE)^2$ provides an alternative test for this hypothesis. It is also a chi-squared statistic with $df = 1$. For these data, $z^2 = (0.286/0.0282)^2 = 102.4$, also showing strong evidence of a positive trend. The correlation statistic (2.5.1) presented in Section 2.5.1 for testing independence with ordinal data is usually similar to the likelihood-ratio and Wald statistics for testing $\beta = 0$ in this model. (In fact, it is the efficient score statistic.) For Table 7.3, it equals 112.6, also based on $df = 1$.

Generalizations of the linear-by-linear association model exist for multi-way tables. We discuss one of these in the following section. Sections 8.2 and 8.3 present other ways of using ordinality, based on models that create logits for an ordinal response variable.

7.3 TESTS OF CONDITIONAL INDEPENDENCE

This section discusses ways of testing the hypothesis of conditional independence in three-way tables. Likelihood-ratio tests compare the fit of two loglinear or logit models. Alternatively, one can use generalizations of the Cochran–Mantel–Haenszel statistic.

7.3.1 Using Models to Test Conditional Independence

Section 6.3.4 showed how to test a partial association by comparing two loglinear models that contain or omit that association. The likelihood-ratio test compares the models by the difference of the G^2 goodness-of-fit statistics, which is identical to the difference of *deviances* (Section 4.5.3).

An important application of this test refers to the null hypothesis of X-Y conditional independence. One compares the model (XZ, YZ) of X-Y conditional independence to the more complex model (XY, XZ, YZ) that contains the X-Y association. The test statistic is $G^2[(XZ, YZ) \mid (XY, XZ, YZ)] = G^2(XZ, YZ) - G^2(XY, XZ, YZ)$. This test assumes that the homogeneous association model (XY, XZ, YZ) holds. It is a test of H_0 : all $\lambda_{ij}^{XY} = 0$ for this model.

When Y is binary, this test relates to logit models. The model for the logit of the probability π that $Y = 1$,

$$\text{logit}(\pi) = \alpha + \beta_i^X + \beta_k^Z,$$

corresponds to loglinear model (XY, XZ, YZ). The null hypothesis of X-Y conditional independence is H_0 : all $\beta_i^X = 0$ for this model. The likelihood-ratio test statistic is

the difference between G^2 statistics for the reduced model $\text{logit}(\pi) = \alpha + \beta_k^Z$ and the full model for the three-way table.

For $2 \times 2 \times K$ tables, the test of conditional independence comparing two loglinear or logit models has the same purpose as the Cochran–Mantel–Haenszel (*CMH*) test (Section 3.2). The *CMH* test works well when the *X-Y* odds ratio is similar in each partial table. In this sense, it is also naturally directed toward the alternative of homogeneous association. For large samples, the model-based likelihood-ratio test usually gives similar results as the *CMH* test. In fact, the *CMH* procedure is an efficient score test (Section 4.5.2) of the hypothesis that the *X-Y* association parameters equal zero in the loglinear or logit model of homogeneous association.

7.3.2 Job Satisfaction and Income Example

Table 7.5, from the 1991 General Social Survey, refers to the relationship between job satisfaction (S) and income (I), stratified by gender (G). The test of the hypothesis of *I-S* conditional independence compares the conditional independence model (GI, GS) to model (IS, GI, GS). This analysis checks whether one can eliminate the *I-S* association term from model (IS, GI, GS), assuming that that model holds. The fit statistics are $G^2(GI, GS) = 19.4$, with $df = 18$, and $G^2(IS, GI, GS) = 7.1$, with $df = 9$. Comparing the models, $G^2[(GI, GS) \mid (IS, GI, GS)] = 19.4 - 7.1 = 12.3$, based on $df = 18 - 9 = 9$. This gives $P = .20$ and does not provide evidence of an association.

7.3.3 Direct Goodness-of-Fit Test

Another way to test the hypothesis of *I-S* conditional independence compares the model (GI, GS) directly to the saturated model. That is, one tests the hypothesis by performing a goodness-of-fit test of the model. For Table 7.5, $G^2(GI, GS) = 19.4$ with $df = 18$. This test of conditional independence has P-value of .37. Again, *I-S* conditional independence is plausible.

Table 7.5 Job Satisfaction and Income, Controlling for Gender

		Job Satisfaction			
Gender	Income	Very Dissatisfied	A Little Satisfied	Moderately Satisfied	Very Satisfied
Female	< 5000	1	3	11	2
	5000–15,000	2	3	17	3
	15,000–25,000	0	1	8	5
	> 25,000	0	2	4	2
Male	< 5000	1	1	2	1
	5000–15,000	0	3	5	1
	15,000–25,000	0	0	7	3
	> 25,000	0	1	9	6

Source: 1991 General Social Survey.

A statistic of form $G^2(XZ, YZ)$ does not require an assumption about homogeneous association. Since it is identical to $G^2[(XZ, YZ) \mid (XYZ)] = G^2(XZ, YZ) - G^2(XYZ)$, the null hypothesis for this test is H_0 : all $\lambda_{ij}^{XY} = 0$ and all $\lambda_{ijk}^{XYZ} = 0$ in the saturated loglinear model. The statistic could be large if there is three-factor interaction, or if there is no three-factor interaction but conditional dependence.

