Log-linear modelling

Jeroen K. Vermunt
Department of Methodology and Statistics
Tilburg University

1 Introduction

Log-linear analysis has become a widely used method for the analysis of multivariate frequency tables obtained by cross- classifying sets of nominal, ordinal, or discrete interval level variables. Examples of textbooks discussing categorical data analysis by means of log-linear models are [4], [2], [14], [15], [16], and [27].

We start by introducing the standard hierarchical log-linear modelling framework. Then, attention is paid to more advanced types of log-linear models that make it possible to impose interesting restrictions on the model parameters, for example, restrictions for ordinal variables. Subsequently, we present "regression-analytic", "path-analytic", and "factor-analytic" variants of log-linear analysis. The last section discusses parameter estimation by maximum likelihood, testing, and software for log-linear analysis.

2 Hierarchical log-linear models

2.1 Saturated models

Suppose we have a frequency table formed by three categorical variables which are denoted by A, B, and C, with indices a, b, and c. The number of categories of A, B, and C is denoted by A^*, B^*, and C^*, respectively. Let m_{abc} be the expected frequency for the cell belonging to category a of A, b of B, and c of C. The saturated log-linear model for the three-way table ABC is given by

$$\log m_{abc} = \lambda + \lambda^A_a + \lambda^B_b + \lambda^C_c + \lambda^{AB}_{ab} + \lambda^{AC}_{ac} + \lambda^{BC}_{bc} + \lambda^{ABC}_{abc} . \qquad (1)$$

It should be noted that the log transformation of m_{abc} is tractable because it restricts the expected frequencies to remain within the admissible range.

The consequence of specifying a linear model for the log of m_{abc} is that a multiplicative model is obtained for m_{abc}, i.e.,

$$\begin{aligned} m_{abc} &= \exp\left(\lambda + \lambda_a^A + \lambda_b^B + \lambda_c^C + \lambda_{ab}^{AB} + \lambda_{ac}^{AC} + \lambda_{bc}^{BC} + \lambda_{abc}^{ABC}\right) \\ &= \tau\tau_a^A\tau_b^B\tau_c^C\tau_{ab}^{AB}\tau_{ac}^{AC}\tau_{bc}^{BC}\tau_{abc}^{ABC}\,. \end{aligned} \tag{2}$$

From Equations 1 and 2, it can be seen that the saturated model contains all interactions terms among A, B, and C. That is, no a priori restrictions are imposed on the data. However, Equations 1 and 2 contain too many parameters to be identifiable. Given the values for the expected frequencies m_{abc}, there is not a unique solution for the λ and τ parameters. Therefore, constraints must be imposed on the log-linear parameters to make them identifiable. One option is to use ANOVA-like constraints, namely,

$$\begin{aligned} &\sum_a \lambda_a^A = \sum_b \lambda_b^B = \sum_c \lambda_c^C = 0\,, \\ &\sum_a \lambda_{ab}^{AB} = \sum_b \lambda_{ab}^{AB} = \sum_a \lambda_{ac}^{AC} = \sum_c \lambda_{ac}^{AC} = \sum_b \lambda_{bc}^{BC} = \sum_c \lambda_{bc}^{BC} = 0\,, \\ &\sum_a \lambda_{abc}^{ABC} = \sum_b \lambda_{abc}^{ABC} = \sum_c \lambda_{abc}^{ABC} = 0\,. \end{aligned}$$

This parameterization in which every set of parameters sums to zero over each of its subscripts is called *effect coding*. In effect coding, the λ term denotes the grand mean of $\log m_{abc}$. The one-variable parameters λ_a^A, λ_b^B, and λ_c^C indicate the relative number of cases at the various levels of A, B, and C as deviations from the mean. More precisely, they describe the partial skewness of a variable, that is, the skewness within the combined categories of the other variables. The two-variable interaction terms λ_{ab}^{AB}, λ_{ac}^{AC}, and λ_{bc}^{BC} indicate the strength of the partial association between A and B, A and C, and B and C, respectively. The partial association can be interpreted as the mean association between two variables within the levels of the third variable. And finally, the three-factor interaction parameters λ_{abc}^{ABC} indicate how much the conditional two-variable interactions differ from one another within the categories of the third variable.

