This article discusses how several hypotheses about change in discrete variables can be tested on data obtained in a longitudinal study. A first class of hypotheses pertain to the invariance of certain characteristics of marginal distributions. A second class of hypotheses derive from assumptions about the causal relations between the variables. In this article, the authors show how all these hypotheses can be tested by means of a generalization of log-linear modeling developed by Lang and Agresti. By means of the same approach, it is also possible to test conjunctions of several hypotheses from both classes.

Analyzing Change in Categorical Variables by Generalized Log-Linear Models

MARCEL A. CROON
WICHER BERGSMA
JACQUES A. HAGENAARS
Tilburg University

1. INTRODUCTION

In a longitudinal panel study, several kinds of hypotheses about change in the variables can be tested. A first type of hypothesis pertains to changes over time in certain aspects of the marginal or conditional distributions of repeatedly observed random variables. This type of hypothesis will be referred to as the homogeneity hypothesis. Testing homogeneity hypotheses does not always require a longitudinal panel design, since many hypotheses about marginal or conditional distributions can also be tested on data from independent samples in a cross-sectional study. Testing these hypotheses in a longitudinal study requires, however, a different approach that takes into account the fact that the marginal distributions are not observed in independent samples but are based on the same sample and, hence, are not statistically independent.

A second class of hypotheses pertain to estimating and testing causal path models for repeatedly observed random variables. In formulating and testing models of this kind, the temporal order between

SOCIOLOGICAL METHODS & RESEARCH, Vol. 29 No. 2, November 2000 195-229
© 2000 Sage Publications, Inc.

different variables has to be taken into account. Moreover, applying causal path models is only possible if some of the variables can be seen as independent variables that have causal effects on other dependent variables.[1]

Although all these hypotheses can be formulated for continuous and for discrete variables, in this article only the case of discrete categorical variables will be considered. Several authors (Duncan 1980, 1981; Clogg, Eliason, and Grego 1990; Hagenaars 1992) have formulated various homogeneity hypotheses that can be studied in longitudinal research with discrete variables.

As an example, let X_1, X_2, Y_1, and Y_2 represent the scores on two variables X and Y measured at two time points, and let Z be a time-invariant explanatory variable that is measured at the first time period. Their joint probability distribution will be denoted by $p(Z, X_1, Y_1, X_2, Y_2)$. To represent marginal distributions, the variables over which the joint distribution is marginalized are replaced by a + symbol. So, $p(+, X_1, +, X_2, +)$ represents the marginal distribution of (X_1, X_2), which is obtained by integrating or summing the original joint distribution over Z, Y_1, and Y_2.

The homogeneity hypotheses discussed by the authors mentioned above can be further classified into two broad categories. First, some homogeneity hypotheses pertain to questions about the invariance of various marginal distributions. The hypothesis of whether there is net change in variable X can be stated in terms of a comparison of the two marginals $p(+, X_1, +, +, +)$ and $p(+, +, +, X_2, +)$. Similarly, the hypothesis that the bivariate distribution of (X, Y) remains invariant over time can be investigated by comparing the marginals $p(+, X_1, Y_1, +, +)$ and $p(+, +, +, X_2, Y_2)$. Moreover, all hypotheses of this kind can be specified separately for fixed values of the time-invariant explanatory variable Z. For instance, a comparison of the marginals $p(z, X_1, Y_1, +, +)$ and $p(z, +, +, X_2, Y_2)$ for specific values of z may indicate for which subpopulations invariance of the distribution of (X, Y) holds and for which subpopulations it does not. Second, some homogeneity hypotheses pertain to certain specified aspects of the marginal distributions of the random variables involved. Instead of asking whether the bivariate marginal distribution of (X, Y) is invariant over time, one can ask whether the association between the variables remains constant. For discrete variables, problems of association are commonly formulated in terms of odds ratios, or coefficients (such as Goodman and Kruskal's gamma or Kendall's tau) derived from these ratios. In an attempt to assess the direction of causal influence, Duncan (1981) extensively discussed the way in which equality of the cross-lagged association could be tested. If the association between X_1 and Y_2 is stronger than the association between X_2 and Y_1, the hypothesis that X causes Y is much more reasonable than the hypothesis that Y causes X. If more than two measurement periods are involved in the panel study, Hagenaars (1992) noted that still more detailed questions can be asked. If the distribution of Y changes over time, is the change from time period 1 to time period 2 the same as the change from time period 2 to time period 3?

A third class of homogeneity hypotheses that will be considered in this article pertain to the invariance of certain conditional distributions. Suppose that X is in some sense an independent variable that causally influences the dependent variable Y. To investigate whether changes in X cause changes in Y, it might be reasonable to assume that the conditional distribution of Y given X does not change over time. This invariance can be investigated by testing whether

$$\frac{p(+,x,y,+,+)}{p(+,x,+,+,+)} = \frac{p(+,+,+,x,y)}{p(+,+,+,x,+)}$$

for all x and y. In a more detailed analysis, one can test whether this invariance holds for different values of the explanatory variable Z. Another example of a situation in which invariance of conditional distributions may be an interesting hypothesis to test is a longitudinal study in which the same variable X is measured at three time periods. In such a study, it may be relevant to test whether the conditional distribution of X_2 given X_1 is the same as the conditional distribution of X_3 given X_2. If this hypothesis of invariance is sustained, the transitions between values of the same variable are governed by a stationary first-order Markov process.

As stated earlier, panel studies not only allow the study of homogeneity hypotheses but also make it possible to investigate hypotheses about the causal relations between repeatedly measured variables. As an example, consider a situation in which a time-invariant explanatory variable Z is measured at the first time period, and in which an independent variable X and a dependent variable Y are measured at two different time periods. For these data, one might be interested in estimating the effects of Z and X on Y at both time periods. A possible model for the causal influences between the five observed variables is depicted in Figure 1.

Note that in this article, the graphical representations of causal models should be read as path diagrams in the sense of Goodman (1973). In such path diagrams, a directed arrow from variable A to variable B represents the direct effect variable A is assumed to have on variable B. These diagrams should not be interpreted as graphical models in the sense of Whittaker (1990) or Lauritzen (1996), where "graphical" actually means "graph-theoretical."

The model shown in Figure 1 makes the following assumptions about the causal relationships between the five variables involved:

1. Z and X_1, whose association is not explained by the postulated model, have a direct effect on Y_1.
2. X_2 is only influenced by X_1, not by Z and Y_1.
3. Y_2 is influenced by Z and X_2, but also by Y_1.

If all variables were continuous, this causal path model could be tested by means of regression analyses for the independent variables Y_1, X_2, and Y_2. For discrete dependent variables, Goodman (1973) proposed a modified path analysis approach in which a logit equation instead of a regression equation is specified and tested for each dependent variable.

In the analysis of discrete variables observed in a longitudinal design, it may also be interesting to study models that combine homogeneity hypotheses and hypotheses about the causal relations between the variables involved. For instance, for the causal model represented by Figure 1, one could ask whether, in addition to the restrictions implied by the causal model, the conditional distributions of Y given (X, Z) are the same at the two measurement points. If one succeeds in developing a causal model in which the number of parameters is further reduced by making some homogeneity assumptions, the final result is a more parsimonious model with a smaller number of parameters to interpret. Combining causal models with homogeneity assumptions would also yield substantively more interesting or informative results, since a joined test of both types of models puts more constraints on the data.

