An alternative strategy to analyze bivariate and multivariate contingency tables is ASSOCIATION MODELS based on LOG-LINEAR MODELING. Log-linear models allow one to specify and analyze different and complex patterns of associations, whereas association coefficients give a summary measure.

Association coefficients measure the strength of the relation between two variables. Users oftentimes demand guidelines to interpret the strength, but general guidelines are impossible. The strength of an association depends on the sample's homogeneity, the reliability of the variables, and the type of relationship between the variables. In psychological experiments (very homogeneous sample, variables with high reliability, direct casual link), a coefficient of 0.4 may signal weak correlation, whereas the same value could be suspiciously high in a sociological survey (heterogeneous sample, lower reliability, indirect causal links). Hence, when fixing threshold values, it is important to consider these factors along with the results of previous studies.

—Johann Bacher

REFERENCES

Daniels, H. E. (1944). The relation between measures of correlation in the universe of sample permutations. *Biometrika, 35,* 129–135.

Healey, J. F. (1995). *Statistics: A tool for social research* (3rd ed.). Belmont, CA: Wadsworth.

Kendall, M. G. (1962). *Rank correlation methods* (3rd ed.). London: Griffin.

SPSS. (2002, August). Crosstabs. In *SPSS 11.0 statistical algorithms* [online]. Available: http://www.spss.com/tech/stat/algorithms/11.0/crosstabs.pdf

ASSOCIATION MODEL

Although the term ASSOCIATION is used broadly, *association model* has a specific meaning in the literature on CATEGORICAL DATA ANALYSIS. By *association model,* we refer to a class of statistical models that fit observed frequencies in a cross-classified table with the objective of measuring the strength of association between two or more ordered categorical variables. For a two-way table, the strength of association being measured is between the two categorical variables that comprise the cross-classified table. For three-way or higher-way tables, the strength of association being measured can be between any pair of ordered categorical variables that comprise the cross-classified table. Although some association models make use of the a priori ordering of the categories, other models do not begin with such an assumption and indeed reveal the ordering of the categories through estimation. The association model is a special case of a LOG-LINEAR MODEL or log-bilinear model.

Leo Goodman should be given credit for having developed association models. His 1979 paper, published in the *Journal of the American Statistical Association,* set the foundation for the field. This seminal paper was included along with other relevant papers in his 1984 book, *The Analysis of Cross-Classified Data Having Ordered Categories.* Here I first present the canonical case for a two-way table before discussing extensions for three-way and higher-way tables. I will also give three examples in sociology and demography to illustrate the usefulness of association models.

GENERAL SETUP FOR A TWO-WAY CROSS-CLASSIFIED TABLE

For the cell of the ith row and the jth column ($i = 1, \ldots, I$, and $j = 1, \ldots, J$) in a two-way table of R and C, let f_{ij} denote the observed frequency and F_{ij} the expected frequency under some model. Without loss of generality, a log-linear model for the table can be written as follows:

$$\log(F_{ij}) = \mu + \mu_i^R + \mu_j^C + \mu_{ij}^{RC}, \qquad (1)$$

where μ is the main effect, μ^R is the row effect, μ^C is the column effect, and μ^{RC} is the interaction effect on the logarithm of the expected frequency. All the parameters in equation (1) are subject to ANOVA-type normalization constraints (see Powers & Xie, 2000, pp. 108–110). It is common to leave μ^R and μ^C unconstrained and estimated nonparametrically. This practice is also called the "saturation" of the marginal distributions of the row and column variables. What is of special interest is μ^{RC}: At one extreme, μ^{RC} may all be zero, resulting in an independence model. At another extreme, μ^{RC} may be "saturated," taking $(I - 1)(J - 1)$ degrees of freedom, yielding exact predictions ($F_{ij} = f_{ij}$ for all i and j).