Another method to identify the log-linear parameters involves fixing the parameters to zero for one category of A, B, and C, respectively. This parameterization, which is called *dummy coding*, is often used in regression models with nominal regressors. Although the expected frequencies under both parameterizations are equal, the interpretation of the parameters is

rather different. When effect coding is used, the parameters must be interpreted in terms of deviations from the mean, while under dummy coding, they must interpreted in terms of deviations from the reference category.

2.2 Non-saturated models

As mentioned above, in a saturated log-linear model, all possible interaction terms are present. In other words, no a priori restrictions are imposed on the model parameters apart from the identifying restrictions. However, in most applications, the aim is to specify and test more parsimonious models, that is, models in which some a priori restrictions are imposed on the parameters. Log-linear models in which the parameters are restricted in some way are called non-saturated models. There are different kinds of restrictions that can be imposed on the log-linear parameters. One particular type of restriction leads to the family of hierarchical log-linear models. These are models in which the log-linear parameters are fixed to zero in such a way that when a particular interaction term is fixed to zero, all higher-order interaction terms containing all its indices as a subset must also be fixed to zero. For example, if the partial association between A and B (λ_{ab}^{AB}) is assumed not to be present, the three-variable interaction λ_{abc}^{ABC} must be fixed to zero as well. Applying this latter restriction to Equation 1 results in the following non-saturated hierarchical log-linear model:

$$\log m_{abc} \;=\; \lambda + \lambda_a^A + \lambda_b^B + \lambda_c^C + \lambda_{ac}^{AC} + \lambda_{bc}^{BC} \,. \tag{3}$$

Another example of a non-saturated hierarchical log-linear model is the (trivariate) independence model

$$\log m_{abc} \;=\; \lambda + \lambda_a^A + \lambda_b^B + \lambda_c^C \,.$$

Hierarchical log-linear models are the most popular log-linear models because, in most applications, it is not meaningful to include higher-order interaction terms without including the lower-order interaction terms concerned. Another reason is that it is relatively easy to estimate the parameters of hierarchical log-linear models because of the existence of simple minimal sufficient statistics (see maximum likelihood estimation).

3 Other types of log-linear models

3.1 General log-linear model

So far, attention has been paid to only one special type of log-linear models, the hierarchical log-linear models. As demonstrated, hierarchical log-linear models are based on one particular type of restriction on the log-linear parameters. But, when the goal is to construct models which are as parsimonious as possible, the use of hierarchical log-linear models is not always appropriate. To be able to impose other kinds of linear restrictions on the parameters, it is necessary to use more general kinds of log-linear models.

As shown by McCullagh and Nelder [23], log-linear models can also be defined in a much more general way by viewing them as a special case of the generalized linear modelling (GLM) family. In its most general form, a log-linear model can be defined as

$$\log m_i \;=\; \sum_j \lambda_j x_{ij}\,, \tag{4}$$

where m_i denotes a cell entry, λ_j a log-linear parameter, and x_{ij} an element of the design matrix. The design matrix provides us with a very flexible tool for specifying log-linear models with various restrictions on the parameters. For detailed discussions on the use of design matrices in log-linear analysis, see, for example, [10], [14], [15], and [25].

Let us first suppose we want to specify the design matrix for an *hierarchical* log-linear model of the form $\{AB, BC\}$. Assume that A^*, B^*, and C^*, the number of categories of A, B, and C, are equal to 3, 3, and 4, respectively. Because in that case model $\{AB, BC\}$ has 18 independent parameters to be estimated, the design matrix will consist of 18 columns: 1 column for the main effect λ, 7 ($[A^*-1]+[B^*-1]+[C^*-1]$) columns for the one-variable terms λ_a^A, λ_b^B, and λ_c^C, and 10 ($[A^*-1]*[B^*-1]+[B^*-1]*[C^*-1]$) columns for the two-variable terms λ_{ab}^{AB} and λ_{bc}^{BC}. The exact values of the cells of the design matrix, the x_{ij}, depend on the restrictions which are imposed to identify the parameters. Suppose, for instance, that column j refers to the one-variable term λ_a^A and that the highest level of A, A^*, is used as the (arbitrary) omitted category. In effect coding, the element of the design matrix corresponding to the ith cell, x_{ij}, will equal 1 if $A = a$, -1 if $A = A^*$, and otherwise 0. On the other hand, in dummy coding, x_{ij} would be 1 if $A = a$, and otherwise 0. The columns of the design matrix referring to the

two-variable interaction terms can be obtained by multiplying the columns for the one-variable terms for the variables concerned (see [10] and [14]).