This test has the advantage of not assuming model structure, such as homogeneous association. A disadvantage is that it often has low power. If there truly is no (or little) three-factor interaction, the *CMH* test and the likelihood-ratio comparison statistic $G^2[(XZ, YZ) \mid (XY, XZ, YZ)]$ are more likely to yield small P-values. In testing that the *X-Y* association parameters alone are zero, those chi-squared tests focus the analysis on fewer degrees of freedom. Capturing an effect with a smaller df value yields a test with greater power (Section 2.5.3). Unless the degree of heterogeneity in *X-Y* partial associations is severe, it is better to use the test having an unsaturated baseline model.

The baseline models in this test and the test using $G^2[(GI, GS) \mid (IS, GI, GS)]$ treat income and job satisfaction as nominal, but they are ordinal. More powerful tests of conditional independence exploit the ordinality. We next construct a test using an ordinal loglinear model.

7.3.4 Detecting Ordinal Conditional Association

Generalizations of the linear-by-linear association model (7.2.1) apply to modeling association between ordinal variables X and Y while controlling for a third variable that may be nominal or ordinal. A useful model,

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \beta u_i v_j + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}, \qquad (7.3.1)$$

is the special case of model (XY, XZ, YZ) that replaces the general *X-Y* association term λ_{ij}^{XY} by the simpler linear-by-linear (abbreviated, $L \times L$) term $\beta u_i v_j$ based on ordered row and column scores. The form of the $L \times L$ association is the same in each partial table, since the term $\beta u_i v_j$ does not depend on k. This model is appropriate when the *X-Y* association has the same positive or negative trend at each level of Z. The model is called a *homogeneous linear-by-linear association model.*

The conditional independence model (XZ, YZ) is the special case of this model with $\beta = 0$. One can use the ordinality of X and Y in testing conditional independence by comparing G^2 for model (7.3.1) to G^2 for model (XZ, YZ), or by forming the Wald statistic $z^2 = (\hat{\beta}/ASE)^2$. These statistics concentrate evidence about the association on a single degree of freedom. Unless model (7.3.1) fits very poorly, these tests are more powerful than tests that ignore the ordering.

For Table 7.5 with scores $\{1, 2, 3, 4\}$ for income and job satisfaction, the homogeneous $L \times L$ model (7.3.1) has $G^2 = 12.3$, with $df = 17$; by comparison, $G^2(GI, GS) = 19.4$ with $df = 18$. The difference equals 7.1, with $df = 1$, and has $P = .01$ for testing conditional independence. This contrasts with the lack of evidence provided by models that ignore the ordinality.

The positive ML estimate $\hat{\beta} = 0.388$, based on $ASE = 0.155$, reveals a tendency for job satisfaction to be greater at higher levels of income. The estimated local odds ratio for each gender is $\exp(0.388) = 1.47$. The Wald statistic $z^2 = (0.388/0.155)^2 = 6.3$, with $df = 1$, also provides strong evidence of an association ($P = .01$).

7.3.5 Generalized Cochran–Mantel–Haenszel Tests

Alternative tests of conditional independence generalize the Cochran–Mantel–Haenszel (*CMH*) statistic (3.2.1) to $I \times J \times K$ tables. Like the *CMH* statistic and the model-based statistic $G^2[(XZ, YZ) \mid (XY, XZ, YZ)]$, these statistics perform well when the X-Y association is similar at each level of Z. There are three versions, according to whether both, one, or neither of X and Y are treated as ordinal.

When X and Y are ordinal, the test statistic generalizes the correlation statistic (2.5.1) for two-way tables. It is designed to detect a linear trend in the X-Y association that has the same direction in each partial table. The statistic takes larger value as the correlation increases in magnitude and as the sample size grows in each table.

The generalized correlation statistic has approximately a chi-squared distribution with $df = 1$. Its formula is complex (Agresti (1990), p. 284), and we omit computational details since it is available in standard software. In fact, it is the efficient score statistic for testing that $\beta = 0$ in model (7.3.1). For large samples, it usually gives similar results as the likelihood-ratio statistic or the Wald statistic for that hypothesis.

For Table 7.5 with the row and column numbers as the scores, the sample correlation between income and job satisfaction equals .171 for females and .381 for males. The generalized correlation statistic equals 6.6 with $df = 1$ ($P = .01$), giving the same conclusion as the likelihood-ratio and Wald tests of the previous subsection.

Often scores other than the row and column numbers are more sensible. For instance, with grouped continuous variables, one might use scores that are midpoints of the class intervals. For Table 7.5, the row scores $(3, 10, 20, 35)$ use midpoints of the middle two categories, in thousands of dollars. Alternatively, midrank scores are based on the relative numbers of observations in the various response categories. When in doubt about scoring, perform a sensitivity analysis by using a few different choices that seem sensible. Unless the categories exhibit severe imbalance in their totals, the choice of scores usually has little impact on the conclusion. Different choices yield similar results for Table 7.5. For the scores $(3, 10, 20, 35)$ for income and $(1, 3, 4, 5)$ for satisfaction, for instance, the generalized correlation statistic equals 6.2 with $df = 1$ ($P = .01$).