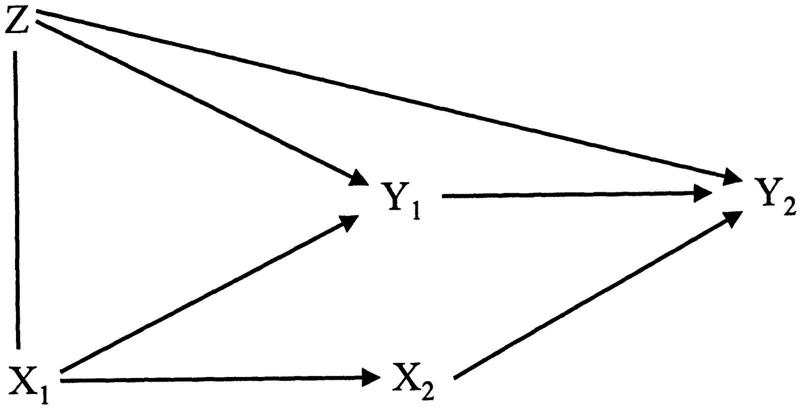

Figure 1: Causal Model 1

This article will discuss how for discrete data homogeneity assumptions can be tested, either on their own or in combination with assumptions that pertain to the causal order between the variables. In general, tests of the homogeneity hypotheses discussed above are not possible by means of a log-linear analysis, since they involve constraints on the expected frequencies themselves rather than constraints on log-linear parameters. However, they can be tested by means of a generalization of log-linear modeling for categorical data as proposed by Lang and Agresti (1994).

Interestingly enough, all the hypotheses considered in this article could also be tested by means of the generalized least squares procedure proposed by Grizzle, Starmer, and Koch (1969). However, their procedure often breaks down for sparse tables (Agresti 1990:462). Moreover, it does not yield estimates of the expected frequencies, so that if a particular model fits the data poorly, no residuals are available. Because an analysis of residuals may indicate how to modify the model, an estimation procedure that yields estimates of the expected frequencies and their standard errors is to be preferred. The maximum likelihood procedure proposed by Lang and Agresti (1994) is such a procedure. Moreover, this maximum likelihood procedure is less vulnerable (but certainly not completely immune) to zeros in the observed contingency table.

The structure of the remainder of this article is as follows. In section 2, the principles of generalized log-linear modeling as proposed by Lang and Agresti (1994) and further extended by Lang (1996a, 1996b) will be discussed. In sections 3 and 4, it will be shown how homogeneity assumptions and assumptions following from modified-path models can be tested by means of generalized log-linear analysis. Section 5 will discuss how homogeneity hypotheses and hypotheses about the causal relations between the variables can be tested simultaneously. Section 6 contains an application of a generalized log-linear model to data obtained in a two-wave panel study. Although this limited example does not show all the possibilities of generalized log-linear modeling in the analysis of discrete longitudinal data, it illustrates some of its flexibility. The article ends with a short discussion.

2. GENERALIZED LOG-LINEAR MODELING

Suppose that in a longitudinal study, K discrete variables are measured on a sample of N subjects. Some of these measurements may refer to the same variable measured at different time periods, whereas other measurements may refer to variables that are measured only once (and hence considered as time invariant). Letting m_k be the number of response categories for measurement k, the number of different response patterns is equal to $M = \Pi_{k=1}^{K} m_k$. A particular response pattern will be denoted by $i = (i_1, \ldots, i_K)$. Let $\pi = (\pi_1, \ldots, \pi_M)$, where π_i is the probability that a randomly selected subject will have response pattern i. The vector $f = (f_1, \ldots, f_M)$ contains the observed frequencies. In the remainder of this article, it is assumed that f follows a multinomial distribution with probabilities π and fixed sample size N. The expected frequencies of the response patterns with respect to a particular model will be denoted by $\mu = (\mu_1, \ldots, \mu_M)$. They should satisfy the constraint $\Sigma_k \mu_k = N$. It will be assumed that $\mu > 0$, that is, that all expected frequencies are strictly positive.

Log-linear models impose a linear structure on $\log\mu = (\log\mu_1, \ldots, \log\mu_M)$; that is, it is assumed that for some specified design matrix X, one may write

$$\log\mu = X\beta \tag{1}$$

for some vector β of unknown parameters. The matrix X can always be chosen so that it is of full column rank. Let U be a full column rank matrix such that the vector space spanned by the columns of U is the orthocomplement of the vector space spanned by the columns of X. One then has $w = Xa$ for some a if and only if $w'U = 0$. Hence, it follows that $X'U = 0$. Model (1) can then equivalently be specified by the constraints

$$U'\log\mu = 0.$$

Log-linear models can thus be formulated in two equivalent ways. In the first kind of formulation, the natural logarithms of the expected frequencies are written as functions of the unknown log-linear parameters. In the second kind of formulation, the same log-linear model is specified by means of the constraints it imposes on the expected frequencies. As a simple specific example, consider the independence model in a 2×2 table. First, it can be written log linearly in the following way:

$$\begin{pmatrix} \mu_{11} \\ \mu_{12} \\ \mu_{21} \\ \mu_{22} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}.$$

But, equivalently, the same log-linear model is completely specified by the single constraint it imposes on the expected frequencies:

$$\log\mu_{11} - \log\mu_{12} - \log\mu_{21} + \log\mu_{22} = 0.$$

In this example, one has

$$U' = (1 \quad -1 \quad -1 \quad 1).$$

Taking antilogarithms, the constraint on the $\log\mu$ terms is equivalent to the well-known definition of independence in a 2 × 2 table in terms of an odds ratio:

$$\frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} = 1.$$

The generalization of log-linear modeling considered by Lang and Agresti (1994) imposes the following linear structure on the expected frequencies:

$$C \log A\mu = X\beta, \qquad (2)$$

in which it is assumed that all elements of the vector $A\mu$ are strictly positive. If the columns of matrix U again span the null-space of X, model (2) is equivalent to

$$U'C \log A\mu = 0. \qquad (3)$$

Generalized log-linear analysis makes it possible to define contrasts between linear combinations of expected frequencies, instead of contrasts between single expected frequencies themselves.

In some applications, the matrix X may be a null matrix; then, matrix U is an identity matrix of appropriate dimensions and the constraints on the expected frequencies can be stated as

$$C \log A\mu = 0.$$

Lang and Agresti (1994) discussed how to obtain the maximum likelihood estimates of the unknown expected frequencies under the constraints given by equation (3) and by the single sampling constraint $\sum\mu = N$. The maximum likelihood estimates determine a stationary point of the augmented log-likelihood function

$$f'\log\mu - \lambda' U' C \log A\mu - \tau(1'\mu - N)$$

with respect to the unknown expected frequencies μ and the unknown Lagrange multipliers in λ and τ. (The row vector $1'$ consists of the appropriate number of 1s to represent the multinomial sampling constraint.) Because this optimization problem has in general no closedform solution, a Newton-Raphson iterative scheme is implemented to obtain the maximum likelihood estimates. Lang and Agresti (1994) also discussed the asymptotic behavior of the estimators and proposed some goodness-of-fit tests. More technical details on their algorithm and the statistical properties of the estimators can be found in their article. Here, it suffices to state that, at least for nonsparse data matrices, the goodness of fit of a particular model can be tested against the general multinomial alternative by means of an asymptotic log-likelihood ratio test. The number of the degrees of freedom for this test is equal to the number of constraints imposed by the model on the expected frequencies. Moreover, if a nonsaturated model is nested within another model, the more restricted model can be tested against the less restricted one by means of a conditional log-likelihood ratio test. The number of the degrees of freedom for this conditional test is equal to the difference between the degrees of freedom for the unconditional tests for each model against the general saturated model. Lang and Agresti (1994) also derived asymptotically valid expressions for the standard errors of the estimated values of the expected frequencies μ, the model parameters β, and the residuals $f - \mu$.

More recent work by Lang and others has shown that generalized log-linear analysis has many potential applications in the field of multivariate analysis of discrete variables. Lang (1996a) further extended generalized log-linear modeling for the simultaneous analysis of data obtained from different populations. Lang (1996b) studied the conditions under which the goodness-of-fit statistic can be partitioned according to two subhypotheses that are tested simultaneously in a generalized log-linear analysis. In section 6, these conditions will be discussed in the context of the specific application of the generalized log-linear model. Lang and Eliason (1997) used the generalized log-linear model in the analysis of social mobility tables. A still further generalization of log-linear modeling is described in Bergsma (1997), who discussed how to test invariance of various association coefficients over time.

In the next section, it will be shown how various homogeneity assumptions can be recast in terms of the Lang-Agresti generalization of log-linear modeling.