The design matrix can also be used to specify all kinds of *non-hierarchical* and *non-standard* models. Actually, by means of the design matrix, three kinds of linear restrictions can be imposed on the log-linear parameters: a parameter can be fixed to zero, specified to be equal to another parameter, and specified to be in a fixed ratio to another parameter.

The first kind of restriction, *fixing to zero*, is accomplished by deleting the column of the design matrix referring to the effect concerned. Note that, in contrast to hierarchical log-linear models, parameters can be fixed to be equal to zero without the necessity of deleting the higher-order effects containing the same indices as a subset.

Equating parameters is likewise very simple. Equality restrictions are imposed by adding up the columns of the design matrix which belong to the effects which are assumed to be equal. Suppose, for instance, that we want to specify a model with a symmetric association between the variables A and B,[1] each having three categories. This implies that

$$\lambda_{ab}^{AB} \;=\; \lambda_{ba}^{AB}\,.$$

The design matrix for the unrestricted effect λ_{ab}^{AB} contains four columns, one for each of the parameters λ_{11}^{AB}, λ_{12}^{AB}, λ_{21}^{AB}, λ_{22}^{AB}. In terms of these four parameters, the symmetric association between A and B implies that λ_{12}^{AB} is assumed to be equal to λ_{21}^{AB}. This can be accomplished by summing the columns of the design matrix referring to these two effects.

As already mentioned above, parameters can also be restricted to be in a *fixed ratio* to each other. This is especially useful when the variables concerned can be assumed to be measured on an ordinal or interval level scale, with known scores for the different categories. Suppose, for instance, that we wish to restrict the one-variable effect of variable A to be linear. Assume that the categories scores of A, denoted by a, are equidistant, that is, that they take on the values 1, 2, and 3. Retaining the effect coding scheme, a linear effect of A is obtained by

$$\lambda_{a}^{A} \;=\; (a-\bar{a})\lambda^{A}\,.$$

[1] Log-linear models with symmetric interaction terms may be used for various purposes. In longitudinal research, they may be applied to test the assumption of marginal homogeneity (see [3] and [16]). Other applications of log-linear models with symmetric association parameters are Rasch models for dichotomous (see [24] and [17]) and polytomous items (see [4]).

Here, $\bar{a}$ denotes the mean of the category scores of A, which in this case is 2. Moreover, λ^A denotes the single parameter describing the one-variable term for A. It can be seen that the distance between the λ_a^A parameters of adjacent categories of A is λ^A. In terms of the design matrix, such a specification implies that instead of including $A^* - 1$ columns for the one-variable term for A, one column with scores $(a - \bar{a})$ has to be included.

These kinds of linear constraints can also be imposed on the bivariate association parameters of a log-linear model. The best known examples are linear-by-linear interaction terms and row- or column-effect models (see [5], [7], [13], and [15]). When specifying a linear-by-linear interaction term, it is assumed that the scores of the categories of both variables are known. Assuming equidistant scores for the categories of the variables A and B and retaining the effect coding scheme, the linear-by-linear interaction between A and B is given by

$$\lambda_{ab}^{AB} = (a - \bar{a})(b - \bar{b})\lambda^{AB} . \tag{5}$$

Using this specification, which is sometimes also called uniform association, the (partial) association between A and B is described by a single parameter instead of using $(A^* - 1)(B^* - 1)$ independent λ_{ab}^{AB} parameters. As a result, the design matrix contains only one column for the interaction between A and B consisting of the scores $(a - \bar{a})(b - \bar{b})$.