7.3.6 Detecting Nominal–Ordinal Conditional Association

When X is nominal and Y is ordinal, scores are relevant only for levels of Y. We summarize the responses of subjects within a given row by the mean of their scores on the ordinal variable and then average this row-wise mean information across the K strata. The test of conditional independence compares the I rows using a statistic based on the variation in those I averaged row mean responses that is designed to detect differences among their true values. It has a large-sample chi-squared distribution

with $df = (I - 1)$. The power is strong when the differences among the row means are similar in each partial table.

The formula for the test statistic is complex (Agresti (1990), pp. 286–287). We present it only for the special case of a two-way table (i.e., $K = 1$), to illustrate the basic idea. For column scores $\{v_j\}$, let $\bar{y}_i = \sum_j v_j n_{ij}/n_{i+}$. The numerator sums the scores on Y for all subjects in row i, and the denominator is the sample size for that row. The measure $\bar{y}_i$ is the sample mean response on the ordinal variable Y in row i, for the chosen scores. Let $\bar{y} = \sum_j v_j n_{+j}/n$ denote the mean response on Y for the overall sample, using the column totals. The test statistic equals

$$(n-1)\frac{\sum_i n_{i+}(\bar{y}_i - \bar{y})^2}{\sum_j v_j^2 n_{+j} - n\bar{y}^2}. \tag{7.3.2}$$

If Y were normally distributed, a one-way ANOVA would compare means of Y for I levels of a nominal variable X. Thus, statistic (7.3.2) is an analog of a one-way ANOVA statistic when Y is ordinal categorical rather than continuous. In fact, with midrank scores for $\{v_j\}$, it is the *Kruskal–Wallis* statistic for comparing mean ranks for I groups. When $I = 2$, it is identical to the correlation statistic (2.5.1), which then compares two row means. For K partial tables, the formula for the test of whether the row mean scores differ is complex, but it is available in standard software. It is the efficient score test for a generalization of model (7.3.1) in which the row scores are parameters.

For Table 7.5, this test treats job satisfaction as ordinal and income (the row variable) as nominal. The test searches for differences among the four income levels in their mean job satisfaction. Using scores $\{1, 2, 3, 4\}$, the mean job satisfaction at the four levels of income equal $(2.82, 2.84, 3.29, 3.00)$ for females and $(2.60, 2.78, 3.30, 3.31)$ for males. For instance, the mean for the 17 females with income < 5000 equals $[1(1) + 2(3) + 3(11) + 4(2)]/17 = 2.82$. The pattern of means is similar for each gender, roughly increasing as income increases. The generalized *CMH* statistic for testing whether the true row mean scores differ equals 9.2 with $df = 3$ ($P = .03$).

Unlike this statistic, the correlation statistic of the previous subsection also treats the rows as ordinal. It detects a linear trend across rows in the row mean scores, and it utilizes the approximate increase in mean satisfaction as income increases. One can use the nominal-ordinal statistic based on variability in row means when X and Y are ordinal, but such a linear trend may not occur. For instance, one might expect responses on Y to tend to be higher in some rows than in others, without the mean of Y increasing consistently or decreasing consistently as the level of X increases. An analogous test treats X as ordinal and Y as nominal and compares J mean scores on X computed within columns. It has $df = J - 1$.

7.3.7 Detecting Nominal–Nominal Conditional Association

Another *CMH*-type statistic, based on $df = (I - 1)(J - 1)$, provides a "general association" test. It is designed to detect *any* type of association that is similar in

Table 7.6 Summary of Generalized Cochran–Mantel–Haenszel Tests of Conditional Independence for Table 7.5

Alternative Hypothesis	Statistic	df	P-value
General association	10.2	9	.34
Row mean scores differ	9.2	3	.03
Nonzero correlation	6.6	1	.01

each partial table. It treats both X and Y as nominal, so does not require category scores. Because of its complexity (Agresti (1990), pp. 234–235), we do not present its formula, but it is available in software. It is an efficient score test of X-Y conditional independence for loglinear model (XY, XZ, YZ). For a single table ($K = 1$), the general association statistic is similar to the Pearson chi-squared statistic, equaling $[n/(n-1)]X^2$.

For Table 7.5, the general association statistic equals 10.2, with $df = 9$ ($P = .34$). We pay a price for ignoring the ordinality of job satisfaction and income. For ordinal variables, the general association test is usually not as powerful as narrower tests with smaller df values that use the ordinality.

Table 7.6 summarizes results of the three generalized *CMH* tests for $I \times J \times K$ tables applied to Table 7.5. (The format is similar to that used by SAS with the CMH option in PROC FREQ.) The *general association* alternative treats both X and Y as nominal, and has $df = (I-1)(J-1) = 9$. It is sensitive to any departure that is similar in each level of Z. The *row mean scores differ* alternative treats the rows of X as nominal and the columns of Y as ordinal and has $df = I - 1 = 3$. It is sensitive to variation among the I mean scores on Y computed within levels of X, when the nature of that variation is similar in each level of Z. Finally, the *nonzero correlation* alternative treats both X and Y as ordinal and has $df = 1$. Its test statistic is sensitive to a linear trend between X and Y that is similar in each level of Z. When $I = J = 2$, all three test statistics have $df = 1$ and simplify to the *CMH* statistic (3.2.1).