3. SPECIFYING HOMOGENEITY ASSUMPTIONS

3.1. HOMOGENEITY ASSUMPTIONS FOR MARGINAL DISTRIBUTIONS

A first class of research hypotheses that can be studied in a longitudinal design pertain to the question of whether the marginal distributions of a set of repeatedly measured variables change over time. Questions of this kind often arise within the context of comparative research in which the rate of change in some characteristic is studied in different groups or subpopulations. If there is change over time in the distribution of the same variable in all subpopulations, a natural question to ask is whether all groups show the same amount of change. By means of a limited number of more specific examples, it is shown below how research hypotheses of the kind considered above can be tested by means of a generalized log-linear analysis.

Example 1

Suppose that the same discrete variable X with m response categories is measured at two time periods. The expected frequencies can be written as a vector

$$\mu = (\mu_{11}, \mu_{12}, \mu_{13}, \dots, \mu_{1m}, \mu_{21}, \dots, \mu_{m-1,m}, \mu_{mm})$$

in which μ_{ij} is the expected frequency associated with the joint event $(X_1 = i, X_2 = j)$. (Note that in this vectorized notation, the last subscripts change the fastest. This convention will be used throughout this article.)

The assumption that the marginal distribution of X did not change over time implies that

$$\sum_j \mu_{ij} = \sum_j \mu_{ji}$$

holds for all i. Thus, the ith row sum of the contingency table should be equal to its ith column sum. Using the shorthand notation with the +

sign indicating variables over which the pattern probabilities have been summed, one can write the previous set of constraints as

$$\mu_{i+} = \mu_{+i}. \tag{4}$$

Of course, one could eliminate the common diagonal element μ_{ii} from both sums to obtain as constraints

$$\sum_{j \neq i} \mu_{ij} = \sum_{j \neq i} \mu_{ji}.$$

In the remainder of this article, the shorthand notation (implying summation over the entire range of scores for the relevant variables) will be used.

Equation (4) can also be stated in terms of a contrast between the logarithm of two sums of response probabilities:

$$\log \mu_{i+} - \log \mu_{+i} = 0$$

for all i. Because the sum of all expected frequency is equal to the number of observations, only $m - 1$ such contrasts need to be defined.

As a specific example, consider the test of marginal homogeneity in a 3 × 3 table. To apply the Lang-Agresti algorithm, the matrices A and C should then be defined as follows:

$$A = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \end{pmatrix}$$

$$C = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix}.$$

In this application, the design matrix X is a null matrix of order 2×1. Note also that because of linear dependence, the contrast based on the third row and column sum need not be considered in the construction of the matrices A and C.

Example 2

Suppose the same polytomous variable X is measured twice in different subpopulations or groups. Let G be the number of different groups, with a particular group denoted by g. The expected frequencies are represented by μ_{gij}. The hypothesis of overall univariate homogeneity of X can be tested by imposing contrasts of the following type:

$$\log\mu_{+i+} - \log\mu_{++i} = 0.$$

The hypothesis of univariate homogeneity in a particular subpopulation g can be tested by imposing constraints such as

$$\log\mu_{gi+} - \log\mu_{g+i} = 0.$$

If the hypothesis of univariate homogeneity has to be rejected in each subpopulation, it may be interesting to test whether each subpopulation changed to the same extent. The hypothesis of equal change may be tested by imposing contrasts of the following type on the expected frequencies for two different groups g and h:

$$\log\mu_{gi+} - \log\mu_{g+i} - \log\mu_{hi+} + \log\mu_{h+i} = 0.$$

Example 3

Assume that a polytomous variable X is measured at three time periods, with the expected frequencies given by μ_{ijk}. Univariate homogeneity is satisfied if

$$\mu_{i++} = \mu_{+i+} = \mu_{++i}$$

holds. These constraints can easily be imposed in a generalized loglinear model by defining two contrasts,

$$\log\mu_{i++} - \log\mu_{+i+} = 0$$

$$\log\mu_{+i+} - \log\mu_{++i} = 0,$$

for each $i : 1 \leq i \leq m - 1$.

However, in situations where the same variable is measured three or more times, other interesting homogeneity hypotheses may be considered. One of them is the hypothesis that the bivariate distribution of two consecutive measurements does not change over time. In the present example, this is equivalent to assuming that

$$\mu_{ij+} = \mu_{+ij},$$

which also leads straightforwardly to the set of constraints

$$\log\mu_{ij+} - \log\mu_{+ij} = 0.$$

Example 4

Let X and Y be two polytomous variables, each measured twice. The expected frequencies are represented by μ_{ijkl}, in which the vector $(ijkl)$ of subscripts refers to the joint event $(X_1 = i, Y_1 = j, X_2 = k, Y_2 = l)$. Bivariate marginal homogeneity is obtained if the joint distribution of X and Y remains the same over time, that is, if the following contrasts are satisfied:

$$\log\mu_{ij++} - \log\mu_{++ij} = 0$$

for all pairs (i, j) of scores on (X, Y).

3.2. TESTING INVARIANCE OF ASSOCIATION

In the previous example, the hypothesis that a particular bivariate distribution remained the same over time was tested. In many social science applications, this hypothesis will not be tenable. In such a case, one may be interested in the weaker hypothesis that the association between the variables as measured by the local odds ratios does not change over time.

Example 5

Let, as in Example 4, two discrete variables X and Y be measured twice, and let the joint distribution of the four measurements be represented by P_{ijkl}. In an $m_1 \times m_2$ contingency table for two nominal variables X and Y, the association between the two variables is often described by means of local odds ratios

$$\frac{\Pr(X=i,Y=j)\ \Pr(X=i',Y=j')}{\Pr(X=i,Y=j')\ \Pr(X=i',Y=j)}.$$

To obtain a complete set of independent odds ratios, from which all other odds ratios can be derived, it suffices to take $i' = i + 1$ and $j' = j + 1$, with $i = 1, \ldots, m_1 - 1$ and $j = 1, \ldots, m_2 - 1$. Then, the hypothesis that the association between X and Y does not change over time is equivalent to the following constraints on the expected frequencies:

$$\frac{\mu_{i+1,j+1,+,+}\mu_{i,j,+,+}}{\mu_{i+1,j,+,+}\mu_{i,j+1,+,+}} = \frac{\mu_{+,+,i+1,j+1}\mu_{+,+,i,j}}{\mu_{+,+,i+1,j}\mu_{+,+,i,j+1}}.$$

These restrictions can be translated in terms of contrasts on the logarithms of sums of expected frequencies:

$$\log\mu_{i+1,j+1,+,+} + \log\mu_{i,j,+,+} - \log\mu_{i+1,j,+,+} - \log\mu_{i,j+1,+,+}$$
$$-\log\mu_{+,+,i+1,j+1} - \log\mu_{+,+,i,j} + \log\mu_{+,+,i+1,j} + \log\mu_{+,+,i,j+1} = 0\ .$$

Example 6

For ordinal variables, it may seem more natural to define the structure of association in terms of global instead of local odds ratios (Dale 1986). In an $m_1 \times m_2$ contingency table with ordered categories, a global odds ratio is defined for each pair (a, b) of categories with $1 \leq a \leq m_1 - 1$ and $1 \leq b \leq m_2 - 1$ in the following way:

$$\psi_{a,b} = \frac{\Pr(X \leq a, Y \leq b)\Pr(X > a, Y > b)}{\Pr(X > a, Y \leq b)\Pr(X \leq a, Y > b)}.$$

In a longitudinal study where the same pair of variables is measured twice, one may define global odds ratios $\psi_{a,b,1}$ and $\psi_{a,b,2}$ for the marginal distributions of (X, Y) at the two time periods. It is easy to see that the hypothesis of equal global odds ratios can be tested by defining contrasts of the following kind:

$$\log\psi_{a,b,1} - \log\psi_{a,b,2} = 0,$$

which at the end can be defined as contrasts between the logarithms of sums of expected frequencies.

3.3. HOMOGENEITY OF CONDITIONAL DISTRIBUTIONS

In some social science applications, the investigator may be interested in the question whether certain conditional distributions are constant over time.