A row association structure is obtained by assuming the column variable to be linear. When A is the row variable, it is defined as

$$\lambda_{ab}^{AB} = (b - \bar{b})\lambda_a^{AB} .$$

Note that for every value of A, there is a λ_a^{AB} parameter. Actually, there are $(A^* - 1)$ independent row parameters. Therefore, the design matrix will contain $(A^* - 1)$ columns which are based on the scores $(b - \bar{b})$. The column association model is, in fact, identical to the row association model, only the roles of the column and row variable change.

3.2 Log-rate model

The general log-linear model discussed in the previous section can be extended to include an additional component, viz., a cell weight ([14] and [19]). The log-linear model with cell weights is given by

$$\log\left(\frac{m_i}{z_i}\right) = \sum_j \lambda_j x_{ij}$$

which can also written as

$$\begin{aligned} \log m_i &= \log z_i + \sum_j \lambda_j x_{ij} \,, \\ m_i &= z_i exp(\sum_j \lambda_j x_{ij}) \,, \end{aligned}$$

where the z_i are the fixed a priori cell weights. Sometimes the vector with elements $\log z_i$ is also called the offset matrix.

The specification of a z_i for every cell of the contingency table has several applications. One of its possible uses is in the specification Poisson regression models that take into account the population size or the length of the observation period. This leads to what is called a *log-rate model*, a model for rates instead of frequency counts ([14] and [6]). A rate is a number of events divided by the size of the population exposed to the risk of having the event.

The weight vector can also be used for taking into account sampling or nonresponse weights, in which case the z_i are equated to the inverse of the sampling weights ([3] and [6]). Another use is the inclusion of *fixed effects* in a log-linear model. This can be accomplished by adding the values of the λ parameters which attain fixed values to the corresponding $\log z_i$'s. The last application I will mention is in the analysis of tables with structural zeros, sometimes also called *incomplete tables* ([15]). This simply involves setting the $z_i = 0$ for the structurally zero cells.

3.3 Log-multiplicative model

The log-linear model is one of the GLMs, that is, it is a linear model for the logs of the cell counts in a frequency table. However, extensions of the standard log-linear model have been proposed which imply the inclusion of non-linear terms, the best known example being the log-multiplicative row-column (RC) association models developed by Goodman [13] and Clogg [5] (see also [7]). These RC association models differ from the association models discussed in section 3.1 in that the row and column scores are not a priori fixed, but are treated as unknown parameters which have to be estimated as well. More precisely, a linear-by-linear association is assumed between two variables, given the unknown column and row scores.

Suppose we have a model for a three-way frequency table ABC containing log-multiplicative terms for the relationships between A and B and B and

C. This gives the following log-multiplicative model:

$$\log m_{abc} = +\lambda_a^A + \lambda_b^B + \lambda_c^C + \mu_a^{AB}\phi^{AB}\mu_b^{AB} + \mu_b^{BC}\phi^{BC}\mu_c^{BC}\,. \tag{6}$$

The ϕ parameters describe the strength of the association between the variables concerned. The μ's are the unknown scores for the categories of the variables concerned. As in standard log-linear models, identifying restrictions have to be imposed on the parameters μ. One possible set of identifying restrictions on the log-multiplicative parameters which was also used by Goodman [13] is:

$$\begin{aligned}
&\sum_a \mu_a^{AB} \;=\; \sum_b \mu_b^{AB} \;=\; \sum_b \mu_b^{BC} \;=\; \sum_c \mu_c^{BC} \;=\; 0 \\
&\sum_a \left(\mu_a^{AB}\right)^2 \;=\; \sum_b \left(\mu_b^{AB}\right)^2 \;=\; \sum_b \left(\mu_b^{BC}\right)^2 \;=\; \sum_c \left(\mu_c^{BC}\right)^2 \;=\; 1\,.
\end{aligned}$$

This gives row and column scores with a mean of zero and a sum of squares of one.

On the basis of the model described in Equation 6, both more restricted models and less restricted models can be obtained. One possible restriction is to assume the row and column scores within a particular partial association to be equal, for instance, μ_a^{AB} equal to μ_b^{AB} for all a equal to b. Of course, this presupposes that the number of rows equals the number of columns. Such a restriction is often used in the analysis of mobility tables ([22]). It is also possible to assume that the scores for a particular variable are equal for different partial associations ([5]), for example, $\mu_b^{AB} = \mu_b^{BC}$. Less restricted models may allow for different μ and/or ϕ parameters within the levels of some other variable ([5]), for example, different values of μ_a^{AB}, μ_b^{AB}, or ϕ^{AB} within levels of C. To test whether the strength of the association between the variables father's occupation and son's occupation changes linearly with time, Luijkx [22] specified models in which the ϕ parameters are a linear function of time.