7.4 EFFECTS OF SPARSE DATA

This section discusses the effects of small cell counts on the fitting of loglinear models and logit models to contingency tables. Tables having many cells with small counts are said to be *sparse*. Sparse tables occur when the sample size is small. They also occur when the sample size is large but so is the number of cells. Sparseness is common in tables with many variables or with classifications having several categories.

7.4.1 Empty Cells

Sparse tables usually contain cells with zero counts. Such cells are called *empty cells* and are of two types: *sampling zeroes* and *structural zeroes*.

In most cases, even though a cell is empty, its true probability is positive. That is, it is theoretically possible to have observations in the cell, and a positive count would occur if the sample size were sufficiently large. This type of empty cell is called a *sampling zero*. The empty cells in Table 7.1 on drug use and in Table 7.5 on job satisfaction are sampling zeroes.

An empty cell in which observations are theoretically impossible is called a *structural zero*. Such cells have true probabilities equal to zero, and the cell count is zero regardless of the sample size. To illustrate, suppose that professors employed in a given department at the University of Rochester for at least five years were cross-classified on their current rank (assistant professor, associate professor, professor) and their rank five years ago. Professors cannot be demoted in rank, so three of the nine cells in the table contain structural zeroes. One of these is the cell corresponding to the rank of professor five years ago and assistant professor now; it cannot contain any observations. Contingency tables containing structural zeroes are called *incomplete tables*.

Sampling zeroes are part of the observed data set. For instance, a count of 0 is a possible outcome for a Poisson variate. It contributes to the likelihood function and the model-fitting process. A structural zero, on the other hand, is not an observation and is not part of the data. Sampling zeroes are much more common than structural zeroes, and the remaining discussion refers to them.

Sampling zeroes can affect the existence of ML estimates of loglinear and logit model parameters. When all cell counts are positive, parameter estimates are necessarily finite. When any marginal counts corresponding to qualitative terms in a model equal zero, infinite estimates occur for that term. For instance, when any X-Y marginal totals equal zero, infinite estimates occur among $\{\hat{\lambda}_{ij}^{XY}\}$ for loglinear models such as (XY, XZ, YZ) and (XY, XZ), and infinite estimates occur among $\{\hat{\beta}_i^X\}$ for the effect of X on Y in logit models.

A value of ∞ (or $-\infty$) for a parameter estimate means that the likelihood function keeps increasing as the parameter moves toward ∞ ($-\infty$). Such results imply that ML fitted values equal 0 in some cells, and some odds ratio estimates have values of ∞ or 0. Most software cannot distinguish and does not indicate when infinite estimates occur. One potential sign is when the iterative process for fitting the model does not converge, typically because a parameter estimate keeps getting larger from cycle to cycle. Another sign is when the software reports large estimates, in relative terms, with very large estimated standard errors. Slight changes in the data then often cause dramatic changes in the estimates and their standard errors. A danger with sparse data is that one might not realize that a true estimated effect is infinite and, as a consequence, report estimated effects and results of statistical inferences that are invalid and highly unstable.

Empty cells and sparse tables can cause severe bias in estimators of odds ratios and poor chi-squared approximations for goodness-of-fit statistics. One remedy to estimation bias is to add a small constant to cell counts before conducting an analysis. For saturated models, adding $\frac{1}{2}$ to each cell reduces the bias in sample odds ratio estimators (Section 2.3.3). For instance, this shrinks infinite (or zero) estimates of odds ratios to finite values corresponding to positive probabilities in all the cells. For

unsaturated models, though, adding $\frac{1}{2}$ to each cell before fitting the model smooths the data too much, causing havoc with sampling distributions. This operation has too conservative an influence on fitted odds ratios and test statistics.

Many ML analyses for unsaturated models are unharmed by empty cells. For instance, when a single cell is empty, finite estimates exist for all parameters in unsaturated models presented so far. In fact, they usually exist when all the marginal totals corresponding to terms in the model are positive. Even when a parameter estimate is infinite, this is not fatal to data analysis. Though an infinite estimate for an odds ratio is rather unsatisfactory, one can construct a confidence interval for the true odds ratio for which one bound is finite (Section 5.7.3).

When iterative fitting processes fail to converge because of infinite estimates, adding a very small constant (such as 10^{-8}) is adequate for ensuring convergence. One can then estimate parameters for which the true estimates are finite and are not affected by the empty cells, as the example in the following subsection shows. When in doubt about the effect of empty cells, one should perform a sensitivity analysis. Repeat the analysis by adding constants of various sizes, (say .00000001, .0001, .01, .1) in order to gauge their effect on parameter estimates and goodness-of-fit statistics. The total count added should be only a tiny percentage of the total sample size. Also, for each possibly influential observation, delete it or move it to another cell to see how much the results vary with small perturbations to the data. Often, some associations are not affected by the empty cells and give stable results for the various analyses, whereas others that are affected are highly unstable. Use caution in making conclusions about an association if small changes in the data are highly influential. In some cases, it makes sense to fit the model by excluding part of the data containing empty cells or by combining that part with other parts of the data.