Example 7

Let X and Y be two variables measured twice, and assume that X may be thought of as acting as an independent variable that has a causal effect on the dependent variable Y. In an attempt to prove that changes in Y are caused or explained by changes in X, it may be interesting to test whether the conditional distribution of Y given X remains invariant over time. If the latter hypothesis can be accepted, then changes in Y cannot be explained by changes in the effect that X has on Y. Equality of the conditional distribution of Y given X can now be stated as follows:

$$\frac{\mu_{i,j,+,+}}{\mu_{i,+,+,+}} = \frac{\mu_{+,+,i,j}}{\mu_{+,+,i,+}},$$

which also can be written as a contrast between the logarithms of sums of expected frequencies.

3.4. SOME EQUIVALENCES

Because the joint distribution of two random variables X and Y is uniquely defined if the univariate distribution of X and the conditional distribution of Y given X are specified, bivariate marginal homogeneity for the distribution of (X, Y) is equivalent to the conjunction of homogeneity of the conditional distribution of Y given X and univariate homogeneity of the distribution of X.

Similarly, the same joint distribution is also uniquely defined if both the univariate distributions of X and Y are specified and either all local or global odds ratios (Dale 1986) are given. Hence, bivariate marginal homogeneity is also equivalent to the conjunction of univariate homogeneity for the distributions of both X and Y, and of equality of either the local or the global odds ratios.

3.5. GENERAL REPRESENTATION OF HOMOGENEITY CONSTRAINTS

All the different kinds of homogeneity assumptions that have been considered hitherto can be represented as contrasts between the logarithms of sums of expected frequencies. Let $\mathcal{M} = \{1, 2, \ldots, M\}$ denote the index set representing the response patterns. With each contrast α corresponds a set of n_α subsets of $\mathcal{M}$:

$$\{\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_{n_\alpha}\}.$$

With each of these subsets $\mathcal{M}_n$ there corresponds a weight $c_{\alpha n}$ that is either 1 or –1. Using this notation, each contrast can be written as

$$\sum_{n=1}^{n_\alpha} c_{\alpha n} \log\left(\sum_{i \in \mathcal{M}_n} \mu_i\right) = 0.$$

By appropriately defining matrices A, C, X, and U, constraints of this kind can easily be handled by the Lang-Agresti algorithm. Note that C and A should be conformable matrices: If C is of order $r_1 \times r_2$, then A is of order $r_2 \times r_3$. Moreover, in this application the matrix X is a r_1-dimensional zero vector. As a consequence, U is a $r_1 \times r_1$ identity matrix. Note, however, that combining different homogeneity constraints in a single analysis may lead to a situation in which some of the constraints are mutually redundant or incompatible. This problem will be discussed more thoroughly below.

4. MODIFIED PATH MODELS FOR DISCRETE VARIABLES

Goodman (1973) introduced modified path models for the causal analysis of discrete variables when that all variables involved are observed. Basic to Goodman's modified path models is a set of logit equations relating a set of explanatory variables to a sequence of response variables. Because the causal order among the response variables can be taken into account, previous responses can act as explanatory variables for responses that occur later in the causal chain. More recently, a similar class of models was described by Gilula and Haberman (1994, 1995). Hagenaars (1990) discussed an extension of the modified path approach in which some of the variables are unobserved. In general, a modified path model is not equivalent to a single log-linear model defined on the total table; instead, it is equivalent to a conjunction of several log-linear models defined on various subtables, including the total table itself. As a consequence, any modified path model can be formulated as a generalized log-linear model.

4.1. SOME EXAMPLES

Figure 2 shows a path diagram of a causal model for four variables A, B, C, and D. It is assumed that all four variables are dichotomous. The arrows between two variables indicate in which direction the causal influences run.

Causal models of this kind imply a particular decomposition of the joint probability distribution p_{abcd} of the four variables. For four variables A, B, C, and D, one may decompose their joint distribution in a tautological way as the product of a set of conditional distributions:

$$p_{abcd} = p_a p_{b|a} p_{c|ab} p_{d|abc}.$$

Because the causal model shown in Figure 2 deletes the direct arrow from A to D, it implies the nontautological decomposition

$$p_{abcd} = p_a p_{b|a} p_{c|ab} p_{d|bc}.$$

Moreover, in Goodman's modified path approach, it is always assumed that the conditional probabilities satisfy a main effects logit model. For the present example with its four dichotomous variables, this further restricts the model in the following way:

$$\text{logit}(B|A = a) = \beta_0 + \beta_1 a$$

$$\text{logit}(C|A = a, B = b) = \gamma_0 + \gamma_1 a + \gamma_2 b$$

$$\text{logit}(D|B = b, C = c) = \delta_0 + \delta_1 b + \delta_2 c,$$

with, in general,

$$\text{logit}\,(Y|X = x) = \ln\left(\frac{\Pr(Y = 1|X = x)}{\Pr(Y = 0|X = x)}\right).$$

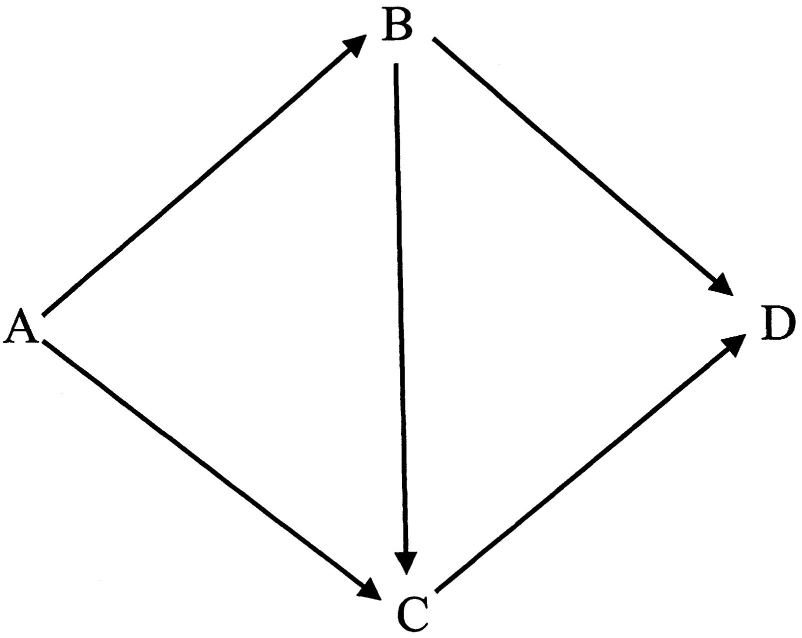

Figure 2: Causal Model 2

The modified path model represented by Figure 2 is equivalent with a single log-linear model defined on the total four-dimensional contingency table. Figure 3 represents this log-linear model in the form of a diagram with undirected edges.

This model, which will usually be denoted as [*AB*, *AC*, *BC*, *BD*, *CD*], is easily derived from Figure 2 by converting the directed arrows into undirected edges.

In general, however, a modified path model will not be equivalent to a single log-linear model for the total table, but will be equivalent to the intersection of several log-linear models that pertain to various marginal tables that display the joint response frequencies for each response variable and its causes in the causal model. Consider the causal model shown in Figure 4.