As mentioned above, the RC association models assume a linear-by-linear interaction in which the row and column scores are unknown. Xie [31] demonstrated that the basic principle behind Goodman's RC association models, i.e., linearly restricting log-linear parameters with unknown scores for the linear terms, can be applied to any kind of log-linear parameter. He proposed a general class of log-multiplicative models in which higher-order interaction terms can be specified in a parsimonious way.

4 Regression-, path-, and factor-analytic models

4.1 Log-linear regression analysis: the logit model

In the log-linear models discussed so far, the relationships between the categorical variables are modelled without making a priori assumptions about their 'causal' ordering: no distinction is made between dependent and independent variables. However, one is often interested in predicting the value of a categorical response variable by means of explanatory variables. The logit model is such a 'regression analytic' model for a categorical dependent variable.

Suppose we have a response variable denoted by C and two categorical explanatory variables denoted by A and B. Moreover, assume that both A and B influence C, but that their effect is equal within levels of the other variable. In other words, it is assumed that there is no interaction between A and B with respect to their effect on C. This gives the following logistic model for the conditional probability of C given A and B, $\pi_{c|ab}$:

$$\pi_{c|ab} = \frac{\exp\left(\lambda_c^C + \lambda_{ac}^{AC} + \lambda_{bc}^{BC}\right)}{\sum_c \exp\left(\lambda_c^C + \lambda_{ac}^{AC} + \lambda_{bc}^{BC}\right)}. \tag{7}$$

When the response variable C is dichotomous, the logit can also be written as

$$\begin{aligned} \log\left(\frac{\pi_{1|ab}}{1-\pi_{1|ab}}\right) &= \log\left(\frac{\pi_{1|ab}}{\pi_{2|ab}}\right) \\ &= (\lambda_1^C - \lambda_2^C) + (\lambda_{a1}^{AC} - \lambda_{a2}^{AC}) + (\lambda_{b1}^{BC} - \lambda_{b2}^{BC}) \\ &= \beta + \beta_a^A + \beta_b^B . \end{aligned}$$

It should be noted that the logistic form of the model guarantees that the probabilities remain in the admissible interval between 0 and 1.

It has been shown that a logit model is equivalent to a log-linear model which not only includes the same λ terms, but also the effects corresponding to the marginal distribution of the independent variables ([3], [11], [14]). For example, the logit model described in Equation 7 is equivalent to the following log-linear model

$$\log m_{abc} = \alpha_{ab}^{AB} + \lambda_c^C + \lambda_{ac}^{AC} + \lambda_{bc}^{BC}, \tag{8}$$

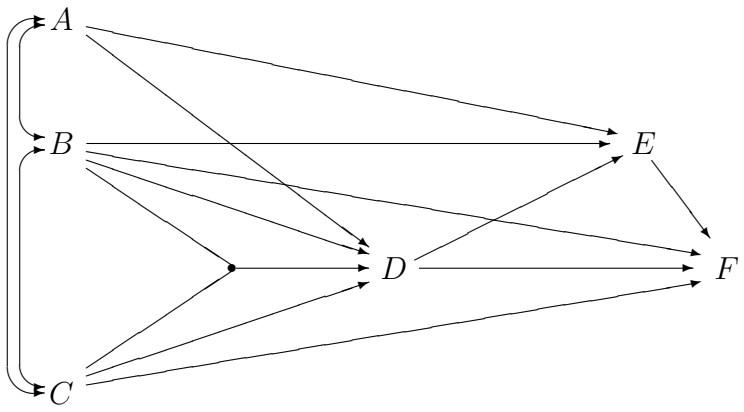

Figure 1: Modified path model

where

$$\alpha_{ab}^{AB} \;=\; \alpha + \lambda_a^A + \lambda_b^B + \lambda_{ab}^{AB}\,.$$

In other words, it equals log-linear model $\{AB, AC, BC\}$ for the frequency table with expected counts m_{abc}. With polytomous response variables, the log-linear or logit model of the form given in Equation 8 is sometimes referred to as a multinomial response model. As shown by Haberman [15], in its most general form, the multinomial response model may be written as

$$\log m_{ik} \;=\; \alpha_k + \sum_j \lambda_j x_{ijk}\,, \tag{9}$$

where k is used as the index for the joint distribution of the independent variables and i as an index for the response variable.