An alternative to ML estimation, using Bayesian methods, provides a way of smoothing data in a less *ad hoc* manner than adding arbitrary constants to cells. Bayesian methods are beyond the scope of this text, but Bishop et al. ((1975), Ch. 12) and Agresti ((1990), Sec. 13.4) describe their use for dealing with sparse data.

7.4.2 Clinical Trials Example

Table 7.7 shows results of a clinical trial conducted at five centers. The purpose was to compare an active drug to placebo in terms of a binary (success, failure) response for treating fungal infections. For these data, let C = Center, T = Treatment (Active drug or Placebo), and R = Response.

Centers 1 and 3 had no successes. Thus, the 6×2 marginal table relating center to response, collapsed over treatment, contains zero counts. This marginal table is shown in the last two columns of Table 7.7. Infinite ML estimates occur for terms in loglinear or logit models containing the C-R association. An example is the logit model containing main effects for C and T in their effects on R. The likelihood function continually increases as the parameters for Centers 1 and 3 decrease toward $-\infty$; that is, as the logit decreases toward $-\infty$, so the fitted probability of success decreases toward 0 for those centers. Most software reports estimates that

Table 7.7 Clinical Trial Relating Treatment (T) to Response (R) for Five Centers (C), with T-R and C-R Marginal Tables

		Response		C-R Marginal	
Center	Treatment	Success	Failure	Success	Failure
1	Active Drug	0	5	0	14
	Placebo	0	9		
2	Active Drug	1	12	1	22
	Placebo	0	10		
3	Active Drug	0	7	0	12
	Placebo	0	5		
4	Active Drug	6	3	8	9
	Placebo	2	6		
5	Active Drug	5	9	7	21
	Placebo	2	12		
T-R	Active Drug	12	36		
Marginal	Placebo	4	42		

Source: Diane Connell, Sandoz Pharmaceuticals Corp.

are truly infinite as large numbers with large standard errors. For instance, when SAS (GENMOD) fits the logit model (setting the center estimate to be 0 for Center 5), the reported center estimates for Centers 1 and 3 are both about -26 with standard errors of about 200,000.

The counts in the 2×2 marginal table relating treatment to response, shown in the bottom panel of Table 7.7, are all positive. The empty cells in Table 7.7 affect the center estimates, but not the treatment estimates, for this logit model. For instance, if we add any positive constant to each cell, the fitting process converges, all center parameter estimates being finite; moreover, the treatment effects and goodness of fit are stable, as the addition of any such constant less than 0.001 yields an estimated log odds ratio equal to 1.55 for the treatment effect ($ASE = 0.70$) and a G^2 goodness-of-fit statistic equal to 0.50.

This treatment estimate also results from deleting Centers 1 and 3 from the analysis. When a center contains responses of only one type, it provides no information about the association between treatment and response. In fact, such tables also make no contribution to the Cochran–Mantel–Haenszel test (Section 3.2.1) or to the exact test of conditional independence between treatment and response (Section 3.3.1).

An alternative strategy in multi-center analyses combines centers of a similar type. Then, if each resulting partial table has responses with both outcomes, the inferences use all data. This, however, affects somewhat the interpretations and conclusions made from those inferences. For Table 7.7, perhaps Centers 1 and 3 are similar to Center 2, since the success rate is very low for that center. Combining these three centers and re-fitting the model to this table and the tables for the other two centers yields an estimated treatment effect of 1.56 ($ASE = 0.70$), with $G^2 = 0.56$.

7.4.3 Effect of Small Samples on X^2 and G^2

The true sampling distributions of goodness-of-fit statistics converge to chi-squared as the sample size $n \to \infty$, for a fixed number of cells N. The adequacy of the chi-squared approximation depends both on n and N. It tends to improve as n/N, the average number of observations per cell, increases.

The quality of the approximation has been studied carefully for the Pearson X^2 test of independence for two-way tables. Most guidelines refer to the fitted values. When $df > 1$, a minimum fitted value of about 1 is permissible as long as no more than about 20% of the cells have fitted values below 5. The size of permissible fitted values decreases as N increases. However, the chi-squared approximation can be poor for sparse tables containing both very small and very large fitted values. Unfortunately, a single rule cannot cover all cases.

The X^2 statistic tends to be valid with smaller samples and sparser tables than G^2. The distribution of G^2 is usually poorly approximated by chi-squared when n/N is less than 5. Depending on the sparseness, P-values based on referring G^2 to a chi-squared distribution can be too large or too small. When most fitted values are smaller than 0.5, treating G^2 as chi-squared gives a highly conservative test; that is, when H_0 is true, reported P-values tend to be much larger than true ones. When most fitted values are between about .5 and 5, G^2 tends to be too liberal; the reported P-value tends to be too small.

For fixed values of n and N, the chi-squared approximation is better for tests with smaller values of df. For instance, consider tests of conditional independence in $I \times J \times K$ tables. The statistic $G^2[(XZ, YZ) \mid (XY, XZ, YZ)]$, which has $df = (I-1)(J-1)$, is closer to chi-squared than $G^2(XZ, YZ)$, which has $df = K(I-1)(J-1)$. The ordinal test based on the homogeneous $L \times L$ association model (7.3.1) has $df = 1$, and behaves even better. It is difficult to provide general guidelines about how large n must be. The adequacy of model-comparison tests depends more on the two-way marginal totals than on cell counts. Cell counts can be small (which often happens when K is large) as long as most totals in the two-way marginal tables exceed about 5.