Typically, the researcher is interested in fitting models between the two extreme cases by altering specifications for μ^{RC}. It is easy to show that all ODDS RATIOS in a two-way table are functions of the interaction parameters (μ^{RC}). Let θ_{ij} denote a local log-odds ratio for a 2 × 2 subtable formed from four adjacent cells obtained from two adjacent row categories and two adjacent column categories:

$$\theta_{ij} = \log\{[F_{(i+1)(j+1)}F_{ij}]/[F_{(i+1)j}F_{i(j+1)}]\},$$
$$i = 1, \ldots, I-1, j = 1, \ldots, J-1.$$

Let us assume that the row and column variables are ordinal on some scales x and y. The scales may be observed or latent. A linear-by-linear association model is as follows:

$$\log(F_{ij}) = \mu + \mu_i^R + \mu_j^C + \beta x_i y_j, \qquad (2)$$

where β is the parameter measuring the association between the two scales x and y representing, respectively, the row and column variables. If the two scales x and y are directly observed or imputed from external sources, estimation of equation (2) is straightforward via MAXIMUM LIKELIHOOD ESTIMATION for log-linear models.

ASSOCIATION MODELS FOR A TWO-WAY TABLE

If we do not have extra information about the two scales x and y, we can either impose assumptions about the scales or estimate the scales internally. Different approaches give rise to different association models. Below, I review the most important ones.

Uniform Association

If the categories of the variables are correctly ordered, the researcher may make a simplifying assumption that the ordering positions form the scales (i.e., $x_i = i, y_j = j$). Let the practice be called *integer scoring*. The integer-scoring simplification results in the following uniform association model:

$$\log(F_{ij}) = \mu + \mu_i^R + \mu_j^C + \beta_{ij}. \qquad (3)$$

The researcher can estimate the model with actual data to see whether this assumption holds true.

Row Effect and Column Effect Models

Although the uniform association model is based on integer scoring for both the row and column variables, the researcher may wish to invoke integer scoring only for the row *or* the column variable. When integer scoring is used only for the column variable, the resulting model is called the *row effect model*. Conversely, when integer scoring is used only for the row variable, the resulting model is called the *column effect model*. Taking the row effect model as an example, we can derive the following model from equation (2):

$$\log(F_{ij}) = \mu + \mu_i^R + \mu_j^C + j\phi_i. \qquad (4)$$

This model is called the row effect model because the latent scores of the row variable ($\phi_i = \beta x_i$) are revealed by estimation after we apply integer scoring for the column variable. That is, ϕ_i is the "row effect" on the association between the row variable and the column variable. Note that the terms *row effect* and *column effect* here have different meanings than μ_i^R and μ_j^C, which are fitted to saturate the marginal distributions of the row and column variables.

Goodman's *RC* Model

The researcher can take a step further and treat both the row and column scores as unknown. Two of Goodman's (1979) association models are designed to estimate such latent scores. Goodman's Association Model I simplifies equation (1) to the following:

$$\log(F_{ij}) = \mu + \mu_i^R + \mu_j^C + j\phi_i + i\varphi_j, \qquad (5)$$

where ϕ_i and φ_j are, respectively, unknown row and column scores as in the row effect and column effect models. However, it is necessary to add three normalization constraints to uniquely identify the $(I + J)$ unknown parameters of ϕ_i and φ_j.

Goodman's Association Model I requires that both the row and column variables be correctly ordered a priori because integer scoring is used for both, as shown in equation (5). This requirement means that the model is not invariant to positional changes in the categories of the row and column variables. If the researcher has no knowledge that the categories are correctly ordered or in fact needs to determine the correct ordering of the categories, Model I is not appropriate. For this reason, Goodman's Association Model II has received the most attention. It is of the following form:

$$\log(F_{ij}) = \mu + \mu_i^R + \mu_j^C + \beta\phi_i\varphi_j, \qquad (6)$$

where β is the association parameter, and ϕ_i and φ_j are unknown scores to be estimated. Also, ϕ_i and φ_j are

Table 1 Comparison of Association Models

Model	μ^{RC}	DF_m	θ_{ij}
Uniform association	βij	1	β
Row effect	$j\phi_i$	$(I-1)$	$\phi_{i+1}-\phi_i$
Column effect	$i\phi_j$	$(J-1)$	$\phi_{j+1}-\phi_j$
Association Model I	$j\phi_i + i\phi_j$	$I+J-3$	$(\phi_{i+1}-\phi_i)+(\phi_{j+1}-\phi_j)$
Association Model II (*RC*)	$\beta\phi_i\phi_j$	$I+J-3$	$(\phi_{i+1}-\phi_i)(\phi_{j+1}-\phi_j)$

subject to four normalization constraints because each requires the normalization of both location and scale.