4.2 Log-linear path analysis

After presenting a "regression analytic" extension, we will now discuss a "path-analytic" extension of log-linear analysis introduced by Goodman [12]. As is shown below, his "modified path analysis approach" that makes it possible to take into account information on the causal and/or time ordering between the variables involves specificating a series of logit models.

Suppose we want to investigate the causal relationships between six categorical variables denoted by A, B, C, D, E, and F. Figure 1 shows the assumed causal ordering and relationships between these variables, where a

pointed arrow indicates that variables are directly related to each other, and a 'knot' that there is a higher order interaction. The variables A, B, and C are exogenous variables. This means that neither their mutual causal order nor their mutual relationships are specified. The other variables are endogenous variables, where E is assumed to be posterior to D, and F is assumed to be posterior to E. From Figure 1, it can be seen that D is assumed to depend on A and on the interaction of B and C. Moreover, E is assumed to depend on A, B, and D, and F on B, C, D, and E.

Let $\pi_{def|abc}$ denote the probability that $D = d$, $E = e$, and $F = f$, given $A = a$, $B = b$, and $C = c$. The information on the causal ordering of the endogenous variables is used to decompose this probability into a product of marginal conditional probabilities ([12] and [30]). In this case, $\pi_{def|abc}$ can also be written as

$$\pi_{def|abc} \quad = \quad \pi_{d|abc}\, \pi_{e|abcd}\, \pi_{f|abcde} \ . \tag{10}$$

This is a straightforward way to indicate that the value on a particular variable can only depend on the preceding variables and not on the posterior ones. For instance, E is assumed to depend only on the preceding variables A, B, C, and D, but not on the posterior variable F. Therefore, the probability that $E = e$ depends only on the values of A, B, C, and D, and not on the value of F.

Decomposing the joint probability $\pi_{def|abc}$ into a set of marginal conditional probabilities is only the first step in describing the causal relationships between the variables under study. In fact, the model given in Equation 10 is still a saturated model in which it is assumed that a particular dependent variable depends on all its posterior variables, including all the higher-order interaction terms. A more parsimonious specification is obtained by using a log-linear or logit parameterization for the conditional probabilities appearing in Equation 10 ([12]). While only simple hierarchical log-linear models will be here used, the results presented apply to other kinds of log-linear models as well, including the log-multiplicative models discussed in section 3.3.

A system of logit models consistent with the path model depicted in Figure 1 leads to the following parameterization of the conditional probabilities appearing in Equation 10:

$$\pi_{d|abc} \quad = \quad \frac{\exp\left(\lambda_d^D + \lambda_{ad}^{AD} + \lambda_{bd}^{BD} + \lambda_{cd}^{CD} + \lambda_{bcd}^{BCD}\right)}{\sum_d \exp\left(\lambda_d^D + \lambda_{ad}^{AD} + \lambda_{bd}^{BD} + \lambda_{cd}^{CD} + \lambda_{bcd}^{BCD}\right)} \, ,$$

$$\pi_{e|abcd} = \frac{\exp\left(\lambda_e^E + \lambda_{ae}^{AE} + \lambda_{be}^{BE} + \lambda_{de}^{DE}\right)}{\sum_e \exp\left(\lambda_e^E + \lambda_{ae}^{AE} + \lambda_{be}^{BE} + \lambda_{de}^{DE}\right)},$$

$$\pi_{f|abcde} = \frac{\exp\left(\lambda_f^F + \lambda_{bf}^{BF} + \lambda_{cf}^{CF} + \lambda_{df}^{DF} + \lambda_{ef}^{EF}\right)}{\sum_f \exp\left(\lambda_f^F + \lambda_{bf}^{BF} + \lambda_{cf}^{CF} + \lambda_{df}^{DF} + \lambda_{ef}^{EF}\right)}.$$

As can be seen, variable D depends on A, B, and C, and there is a three-variable interaction between B, C, and D; E depends on A, B, and D, but there are no higher-order interactions between E and the independent variables; and F depends on B, C, D, and E.