When cell counts are so small that chi-squared approximations may be inadequate, one could combine categories of variables to obtain larger counts. This is usually not advisable unless there is a natural way to combine them and little information loss in defining the variable more crudely. In any case, poor sparse-data performance of chi-squared tests is becoming less problematic because of the development of exact small-sample methods. This text has presented several exact tests, such as Fisher's exact test for 2×2 tables and analogous exact tests for $I \times J$ tables (Section 2.6), an exact test of conditional independence for $2 \times 2 \times K$ tables (Sec. 3.3), and exact tests for logistic regression (Section 5.7).

Exact analyses are now feasible due to recent improvements in computer power and sophistication of algorithms. For instance, the StatXact software conducts many exact inferences for two-way and three-way tables and LogXact handles exact inference in logistic regression. In principle, exact inferences about parameters or about goodness of fit exist for any loglinear or logit model, and software should soon enable us to conduct exact analyses for general situations.

7.5 SOME MODEL-FITTING DETAILS*

Most nonstatisticians will not have the interest or the prerequisite theoretical background to understand all the technical details underlying loglinear and logit model-fitting and inference. This is not a handicap for applying the methods. Thus, this text has omitted theoretical derivations in favor of emphasizing application and interpretation of models. With available software, one can fit models and analyze data sets without understanding how one maximizes likelihood functions, derives standard error formulas, proves large-sample chi-squared distributions for X^2 and G^2, and so forth. This section provides a brief introduction to these topics by giving a heuristic discussion of some loglinear and logit model-fitting details.

7.5.1 Sufficient Statistics and Likelihood Equations

Loglinear and logit models have fitted values and ML estimates of model parameters that depend on the data only through certain *sufficient statistics*. One can replace the data by these summary statistics without losing any information needed to fit the model.

For models with qualitative factors, the sufficient summaries of the data are marginal counts for terms in the model. For loglinear model (XZ, YZ), for instance, the sufficient statistics are the X-Z and Y-Z two-way marginal tables $\{n_{i+k}\}$ and $\{n_{+jk}\}$. Every three-way table having the same entries in these two marginal tables has the same fit.

The ML parameter estimates provide fitted values that satisfy the model and maximize the likelihood function. The estimates are solutions to a set of *likelihood equations*, which equate the fitted values to the sufficient statistics. For instance, model (XZ, YZ) has likelihood equations

$$\hat{\mu}_{i+k} = n_{i+k}, \qquad \hat{\mu}_{+jk} = n_{+jk}.$$

The fitted values satisfy the model but have the same X-Z and Y-Z marginal totals as the observed data. The formula for the fitted values for model (XZ, YZ) is $\hat{\mu}_{ijk} = n_{i+k}n_{+jk}/n_{++k}$. One can verify from this formula that the likelihood equations hold; that is, the X-Z and Y-Z marginals of $\{\hat{\mu}_{ijk}\}$ equal the corresponding observed totals.

Fitted values for loglinear models have similarities to the sample data, since certain marginal totals are the same for each. The fitted values are smoothed versions of the sample counts that match them in those margins, but which have associations and interactions satisfying the model. For instance, though the fitted values for model (XZ, YZ) match the data in the X-Z and Y-Z margins, they have X-Y conditional odds ratios equal to 1.0 at each level of Z. Analogous results apply to logistic regression models. For either model type, a parameter estimate is infinite when its sufficient statistic takes its maximum or minimum possible value. This happens, for instance, when a sufficient marginal total for a loglinear model equals zero.

Many loglinear and logit models do not have direct ML estimates. Unlike model (XZ, YZ), they require iterative algorithms for solving likelihood equations to produce fitted values and parameter estimates. The most popular iterative procedure is

the *Newton–Raphson* method. This method, described in Section 4.5.1, maximizes successive parabolic approximations for the log likelihood function. Calculations for doing this involve solving a system of linear equations at each step. The parameter values that maximize the parabolic functions serve as successive approximations for the ML estimates.

7.5.2 Asymptotic Chi-Squared Distributions

We next sketch the reason that X^2 and G^2 statistics for testing model fit have large-sample chi-squared distributions. Consider the sampling scheme by which N cell counts $\{n_i\}$ are Poisson variates with means $\{\mu_i\}$. For simplicity, we use a single subscript, though the table may have any dimension.

The X^2 statistic has form $X^2 = \sum e_i^2$, where $e_i = (n_i - \hat{\mu}_i)/\sqrt{\hat{\mu}_i}$ is the Pearson residual for cell i. This residual estimates $(n_i - \mu_i)/\sqrt{\mu_i}$, which has a large-sample standard normal distribution, since the Poisson distribution is approximately normal when its mean, which is also its variance, is large. Squaring standard normal variates produces chi-squared variates with $df = 1$. Adding N independent chi-squared variates with $df = 1$ yields a chi-squared variate with $df = N$. Thus, $\sum(n_i - \mu_i)^2/\mu_i$ has an approximate chi-squared distribution with df equal to the number of cells, N.