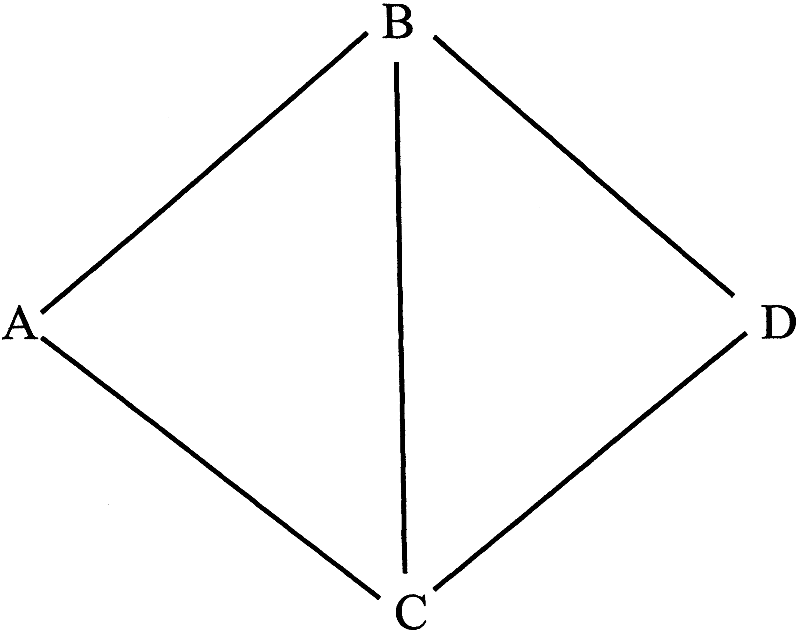

Figure 3: Log-Linear Model 3

Compared to the model shown in Figure 2, a single arrow from A to D has been added. This model implies the following decomposition of the joint probability distribution p_{abcd}:

$$p_{abcd} = p_a p_{b|a} p_{c|ab} p_{d|abc},$$

with the last logit equation altered to

$$\text{logit}(D|A = a, B = b, C = c) = \delta_0 + \delta_1 a + \delta_2 b + \delta_3 c.$$

This model is definitely not equivalent to the log-linear model

$$[AB, AC, AD, BC, BD, CD],$$

which is shown in Figure 5.

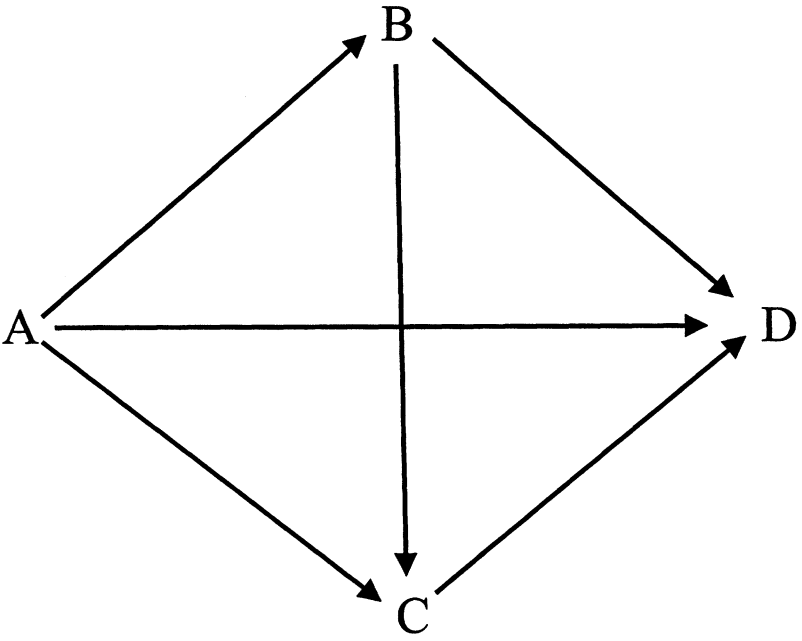

Figure 4: Causal Model 3

The modified path model is equivalent to the intersection of the following log-linear models. One pertains to the total table (*A*, *B*, *C*, *D*), and the other pertains to the marginal table (*A*, *B*, *C*):

1. In table (*A*, *B*, *C*), the model [*AB*, *AC*, *BC*] is satisfied.
2. In table (*A*, *B*, *C*, *D*), the model [*ABC*, *AD*, *BD*, *CD*] is satisfied.

The latter example illustrates the fact that in general, modified path models cannot be tested by a single log-linear analysis.

4.2. ESTIMATING MODIFIED PATH MODELS BY THE LANG-AGRESTI ALGORITHM

Goodman (1973) showed how the iterative proportional fitting algorithm for fitting log-linear models on total tables can be used to fit the intersection of log-linear models defined on the total or marginal

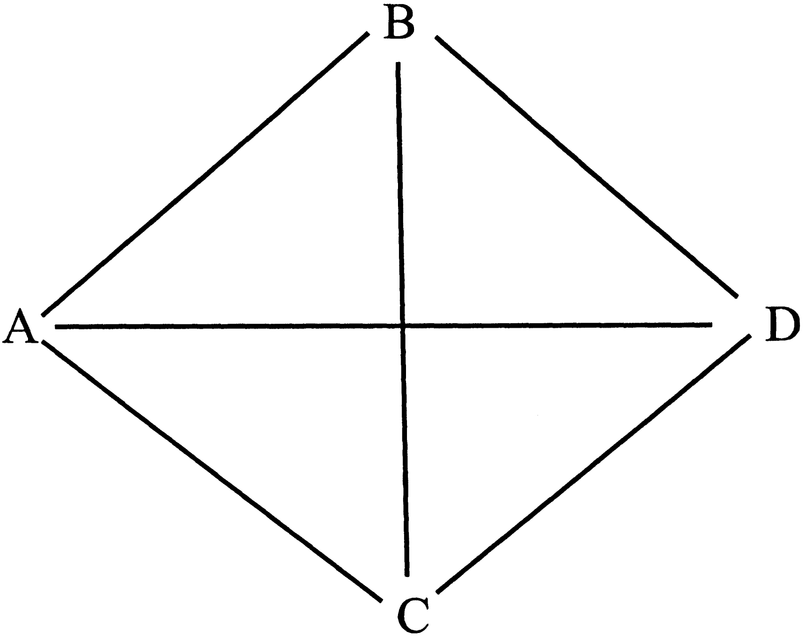

Figure 5: Log-Linear Model 3

tables. Gilula and Haberman (1994, 1995) estimated the parameters of what they called conditional log-linear models by means of an iterative Newton-Raphson procedure.

Here, it will be shown that modified path models can also be estimated and tested by a generalized log-linear analysis. The main advantage of this approach to the estimation of the parameters of a modified path model is that it can easily be extended to cases in which the modified path model itself is combined with other type of hypotheses on some marginal or conditional distributions. Neither Goodman's (1973) approach nor the approach advocated by Gilula and Haberman (1994, 1995) can be used when estimating the parameters of such a modified path model.

Consider first the estimation of the modified path model represented by Figure 4 and assume that all four variables involved are dichotomous. Several vectors and matrices have to be defined. Let the vector μ contain the expected frequencies μ_{abcd}, with the last subscript changing the fastest:

$$\mu = (\mu_{0000}, \mu_{0001}, \mu_{0010}, \ldots, \mu_{1110}, \mu_{1111}).$$

The log-linear model for the marginal (A, B, C) table can be implemented in the following way. First, an 8 × 16 matrix A_1 is defined as follows:

$$A_1 = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix}.$$

The product $A_1\mu$ gives the expected frequencies for the marginal (A, B, C) table. Next, one defines $C_1 = I_{8 \times 8}$ and

$$X_1 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}.$$

The one-dimensional null-space of X_1 is then spanned by the matrix $U_1' = (1, -1, -1, 1, -1, 1, 1, -1)$.

To represent the log-linear model for the total table, the definitions of the matrices A and C are simple: $A_2 = C_2 = I_{16 \times 16}$. The model design matrix X_2 is of order 16 × 12. Apart from the constant column corresponding to the overall log-linear parameter, the columns of X_2 correspond to the four main effects, the six two-variable interaction effects, and the one three-variable interaction effect for the triple (A, B, C). The null-space of X_2 is spanned by the columns of a 16×4 matrix U_2.