4.3 Log-linear factor analysis: the latent class model

As many concepts in the social sciences are difficult or impossible to measure directly, several directly observable variables, or indicators, are often used as indirect measures of the concept to be measured. The values of the indicators are assumed to be determined only by the unobservable value of the underlying variable of interest and by measurement error. In latent structure models, this principle is implemented statistically by assuming probabilistic relationships between latent and manifest variables and by the assumption of local independence. Local independence means that the indicators are assumed to be independent of each other given a particular value of the unobserved or latent variable; in other words, they are only correlated because of their common cause.

Latent structure models can be classified according to the measurement level of the latent variable(s) and the measurement level of the manifest variables. When both the latent and observed variables are categorical one obtains a model called latent class model. As shown by Haberman [15], the latent class model be can defined as a log-linear model with one or more unobserved variables, yielding a "factor-analytic" variant of the log-linear model.

Suppose there is, as depicted in Figure 2, a latent class model with one latent variable W with index w and 4 indicators A, B, C, and D with indices a, b, c, and d. Moreover, let W^* denote the number of latent classes. This latent class model is a equivalent to the hierarchical log-linear model $\{WA, WB, WC, WD\}$; that is,

$$\begin{aligned}\log m_{wabcd} = {} & \lambda + \lambda_w^W + \lambda_a^A + \lambda_b^B + \lambda_c^C + \lambda_d^D \\ & + \lambda_{wa}^{WA} + \lambda_{wb}^{WB} + \lambda_{wc}^{WC} + \lambda_{wd}^{WD}. \end{aligned} \quad (11)$$

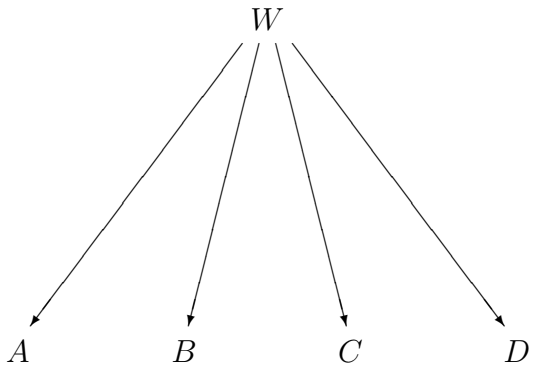

Figure 2: Latent class model

In addition to the overall mean and the one-variable terms, it contains only the two-variable associations between the latent variable W and the manifest variables. As none of the interactions between the manifest variables are included, it can be seen that they are assumed to be conditionally independent of each other given W.

In its classical parameterization proposed by Lazarsfeld [21], the latent class model is defined as

$$\pi_{wabcd} \;=\; \pi_w \pi_{a|w}\, \pi_{b|w}\, \pi_{c|w}\, \pi_{d|w}\,. \tag{12}$$

It can be seen that again the observed variables A, B, C, and D are postulated to be mutually independent given a particular score on the latent variable W. Note that this is in fact a log-linear path model in which one variable is unobserved. The relation between the conditional probabilities appearing in Equation 12 and the log-linear parameters appearing in Equations 11 is

$$\pi_{a|w} \;=\; \frac{\exp\left(\lambda^A_a + \lambda^{WA}_{wa}\right)}{\sum_a \exp\left(\lambda^A_a + \lambda^{WA}_{wa}\right)}. \tag{13}$$

5 Estimation, testing, and software

5.1 Maximum likelihood estimation

Maximum likelihood (ML) estimates for the expected frequencies of a specific log-linear model are most easily derived assuming a Poisson sampling

scheme, but the same estimates are obtained with a multinomial or product-multinomial sampling scheme. Denoting an observed frequency in a three-way table by n_{abc}, the relevant part of the Poisson log-likelihood function is

$$\log \mathcal{L} \;=\; \sum_{abc} \left(n_{abc} \log m_{abc} - m_{abc} \right), \tag{14}$$

where the expected frequencies m_{abc} are a function of the unknown λ parameters.