The df for X^2 do not equal N, however, because substituting $\{\hat{\mu}_i\}$ for $\{\mu_i\}$ in the Pearson residuals $\{e_i\}$ reduces their variance and yields correlated values. The df equal the rank of the covariance matrix of these residuals. This depends on the complexity of the model, equaling the number of cells minus the number of nonredundant model parameters.

Expanding G^2 in a Taylor series approximation, one obtains X^2 from the first two terms. Under the null hypothesis that the model holds, the higher-order terms in the expansion are negligible as the sample size increases. In other words, X^2 is then a quadratic approximation for G^2. The two statistics have the same large-sample chi-squared distribution as the sample size n increases. When the null hypothesis is false, X^2 and G^2 tend to increase as n increases, but their limiting *noncentral* chi-squared distributions can be quite different.

7.5.3 Comparing Nested Models

Many model-building procedures involve comparing the fits of two models, when one is a special case of the other. Our discussion here pertains to a GLM of any type for categorical data. For an arbitrary model, denoted by M, let $G^2(M)$ denote the value of G^2 for testing the fit of the model. In GLM terminology, this is the model's *deviance* (Section 4.5.3).

Consider two models, M_0 and M_1, such that M_0 is simpler than M_1. For instance, M_1 could be loglinear model (XY, XZ, YZ), and M_0 could be model (XZ, YZ). Model M_0 is *nested* within M_1, being a special case in which certain parameters equal zero. Since M_1 is more complex than M_0, it has a larger set of parameter values to search over in maximizing the likelihood function. Thus, the fit of M_1 is better, in the sense

that necessarily

$$G^2(M_1) \leq G^2(M_0).$$

A test comparing the models checks whether the more complex model M_1 gives a better fit than the simpler model M_0. The test of (H_0 : M_0 holds) against (H_a : M_1 holds) analyzes whether the extra terms in M_1 that are not in M_0 equal zero. Assuming that model M_1 holds, the likelihood-ratio approach for testing that M_0 holds uses test statistic

$$G^2(M_0 \mid M_1) = G^2(M_0) - G^2(M_1).$$

We used statistics of form $G^2(M_0 \mid M_1)$ for loglinear models in Sections 6.3.4, 7.1.5, 7.2.3, and 7.3.1 to test whether a model term equals zero. The simpler model M_0 omits the term, and the more complex model M_1 contains it. We also used this test in Sections 5.3.2 and 5.5.2 to compare nested logistic regression models.

Theory states that $G^2(M_0 \mid M_1)$ is a large-sample chi-squared statistic when the parameter spaces are fixed for M_0 and M_1 as n increases. The df value measures the difference between the number of parameters for the two models. This large-sample theory is not always appropriate in practice. An example is when M_0 is a logistic regression model with continuous predictors and M_1 is the saturated model, so the test refers to the goodness of fit of M_0. In that case, one usually observes data at additional levels of the predictors as the sample size increases. The saturated model has a separate parameter at each combination of predictor levels, so its number of parameters increases with the sample size. Thus, the parameter space is not fixed for M_1, and the df value comparing its size to the number of parameters for the simpler model is not fixed, invalidating the chi-squared theory. One can, however, compare the fits of two unsaturated logistic regression models. Their numbers of parameters and the difference df between them stays fixed as n increases. For a given sample size, the chi-squared approximation tends to be better for smaller values of df, such as when the two models differ by just one term.

Let $\{\hat{\mu}_{0i}\}$ and $\{\hat{\mu}_{1i}\}$ denote fitted values for models M_0 and M_1. One can show that the likelihood-ratio statistic for comparing the models also equals

$$G^2(M_0 \mid M_1) = 2\sum \hat{\mu}_{1i} \log\left(\frac{\hat{\mu}_{1i}}{\hat{\mu}_{0i}}\right). \tag{7.5.1}$$

This statistic has the form of the usual G^2 statistic, but with $\{\hat{\mu}_{1i}\}$ in place of the observed cell counts. In fact, $G^2(M_0)$ is the special case of $G^2(M_0 \mid M_1)$ with M_1 being the saturated model, in which case the fitted values $\{\hat{\mu}_{1i}\}$ for M_1 are simply the cell counts $\{n_i\}$. For the Pearson statistic X^2, the difference $X^2(M_0) - X^2(M_1)$ for nested models is not necessarily nonnegative. A more appropriate Pearson statistic for comparing nested models is

$$X^2(M_0 \mid M_1) = \sum \frac{(\hat{\mu}_{1i} - \hat{\mu}_{0i})^2}{\hat{\mu}_{0i}}. \tag{7.5.2}$$

The Pearson X^2 for testing the fit of a model has this form with M_1 as the saturated model.

These remarks relate to the results in Section 2.4.6 on the partitioning of chi-squared. The likelihood-ratio statistic for the simpler model partitions into

$$G^2(M_0) = G^2(M_1) + [G^2(M_0) - G^2(M_1)] = G^2(M_1) + G^2(M_0 \mid M_1),$$

a statistic for testing M_1 and a statistic for testing M_0 given that M_1 holds. By contrast, $X^2(M_0)$ does not partition exactly into $X^2(M_1) + X^2(M_0 \mid M_1)$.