Finally, to implement the simultaneous analysis of the marginal and total log-linear models, the following supermatrices have to be defined:

$$A = \begin{pmatrix} A_1 \\ A_2 \end{pmatrix},$$

$$C = \begin{pmatrix} C_1 & O_{8\times16} \\ O_{16\times8} & C_2 \end{pmatrix}$$

$$U' = \begin{pmatrix} U_1' & O_{1\times16} \\ O_{4\times8} & U_2' \end{pmatrix}.$$

4.3. SOME GENERAL CONSIDERATIONS

In general, modified path models can be characterized in the following way. In any modified path model, the set of explanatory exogenous variables $\chi = \{X_1, X_2, \ldots, X_q\}$ is assumed to exercise in a well-specified manner causal influences on the set of endogenous response variables $\mathcal{Y} = \{Y_1, \ldots, Y_p\}$. The association among the exogenous variables is not explained by the modified path itself. In general, the association structure among the exogenous variables is not further specified. For discrete variables, it is usually assumed to satisfy the saturated log-linear model, although in the search for a more parsimonious model one might consider more restricted log-linear models for their association without considering hypotheses on their respective causal ordering. The set of dependent variables in $\mathcal{Y}$, on the other hand, can be ordered with respect to causal priority. In the sequel, it is assumed that the response variables are indexed in such a way that Y_1 is causally prior to Y_2, which itself is causally prior to Y_3, and so on. The last variable in the causal chain is then Y_p.

For each dependent variable, Y_k, there exists a subset of explanatory variables $S_k \subseteq \chi$ and a subset of causally prior response variables $\mathcal{T}_k \subseteq \{Y_1, \ldots, Y_{k-1}\}$, which are assumed to be direct causes of Y_k. For each Y_k, a logit model is specified in which all variables from

$C_k = \mathcal{T}_k \cup S_k$ are directly linked to Y_k. The part of the modified path model that corresponds to Y_k is then equivalent to a hierarchical log-linear model on the marginal table defined for the variables in $\{Y_k\} \cup C_k$. This log-linear model contains in the first place all the main and interaction effects that can be defined for the variables in C_k. Furthermore, it contains the two-variable interaction effects that can be defined by pairing Y_k to each variable in C_k. As an example, suppose that in a particular modified path model, Y_3 is influenced by X_1, X_2, and the prior response variable Y_2, but not by the prior response variable Y_1, then the log-linear model for the marginal table (X_1, X_2, Y_2, Y_3) is the model $[X_1X_2Y_2, X_1Y_3, X_2Y_3, Y_2Y_3]$.

To estimate the modified path model by means of a generalized log-linear analysis, appropriate matrices A_k, C_k, and X_k have to be defined for each logit model. By means of the matrix A_k, one defines the expected frequencies in the marginal table that corresponds with the kth logit equation. The matrix C_k is always an identity matrix with the appropriate dimensions. The design matrix X_k defines the log-linear model that corresponds with the kth logit equation. The matrix whose columns span the null-space of X_k is U_k.

The complete modified path model is then equivalent to the intersection of all log-linear models that correspond to the set of logit equations. For a generalized log-linear analysis, one needs to define the following supermatrices:

$$A = \begin{pmatrix} A_1 \\ A_2 \\ \vdots \\ A_p \end{pmatrix}$$

and

$$U = \bigoplus_k U_k,$$

in which $\oplus$ represents the direct sum of matrices:

$$\bigoplus_{k=1}^{2} M_k = \begin{pmatrix} M_1 & O \\ O & M_2 \end{pmatrix}.$$

Note that this way of combining various log-linear models defined on different marginal tables derived from the same total table is only valid if the parameter spaces of the different log-linear models do not intersect. By formulating equality constraints on log-linear parameters from different submodels, the different parameter spaces would have common elements. In that case, the Lang-Agresti algorithm can still be used, but now one has to create a supermatrix X that incorporates the assumed equality constraints. The final U matrix, then, contains the basis of the null-space of X.

The maximum likelihood estimates of the expected frequencies under the modified path model are now determined by maximizing the log-likelihood function under the constraints

$$U' \log A\mu = 0.$$

5. COMBINING MODIFIED PATH MODELS AND HOMOGENEITY ASSUMPTIONS

The strength of the generalized log-linear model for the analysis of contingency tables resides in the fact that several log-linear or linear models defined on partial or marginal tables and on the total table can be fitted simultaneously. Consider the model represented by Figure 6. The four variables involved in this modified path model correspond to an independent variable X measured at two time points and a dependent variable Y also measured twice at the same time points. The four variables involved in this model are represented by X_1, X_2, Y_1, and Y_2. The model assumes that the independent variable X_1 has a direct effect on the dependent variable Y_1 and on the independent variable X_2. The independent variable Y_2 depends causally on X_2 and on the value of Y_1. In this model, X_1 is the only exogenous variable; all three other variables are endogenous. The modified path model represented by Figure 6 is equivalent to the intersection of the following two log-linear models:

1. In the marginal table (X_1, Y_1, X_2), log-linear model $[X_1Y_1, X_1X_2]$ holds.
2. In the total table (X_1, Y_1, X_2, Y_2), log-linear model $[X_1Y_1X_2, Y_1Y_2, X_2Y_2]$ holds.

The modified path model described above allows the investigator to assess how strongly scores on a response variable are related to certain explanatory variables and to prior response variables. In longitudinal studies, it might also be interesting to investigate whether changes in the dependent response variables are caused by changes in some independent variables. In this context, one could consider the hypothesis that the conditional distribution of the dependent variable given the independent variables remains constant over time. In the present example, this implies testing the equality of the conditional distributions of Y_1 given X_1 and of Y_2 given X_2. In terms of marginal distributions defined on the total table containing the expected frequencies μ_{ijkl} for the random vector (X_1, Y_1, X_2, Y_2), the equality of these two conditional distributions leads to the following constraints:

$$\frac{\mu_{i,j,+,+}}{\mu_{i,+,+,+}} = \frac{\mu_{+,+,i,j}}{\mu_{+,+i}}$$

for all i and j. As discussed in a previous section, homogeneity conditions of this kind can easily be handled by the generalized log-linear model. Moreover, they can be combined with the constraints implied by the formal structure of a modified path model. Let the pair of matrices (A_1, U_1) characterize the constraints implied by the modified path model, and let the pair of matrices (A_2, U_2) characterize the constraints implied by the homogeneity assumptions. Then, the model that is the combination of both submodels is characterized by the matrices

$$A = \begin{pmatrix} A_1 \\ A_2 \end{pmatrix}$$

and $U = U_1 \oplus U_2$.

A potential problem with combining different models in a simultaneous analysis is that they may contain redundant constraints. Lang (1996b) discussed this problem and formulated a general condition that is sufficient to ensure nonredundancy. In the present context of

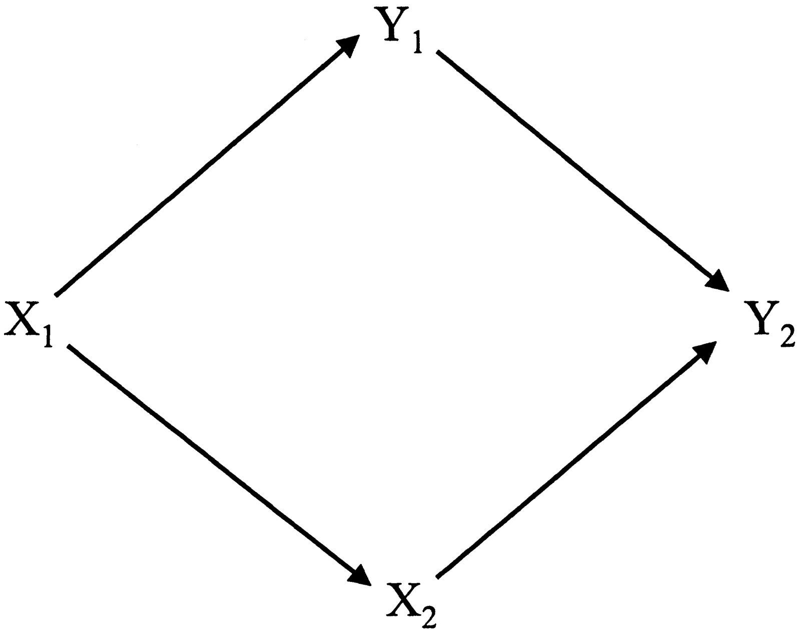

Figure 6: Causal Model 3

combining a homogeneity model and a path model, this condition leads to the following result.

Let $D_{A_1\mu}$ and $D_{A_2\mu}$ be diagonal matrices containing the elements from $A_1\mu$ and $A_2\mu$. Define the matrices

$$H_1 = U_1' D_{A_1\mu} A_1$$

and

$$H_2 = U_2' D_{A_2\mu} A_2.$$

If the matrix $H = (H_1 \vdots H_2)$ is of full column rank, the constraints implied by the two different models are nonredundant.