As equation (6) shows, the interaction component (μ^{RC}) of Goodman's Association Model II is in the form of multiplication of unknown parameters—log-bilinear specification. The model is also known as the *log-multiplicative model*, or simply the RC model. The RC model is very attractive because it allows the researcher to estimate unknown parameters even when the categories of the row and the column variables may not be correctly ordered. All that needs to be assumed is the existence of the ordinal scales. The model can reveal the orders through estimation.

Table 1 presents a summary comparison of the aforementioned association models. The second column displays the model specification for the interaction parameters (μ^{RC}). The number of degrees of freedom for each μ^{RC} specification is given in the third column (DF_m). If there are no other model parameters to be estimated, the degrees of freedom for a model are equal to $(I-1)(J-1) - DF_m$. The formula for calculating the local log-odds ratio is shown in the last column.

Goodman's Association Model II (RC model) can be easily extended to have multiple latent dimensions so that μ^{RC} of equation (1) is specified as

$$\mu_{ij}^{RC} = \sum \beta_m \phi_{im} \varphi_{jm}, \qquad (7)$$

where the summation sign is with respect to all possible m dimensions, and the parameters are subject to necessary normalization constraints. Such models are called $RC(M)$ models. See Goodman (1986) for details.

ASSOCIATION MODELS FOR THREE-WAY AND HIGHER-WAY TABLES

Below, I mainly discuss the case of a three-way table. Generalizations to a higher-way table can be easily made. Let R denote row, C denote column, and L denote layer, with their categories indexed respectively by i ($i = 1, \ldots, I$), j ($j = 1, \ldots, J$), and k ($k = 1, \ldots, K$). In a common research setup, the researcher is interested in understanding how the two-way association between R and C depends on levels of L. For example, in a trend analysis, L may represent different years or cohorts. In a comparative study, L may represent different nations or groups. Thus, research attention typically focuses on the association pattern between R and C and its variation across layers.

Let F_{ijk} denote the expected frequency in the ith row, the jth column, and the kth layer. The saturated log-linear model can be written as follows:

$$\log(F_{ijk}) = \mu + \mu_i^R + \mu_j^C + \mu_k^L + \mu_{ij}^{RC} + \mu_{ik}^{RL} + \mu_{jk}^{CL} + \mu_{ijk}^{RCL}. \qquad (8)$$

In a typical research setting, interest centers on the variation of the RC association across layers. Thus, the baseline (for the null hypothesis) is the following conditional independence model:

$$\log(F_{ijk}) = \mu + \mu_i^R + \mu_j^C + \mu_k^L + \mu_{ik}^{RL} + \mu_{jk}^{CL}. \qquad (9)$$

That is, the researcher needs to specify and estimate μ^{RC} and μ^{RCL} to understand the layer-specific RC association.

There are two broad approaches to extending association models to three-way or higher-way tables. The first is to specify an association model for the typical association pattern between R and C and then estimate parameters that are specific to layers or test whether they are invariant across layers (Clogg, 1982a). The general case of the approach is to specify μ^{RC} and μ^{RCL} in terms of the RC model so as to change equation (8) to the following:

$$\log(F_{ijk}) = \mu + \mu_i^R + \mu_j^C + \mu_k^L + \mu_{ik}^{RL} + \mu_{jk}^{CL} + \beta_k \phi_{ik} \varphi_{jk}. \qquad (10)$$

That is, the β, ϕ, and φ parameters can be layer specific or layer invariant, subject to model specification and statistical tests. The researcher may also wish to test special cases (i.e., the uniform association, column effect, and row effect models) where ϕ and/or φ parameters are inserted as integer scores rather than estimated.