7.5.4 Distribution of Parameter Estimators

Finally, we discuss standard errors for ML estimates of Poisson loglinear model parameters and binomial logit parameters. The standard errors are square roots of variances, which are the diagonal elements of the covariance matrix of the parameter estimators. The estimated covariance matrix is the inverse of a matrix called the *information matrix*, which is a by-product of the Newton–Raphson fitting procedure. The information matrix measures the curvature of the log likelihood function at the ML estimates. More highly curved log likelihood functions yield greater information about the parameter values; this results in smaller elements of the inverse of the information matrix and smaller standard errors.

We first illustrate this for loglinear models for Poisson counts. Let $\boldsymbol{\mu}$ denote a column vector of expected frequencies. Loglinear models for Poisson cell means have form

$$\log \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}. \tag{7.5.3}$$

The matrix $\mathbf{X}$, called a *model matrix* or *design matrix*, contains known constants. The column vector $\boldsymbol{\beta}$ contains the parameters. The log means are linearly related to the parameters, so the model is "loglinear." To illustrate, consider the independence model, $\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y$, for a 2×2 table. The model has three nonredundant parameters. For the constraints $\lambda_2^X = \lambda_2^Y = 0$, these are $\boldsymbol{\beta} = (\lambda, \lambda_1^X, \lambda_1^Y)$. Expression (7.5.3) is then

$$\begin{pmatrix} \log \mu_{11} \\ \log \mu_{12} \\ \log \mu_{21} \\ \log \mu_{22} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} \lambda \\ \lambda_1^X \\ \lambda_1^Y \end{pmatrix}$$

For instance, $\log \mu_{12} = \lambda + \lambda_1^X + \lambda_2^Y = \lambda + \lambda_1^X$, so the second row of $\mathbf{X}$ is $(1, 1, 0)$.

The model matrix $\mathbf{X}$ occurs in the expression for the covariance matrix of parameter estimators. As the sample size n increases, the ML estimator $\hat{\boldsymbol{\beta}}$ has distribution approaching the normal. The estimated covariance matrix of $\hat{\boldsymbol{\beta}}$ is

$$\hat{\text{C}}\text{ov}(\hat{\boldsymbol{\beta}}) = [\mathbf{X}'\mathbf{Diag}(\hat{\boldsymbol{\mu}})\mathbf{X}]^{-1}, \tag{7.5.4}$$

where the diagonal matrix has the fitted values on the main diagonal. The elements on the main diagonal of the covariance matrix are the large-sample variances, and their square roots are the standard errors. As the fitted values increase, the standard errors decrease.

Logistic regression models have a similar formula. Let $\boldsymbol{\pi}$ denote the vector of "success" probabilities at the various settings of the explanatory variables. The model has form

$$\text{logit}(\boldsymbol{\pi}) = \mathbf{X}\boldsymbol{\beta}.$$

Let $\hat{\boldsymbol{\pi}}$ denote the estimated probabilities for the model fit, with value $\hat{\pi}_i$ for the n_i observations at the ith setting of explanatory variables. The large-sample estimated covariance matrix of the ML estimator $\hat{\boldsymbol{\beta}}$ equals

$$\hat{\text{C}}\text{ov}(\hat{\boldsymbol{\beta}}) = [\mathbf{X}'\mathbf{DiagX}]^{-1}, \tag{7.5.5}$$

where **Diag** denotes a diagonal matrix having elements $n_i\hat{\pi}_i(1 - \hat{\pi}_i)$ on the main diagonal. The standard errors decrease as the sample size increases. This can happen by $\{n_i\}$ increasing or by more observations occurring at additional settings of the explanatory variables.

PROBLEMS

7.1. Draw the association graph for loglinear model (WXZ, WYZ). Which, if any, variables are conditionally independent in this model?

7.2. For a four-way table, are X and Y independent, given Z alone, for model (i) (WX, XZ, YZ, WZ), (ii) (WX, XZ, YZ, WY)?

7.3. Refer to Problem 6.3 with Table 3.1. Show the association graph for model (DV, PV), and fit the model. Using the fitted values, compute the estimated P-V odds ratios at the two levels of D, and compute the marginal P-V odds ratio. Compare the values. Are the collapsibility conditions satisfied?

7.4. Refer to the clinical trial in Problem 3.10 with Table 3.6.

a. Fit logit model (7.1.1). Using the fitted values, compute the estimated odds ratio between group and response (i) for each center, (ii) for the marginal table. Why do they differ?

b. Fit the simpler logit model deleting the center effects to this $2 \times 2 \times 3$ table. Using the fitted values, compute the estimated odds ratio between group and response (i) for each center, (ii) for the marginal table. When this model fits well, can we collapse over centers? Compare the fit to the model in (a). Is the simpler model adequate?

7.5. Refer to Problem 6.13 with Table 6.17.

a. Show that model (CE, CH, CL, EH, EL, HL) fits well. Show that model (CEH, CEL, CHL, EHL) also fits well but does not provide a significant improvement. Beginning with (CE, CH, CL, EH, EL, HL), show that backward elimination yields (CE, CL, EH, HL). Interpret its fit.