In general, the constraints implied by a marginal model and by a path model will be nonredundant. Only in special cases will a homogeneity hypothesis be implied by a log-linear model. An example of a situation in which a log-linear model implies marginal homogeneity is the following. Consider a $R \times R$ table in which a symmetric independence model is being fitted. This log-linear model, which can be written in a multiplicative form as $\mu_{ij} = \beta_i \bullet \beta_j$, implies marginal homogeneity $\mu_{i+} = \mu_{+i}$. Hence, in this model the marginal homogeneity constraints are redundant given the constraints implied by the log-linear model.

Note, however, that redundancy can also arise within each group of constraints itself. For example, the log-linear model $[AB, C]$ for the three-dimensional table (ABC) implies independence between A and C in the two-dimensional table (AC). So, each group of constraints should also be checked on its own for redundancy.

6. *APPLICATION*

In a large-scale survey titled Social Security: Research on Demographic and Psychological Aspects, sponsored by the Dutch Ministry of Education, several social and psychological effects of unemployment were studied. The data for a sample of 427 respondents refer to the variables labor market status (1 = unemployed, 2 = employed) and structuration of daily activities (1 = poor structuration, 2 = good structuration) measured at two time points. In the sequel, the variable labor market status will be denoted by X and the variable structuration of daily activities will be denoted by Y. Table 1 gives the $2 \times 2 \times 2 \times 2$ contingency table for X_1, Y_1, X_2, and Y_2.

By means of the generalized log-linear model, several hypotheses were tested on these data. All the analyses reported in this section used specific software that was developed in Mathematica. A listing of the Mathematica source code for maximum likelihood fitting of the class of models described in this article can be found in Bergsma (1997, appendix E).

A first series of tests were concerned with several homogeneity hypotheses. The first homogeneity hypothesis concerned whether the joint distribution of (X, Y) remains the same over time. Using the notation introduced earlier, one tests the hypothesis that the equality

$$\mu_{ij++} = \mu_{++ij}$$

TABLE 1: Response Patterns and Their Frequencies

1111	41	2111	1
1112	23	2112	2
1121	12	2121	39
1122	17	2122	28
1211	20	2211	2
1212	48	2212	2
1221	2	2221	22
1222	25	2222	143

holds for all pairs of scores (i, j) on the variables (X, Y). The null hypothesis that the joint distribution of (X, Y) is invariant over time had to be rejected with $L^2 = 45.3102$ with 3 degrees of freedom ($p < .001$).

Note that the same hypothesis can also be tested by testing for row and column homogeneity in the 4 × 4 table formed by taking the four different response patterns on the joint variables (X, Y) as the categories of a new variable. In this way, the rows of the 4 × 4 table would correspond to the measurements at the first time point; its column would correspond to the measurements at the second time point. However, the outcome of the homogeneity test on this new table would be exactly the same as that of the test that was actually performed on the 2^4 table.

The second homogeneity hypothesis concerned whether the univariate distributions of X and Y were invariant over time. Testing univariate homogeneity for X tests whether

$$\mu_{i+++} = \mu_{++i+}$$

for all scores i on X. Similarly, for Y one tests whether

$$\mu_{+j++} = \mu_{+++j}$$

holds for all scores j on Y.

For the data at hand, both hypotheses of univariate homogeneity had to be rejected. The value of L^2 was equal to 43.3837 for the distribution of X and 5.0016 for the distribution of Y. Both tests are with 1 degree of freedom and significant at the 5 percent level. The simultaneous test of both homogeneity hypotheses yielded $L^2 = 45.0984$, which with 2 degrees of freedom is highly significant.

Table 2, which gives the joint distribution of the scores on X for the two time periods, clearly shows that many more people change from unemployment to employment than the other way around. Table 3, which gives the joint distribution of the scores on Y for the two time periods, shows that there is also a significant change in the proportion of people reporting well-structured daily activities.

Both hypotheses of univariate homogeneity could also have been tested by McNemar's (1947) test for the difference between correlated proportion, or by Bhapkar's (1966) variant of it. The results of these alternative tests were similar to those reported above. McNemar's test (with 1 degree of freedom) was equal to 38.11 for variable X and 4.97 for variable Y. This similarity between the outcomes of the two test procedures should not come as a surprise, since McNemar's test is a Wald test statistic that is asymptotically equivalent to the log-likelihood ratio test (Agresti 1990:359-60).

Note also that the univariate marginal distributions of X_2 and Y_2 are exactly identical in these data. However, this is a mere coincidence in the data and cannot be given a substantive interpretation.

In a third analysis using the generalized log-linear model, we tested the hypothesis of whether the association (as measured by means of the odds ratio) between X and Y remains the same over time. For the present data, this hypothesis is equivalent to the constraint that the expected frequencies satisfy

$$\frac{\mu_{11++}\mu_{22++}}{\mu_{12++}\mu_{21++}} = \frac{\mu_{++11}\mu_{++22}}{\mu_{++12}\mu_{++21}} .$$

The null hypothesis of no change in association between X and Y could not be rejected, since $L^2 = 0.0101$ with 1 degree of freedom. Inspection of the data matrix showed that the observed odds ratios were almost identical at the two time points: 2.3635 at the first and 2.4234 at the second time point.

Finally, in a fourth analysis, we tested the hypothesis of whether the conditional distributions $p(Y|X = 1)$ and $p(Y|X = 2)$ remain the same over time. For the present data, this hypothesis is equivalent to the following two constraints on the expected frequencies:

TABLE 2: **Labor Market Status at Time 1 and Time 2**

	$X_2 = 1$	$X_2 = 2$	
$X_1 = 1$	132	56	188
$X_1 = 2$	7	232	239
	139	288	427

TABLE 3: **Structuration Activities at Time 1 and Time 2**

	$Y_2 = 1$	$Y_2 = 2$	
$Y_1 = 1$	93	70	163
$Y_1 = 2$	46	218	264
	139	288	427

$$\frac{\mu_{12++}}{\mu_{1+++}} = \frac{\mu_{++12}}{\mu_{++1+}}$$

$$\frac{\mu_{22++}}{\mu_{2+++}} = \frac{\mu_{++22}}{\mu_{++2+}}.$$

The null hypothesis that the two conditional distributions are invariant over time could not be rejected: the analysis yielded $L^2 = 2.1392$, which with 2 degrees of freedom corresponds to a probability level of $p = .343$.

The conclusion to the analyses based on various homogeneity assumptions is that the univariate distributions of both the independent variable X and the dependent variable Y change over time, but that the conditional distributions $p(Y|X = 1)$ and $p(Y|X = 2)$ are invariant. Consequently, the strength of the association between both variables, as measured by the odds ratio, is also invariant over time.

A second series of analyses studied whether the relations between the four variables could be explained by means of a particular causal model. The model being tested consisted of two separate causal hypotheses and is shown in Figure 6. First, it was assumed that labor market status at time point 1 has a direct effect on structuration at the same time point and a direct effect on labor market status at time point 2. This part of the model is equivalent to the log-linear model $[X_1Y_1, X_1X_2]$ in the three-dimensional table $(X_1Y_1X_2)$. Second, it was postulated that structuration at time point 2 is affected by structuration at the previous time point 1 and by labor market status at time point 2. This part of the model is equivalent to the log-linear model $[X_1Y_1X_2, Y_1Y_2, X_2Y_2]$ in the total table $(X_1Y_1X_2Y_2)$.

This modified path model was fitted to the data by means of the Lang-Agresti algorithm. The model provides an excellent fit to the data, since the analysis resulted in a test statistic of $L^2 = 8.1506$, which with 7 degrees of freedom corresponds to a probability level of $p = .320$. The estimated logit equation is

$$\hat{\text{logit}}(Y_2 \mid y_1, x_2) = -3.46 + 1.78y_1 + 0.75x_2.$$

The standard errors of the coefficients of Y_1 and X_2 are 0.23 and 0.24, respectively. Both coefficients are significantly different from zero at the 1 percent level.