The second approach, called the *log-multiplicative layer-effect model,* or the "unidiff model," is to allow a flexible specification for the typical association pattern between R and C and then to constrain its cross-layer variation to be log-multiplicative (Xie, 1992). That is, we give a flexible specification for μ^{RC} but constrain μ^{RCL} so that equation (8) becomes the following:

$$\log(F_{ijk}) = \mu + \mu_i^R + \mu_j^C + \mu_k^L + \mu_{ik}^{RL} + \mu_{jk}^{CL} + \phi_k \psi_{ij}. \qquad (11)$$

With the second approach, the RC association is not constrained to follow a particular model and indeed can be saturated with $(I - 1)(J - 1)$ dummy variables. In a special case in which the typical association pattern between R and C is the RC model, the two approaches coincide, resulting in the three-way RCL log-multiplicative model. Powers and Xie (2000, pp. 140–145) provide a more detailed discussion of the variations and the practical implications of the second approach. It should be noted that the two approaches are both special cases of a general framework proposed by Goodman (1986) and extended in Goodman and Hout (1998).

APPLICATIONS

Association models have been used widely in sociological research. Below, I give three concrete examples. The first example is one of scaling. See Clogg (1982b) for a detailed illustration of this example. Clogg aimed to scale an ordinal variable that measures attitude on abortion. The variable was constructed from a Guttman scale, and the cases that did not conform to the scale response patterns were grouped into a separate category, "error responses." As is usually the case, scaling required an "instrument." In this case, Clogg used a measure of attitude on premarital sex that was collected in the same survey. The underlying assumption was that the scale of the attitude on abortion could be revealed from its association with the attitude on premarital sex. Clogg used the log-multiplicative model to estimate the scores associated with the different categories of the two variables. Note that the log-multiplicative RC model assumes that the categories are ordinal but not necessarily correctly ordered. So, estimation reveals the scale as well as the ordering. Through estimation, Clogg showed that the distances between the adjacent categories were unequal and that those who gave "error responses" were in the middle in terms of their attitudes on abortion.

The second example is the application of the log-multiplicative layer-effect model to the cross-national study of intergenerational mobility (Xie, 1992). The basic idea is to force cross-national differences to be summarized by layer-specific parameters [i.e., ϕ_k of equation (11)] while allowing and testing different parameterizations of the two-way association between father's occupation and son's occupation (i.e., ψ_{ij}). The ϕ_k parameters are then taken to represent the social openness or closure of different societies.

The third example, which involves the study of human fertility, is nonconventional in the sense that the basic setup is not log-linear but log-rate. The data structure consists of a table of frequencies (births) cross-classified by age and country and a corresponding table of associated exposures (women-years). The ratio between the two yields the country-specific and age-specific fertility rates. The objective of statistical modeling is to parsimoniously characterize the age patterns of fertility in terms of fertility level and fertility control for each country. In conventional demography, this is handled using Coale and Trussell's *Mm* method. Xie and Pimentel (1992) show that this method is equivalent to the log-multiplicative layer-effect model, with births as the dependent variable and exposure as an "offset." Thus, the *M* and *m* parameters of Coale and Trussell's method can be estimated statistically along with other unknown parameters in the model.

ESTIMATION

Estimation is straightforward with the uniform, row effect, column effect, and Goodman's Association Model I models. The user can use any of the computer programs that estimate a log-linear model. What is complicated is when the RC interaction takes the form of the product of unknown parameters—the log-multiplicative or log-bilinear specification. In this case, a reiterative estimation procedure is required. The basic idea is to alternately treat one set of unknown parameters as known while estimating the other and to continue the iteration process until both are stabilized. Special computer programs, such as ASSOC and

CDAS, have been written to estimate many of the association models. User-written subroutines in GLIM and STATA are available from individual researchers. For any serious user of association models, I also recommend Lem, a program that can estimate different forms of the log-multiplicative model while retaining flexibility. See my Web site www.yuxie.com for updated information on computer subroutines and special programs.

