Yu Xie, Institute for Social Research, 426 Thompson Street, University of Michigan, Ann Arbor, MI 48106. Email: yuxie@umich.edu. Tel: (734)936-0039. Fax: (734)998-7415.

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 a three-way or higher-way table, the strength of association being measured can be between any pair of ordered categorical variables that comprise the cross-classified table. While 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. Association model is a special case of LOGLINEAR MODEL or log-bilinear model.

Leo Goodman should be given the credit for having developed association models. His 1979 paper published in the *Journal of 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 multi-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,…I, and j = 1,…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 loglinear model for the table can be written as:

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

where μ is the "main effect," μ^R the "row effect," μ^C the "column effect," and μ^{RC} 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 and Xie

2000, pp.108-110). It is common to leave μ^R and μ^R unconstrained and estimated non-

parametrically. This practice is also called the "saturation" of the marginal distributions

of the row and column variables. What is of special interest is to learn about 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 2x2 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,…I-1, j=1,…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:

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

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 LOGLINEAR 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 me call the practice "integer-scoring." The integer-scoring

simplification results in the uniform association model:

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

The researcher can estimate the model with actual data to see whether or not this

assumption holds true.

Row-Effect and Column-Effect Models. While the uniform association model is

based on integer-scoring for both the row and column variables, the researcher may wish

to invoke it for only 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):

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

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 wish to 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:

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

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 in order 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 are correctly ordered *a priori*, since integer-scoring is used for both, 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 form:

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

where β is the association parameter, and ϕ_i and φ_j are unknown scores to be estimated. ϕ_i and φ_j are subject to four normalization constraints, since 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 is equal to $(I-1)(J-1)-DF_m$. The formula for calculating the local log-odds-ratio is shown in the last column.

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	$i\varphi_j$	$(J-1)$	$\varphi_{j+1}-\varphi_j$
Association Model I	$j\phi_i + i\varphi_j$	$I+J-3$	$(\phi_{i+1}-\phi_i)+(\varphi_{j+1}-\varphi_j)$
Association Model II (RC)	$\beta\phi_i \varphi_j$	$I+J-3$	$(\phi_{i+1}-\phi_i)(\varphi_{j+1}-\varphi_j)$

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

$$\mu_{ij}^{RC} = \Sigma \, \beta_m \phi_{im} \, \varphi_{jm}, \tag{7}$$

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.

<u>Association Models for Three-Way and Higher-Way Tables</u>

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,…I) , j (j=1,…J), and k (k=1,…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 loglinear model can be written as:

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

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:

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

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

There are two broad approaches to extending association models that were initially developed for a two-way table to a three-way or higher-way table. 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:

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

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 "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:

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

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 where 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. Power

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." To scale the variable, it was necessary to have 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 non-conventional in the sense that the basic setup is not loglinear 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 I models. The user can use any of the computer programs that estimate a LOGLINEAR 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 website www.yuxie.com for updated information on computer

subroutines and special programs.

References

Clogg, Clifford 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, Clifford C. 1982b. "Using Association Models in Sociological Research: Some Examples." *American Journal of Sociology* 88:114-134.

Goodman, Leo 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, Leo 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, Leo A. and Michael Hout. 1998. "Understanding the Goodman-Hout Approach to the Analysis of Differences in Association and Some Related Comments." Pp. 249-261 in *Sociological Methodology* 1998, edited by Adrian E. Raftery. Washington, DC: American Sociological Association.

Powers, Daniel A. and Yu Xie. 2000. *Statistical Methods for Categorical Data Analysis*. Academic Press.

Xie, Yu. 1992. "The Log-Multiplicative Layer Effect Model for Comparing Mobility Tables." *American Sociological Review* 57:380-395.