The same model could also have been tested by means of the procedure described in Goodman (1973) and Gilula and Haberman (1994). In this procedure, the two log-linear models are fitted separately to the two tables and the two L^2 test statistics and the degrees of freedom for each model are added to obtain the overall L^2 value and the total degrees of freedom. The results of this more classical approach to testing modified path models are, of course, exactly identical to those obtained by an analysis based on the generalized log-linear model.

As a final analysis, the modified path model given in Figure 6 was combined with the assumption of homogeneity of the conditional distributions of Y given X. This model cannot be estimated by the procedures proposed by Goodman (1974) and Gilula and Haberman (1995), since the homogeneity constraints on the conditional distribution cannot be translated into log-linear terms. (See, however, Vermunt [1997, appendix F] for an adaptation of Goodman's estimation procedure that allows restrictions on the conditional probabilities.)

The analysis based on the generalized log-linear model resulted in $L^2 = 10.2177$ with 9 degrees of freedom ($p = .333$). The value of the conditional log-likelihood ratio test of this model against the hypothesis represented by the modified path model without the homogeneity constraints is 2.0671, which with 2 degrees of freedom is not significant. Adding the assumption of homogeneity of the conditional distributions to the assumptions of the modified path model does not worsen the fit in a significant way.

An interesting observation is that the value of L^2 for the combined model is approximately equal to the sum of the values of L^2 for the separate models:

$$10.2177 \approx 8.1506 + 2.1392.$$

This quasi-equality could be an indication of the fact that the two models are asymptotically separable; that for large sample sizes the combination of the two models can be tested by adding the test statistics of the two separate models. (See Lang [1996b] for a more thorough discussion of the concept of asymptotic separability and conditions that ensure it.)

As to the parameter estimates under the combined model, the conditional distribution of Y given X, which is held constant over time, is given by the following conditional probabilities:

$$\hat{p}(Y_t = 2|X_t = 1) = 0.525$$

$$\hat{p}(Y_t = 2|X_t = 2) = 0.736.$$

The conditional distribution of X_2 given X_1 is given by

$$\hat{p}(X_2 = 2|X_1 = 1) = 0.288$$

$$\hat{p}(X_2 = 2|X_1 = 2) = 0.965.$$

Finally, the conditional distribution of Y_2 given X_2 and Y_1 satisfies the following logit model:

$$\widehat{\text{logit}}(Y_2|y_1, x_2) = -3.32 + 1.77y_1 + 0.75x_2.$$

The standard error for the coefficient of Y_1 in the logit equation above was equal to 0.23; the standard coefficient of X_2 in the same equation was 0.17. Both coefficients are significantly different from zero at the 1 percent level. The estimated logit equation and the standard errors of the coefficients from the joint analysis are similar to their counterparts obtained in the previous analysis. Only the standard error of the coefficient of X_2 becomes somewhat smaller in the joint analysis. The two logit equations are similar because the extra homogeneity conditions imposed during the joint analysis are well satisfied by the data.

7. DISCUSSION

This article has shown that various homogeneity models and log-linear models, and combinations thereof, can be tested by means of the generalized log-linear model described by Lang and Agresti (1994). Although some of the models considered in this article can also be tested by more classical or traditional methods, the generalized log-linear model has a much broader scope and allows particular combinations of the models being tested. The most interesting and promising point of the generalized log-linear model is the fact that one may test models that simultaneously impose linear constraints on the expected frequencies and on their logarithms.

NOTE

[1] Throughout this article, the term *causal* will be used somewhat loosely to denote asymmetrical relationships between the variables. For more precise definitions of and views on causality, see, among others, Rubin (1974), Sobel (1995), and Pearl (1995).

REFERENCES

Agresti, Alan. 1990. *Categorical Data Analysis*. New York: Wiley.

Bergsma, Wicher. 1997. *Marginal Models for Categorical Data*. Tilburg: Tilburg University Press.

Bhapkar, V. P. 1966. "A Note on the Equivalence of Two Test Criteria for Hypotheses in Categorical Data." *Journal of the American Statistical Association* 61:228-35.

Clogg, Cliff C., Scott R. Eliason, and John M. Grego. 1990. "Models for the Analysis of Change in Discrete Variables." Pp. 409-41 in *Statistical Methods in Longitudinal Research*, vol. 2, *Time Series and Categorical Data*, edited by A. von Eye. New York: Academic Press.

Dale, Jocelyn R. 1986. "Global Cross-Ratio Models for Bivariate, Discrete, Ordered Responses." *Biometrics* 42:909-17.

Duncan, Otis D. 1980. "Testing Key Hypotheses in Panel Analysis." Pp. 279-89 in *Sociological Methodology 1980*, edited by Karl F. Schuessler. San Francisco: Jossey-Bass.

———. 1981. "Two Faces of Panel Analysis: Parallels With Comparative Cross-Sectional Analysis and Time-Lagged Association." Pp. 281-318 in *Sociological Methodology 1981*, edited by Samuel Leinhardt. San Francisco: Jossey-Bass.

Gilula, Zvi and Shelby J. Haberman. 1994. "Conditional Log-Linear Models for Analyzing Categorical Panel Data." *Journal of the American Statistical Association* 89:645-56.

———. 1995. "Prediction Functions for Categorical Panel Data." *Annals of Statistics* 23:1130-42.

Goodman, Leo. 1973. "The Analysis of Multidimensional Contingency Tables When Some of the Variables Are Posterior to Others: A Modified Path Approach." *Biometrika* 60:179-92.

Grizzle, James E., C. Frank Starmer, and Gary G. Koch. 1969. "Analysis of Categorical Data by Linear Models." *Biometrics* 25:489-504.

Hagenaars, Jacques A. 1990. *Categorical Longitudinal Data: Log-Linear Panel, Trend, and Cohort Data*. Newbury Park, CA: Sage.

———. 1992. "Analyzing Categorical Longitudinal Data Using Marginal Homogeneity Models." *Statistica Applicata* 4:763-71.

Lang, Joseph B. 1996a. "Maximum Likelihood Methods for a Generalized Class of Log-Linear Models." *Annals of Statistics* 24:726-52.

———. 1996b. "On the Partitioning of Goodness-of-Fit Statistics for Multivariate Categorical Response Models." *Journal of the American Statistical Association* 91:1017-23.

Lang, Joseph B. and Alan Agresti. 1994. "Simultaneously Modelling the Joint and Marginal Distributions of Multivariate Categorical Responses." *Journal of the American Statistical Association* 89:625-32.

Lang, Joseph B. and Scott R. Eliason. 1997. "Applications of Association-Marginal Models to the Study of Social Mobility." *Sociological Methods & Research* 26:183-212.

Lauritzen, Steffen L. 1996. *Graphical Models*. Oxford, UK: Clarendon.

McNemar, Quinn. 1947. "Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages." *Psychometrika* 12:153-57.

Pearl, Judea. 1995. "Causal Diagrams for Empirical Research." *Biometrika* 82:669-710.

Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66:688-701.

Sobel, Michael E. 1995. "Causal Inferences in the Social and Behavioral Sciences." Pp. 1-38 in *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, edited by Gerard Arminger, Cliff C. Clogg, and Michael E. Sobel. New York: Plenum.

Vermunt, Jeroen K. 1997. *Log-Linear Models for Event Histories*. Thousand Oaks, CA: Sage.

Whittaker, Joe. 1990. *Graphical Models in Applied Multivariate Statistics*. Chichester, UK: Wiley.

Marcel A. Croon is an associate professor in the Methodology Department of the Faculty of Social Sciences at Tilburg University. His research interests are in applied statistics, measurement theory, and research methodology.

Wicher Bergsma is a postdoctoral researcher in the Methodology Department at Tilburg University. His research was supported by the Netherlands Organization for Scientific Research.

Jacques A. Hagenaars is a professor in the Methodology Department at Tilburg University. His research interests are in social statistics, causal models, and research methodology. He is the author of Categorical Longitudinal Data *and* Log-Linear Models With Latent Variables.