—Yu Xie

REFERENCES

Clogg, C. C. (1982a). Some models for the analysis of association in multiway cross-classifications having ordered categories. *Journal of the American Statistical Association, 77,* 803–815.

Clogg, C. C. (1982b). Using association models in sociological research: Some examples. *American Journal of Sociology, 88,* 114–134.

Goodman, L. A. (1979). Simple models for the analysis of association in cross-classifications having ordered categories. *Journal of the American Statistical Association, 74,* 537–552.

Goodman, L. A. (1984). *The analysis of cross-classified data having ordered categories.* Cambridge, MA: Harvard University Press.

Goodman, L. A. (1986). Some useful extensions of the usual correspondence analysis approach and the usual log-linear models approach in the analysis of contingency tables. *International Statistical Review, 54,* 243–309.

Goodman, L. A., & Hout, M. (1998). Understanding the Goodman-Hout approach to the analysis of differences in association and some related comments. In Adrian E. Raftery (Ed.), *Sociological methodology* (pp. 249–261). Washington, DC: American Sociological Association.

Powers, D. A., & Xie, Y. (2000). *Statistical methods for categorical data analysis.* New York: Academic Press.

Xie, Y. (1992). The log-multiplicative layer effect model for comparing mobility tables. *American Sociological Review, 57,* 380–395.

Xie, Y., & Pimentel, E. E. (1992). Age patterns of marital fertility: Revising the Coale-Trussell method. *Journal of the American Statistical Association, 87,* 977–984.

ASSUMPTIONS

Assumptions are ubiquitous in social science. In theoretical work, assumptions are the starting axioms and postulates that yield testable implications spanning broad domains. In empirical work, statistical procedures typically embed a variety of assumptions, for example, concerning measurement properties of the variables and the distributional form and operation of unobservables (such as HOMOSKEDASTICITY or NORMAL DISTRIBUTION of the ERROR). Assumptions in empirical work are discussed in the entries for particular procedures (e.g., ORDINARY LEAST SQUARES); here we focus on assumptions in theories.

The purpose of a scientific THEORY is to yield testable implications concerning the relationships between observable phenomena. The heart of the theory is its set of assumptions. The assumptions embody what Popper (1963) calls "guesses" about nature—guesses to be tested, following Newton's vision, by testing their logical implications. An essential feature of the assumption set is internal logical consistency. In addition, three desirable properties of a theory are as follows: (a) that its assumption set be as short as possible, (b) that its observable implications be as many and varied as possible, and (c) that its observable implications include phenomena or relationships not yet observed, that is, novel predictions.

Thus, a theory has a two-part structure: a small part containing the assumptions and a large and ever-growing part containing the implications. Figure 1 provides a visualization of the structure of a theory.

A theory can satisfy all three properties above and yet be false. That is, one can invent an imaginary world, set up a parsimonious set of postulates about its operation, deduce a wide variety of empirical consequences, and yet learn through empirical test that no known world operates in conformity with the implications derived from the postulated properties of the imaginary world. That is why empirical analysis is necessary, or, put differently, why theoretical analysis alone does not suffice for the accumulation of reliable knowledge.

A note on terminology: *Assumption* is a general term, used, as noted earlier, in both theoretical and empirical work. Sharper terms sometimes used in theoretical work include *axiom,* which carries the connotation of *self-evident,* and *postulate,* which does not, and is therefore more faithful to an enterprise marked by guesses and bound for discovery. Other terms include the serviceable *premise,* the colorful *starting principle,* and the dangerous *hypothesis,* which is used not only as a postulate (as in the HYPOTHETICO-DEDUCTIVE METHOD invented by Newton) but also as an observable proposition to be tested.

Where do assumptions come from? Typically, the frameworks for analyzing topical domains include a variety of relations and FUNCTIONS, some of which may prove to be fruitful assumptions. In general,