Suppose we want to find ML estimates for the parameters of the hierarchical log-linear model described in Equation 3. Substituting Equation 3 into Equation 14 and collapsing the cells containing the same λ parameter, yields the following log-likelihood function:

$$\begin{aligned} \log \mathcal{L} \;=\; & n_{+++}\lambda + \sum_{a} n_{a++}\lambda_a^A + \sum_{b} n_{+b+}\lambda_b^B + \sum_{c} n_{++c}\lambda_c^C \\ & + \sum_{ab} n_{ab+}\lambda_{ab}^{AB} + \sum_{bc} n_{+bc}\lambda_{bc}^{BC} \\ & - \sum_{abc} \exp\left(u + \lambda_a^A + \lambda_b^B + \lambda_c^C + \lambda_{ab}^{AB} + \lambda_{bc}^{BC} \right), \end{aligned} \tag{15}$$

where a + is used as a subscript to denote that the observed frequencies have to be collapsed over the dimension concerned. It can now be seen that the observed marginals n_{+++}, n_{a++}, n_{+b+}, n_{++c}, n_{ab+}, and n_{+bc} contain all the information needed to estimate the unknown parameters. Because knowledge of the bivariate marginals AB and BC implies knowledge of n_{+++}, n_{a++}, n_{+b+}, and n_{++c}, n_{ab+} and n_{+bc} are called the minimal sufficient statistics, the minimal information needed for estimating the log-linear parameters of the model of interest.

In hierarchical log-linear models, the minimal sufficient statistics are always the marginals corresponding to the interaction terms of the highest order. For this reason, hierarchical log-linear models are mostly denoted by their minimal sufficient statistics. The model given in Equation 3 may then be denoted as $\{AB, BC\}$, the independence model as $\{A, B, C\}$, and the saturated model as $\{ABC\}$.

When no closed form expression exists for $\hat{m}_{abc}$, ML estimates for the expected cell counts can be found by means of the iterative proportional fitting algorithm (IPF) [8]. Let $\hat{m}_{abc}^{(\nu)}$ denote the estimated expected frequencies after the νth IPF iteration. Before starting the first iteration, arbitrary starting

values are needed for the log-linear parameters that are in the model. In most computer programs based on the IPF algorithm, the iterations are started with all the λ parameters equal to zero, in other words, with all estimated expected frequencies $\hat{m}_{abc}^{(0)}$ equal to 1. For the model in Equation 3, every IPF iteration consists of the following two steps:

$$\begin{aligned} \hat{m}_{abc}^{(\nu)'} &= \hat{m}_{abc}^{(\nu-1)} \frac{n_{ab+}}{\hat{m}_{ab+}^{(\nu-1)}}, \\ \hat{m}_{abc}^{(\nu)} &= \hat{m}_{abc}^{(\nu)'} \frac{n_{+bc}}{\hat{m}_{+bc}^{(\nu)'}}, \end{aligned}$$

where the $\hat{m}_{abc}^{(\nu)'}$ and $\hat{m}_{abc}^{(\nu)}$ denote the improved estimated expected frequencies after imposing the ML related restrictions. The log-linear parameters are easily computed from the estimated expected frequencies.

Finding ML estimates for the parameters of other types of log-linear models is a bit more complicated than for the hierarchical log-linear model because the sufficient statistics are no longer equal to particular observed marginals. Most program solve this problem using a *Newton-Raphson algorithm*. An alternative to the Newton-Raphson algorithm is the *uni-dimensional Newton algorithm*. It differs from the multi-dimensional Newton algorithm in that it adjusts only one parameter at a time instead of adjusting them all simultaneously. In that sense, it resembles IPF. Goodman [13] proposed using the uni-dimensional Newton algorithm for the estimation of log-multiplicative models.

For ML estimation of latent class models, one can make use of an IPF-like algorithm called the Expectation-Maximization (EM) algorithm, a Newton-Raphson algorithm, or a combination of these.

5.2 Model selection

The goodness of fit of a postulated log-linear model can be assessed by comparing the observed frequencies, n, with the estimated expected frequencies, $\hat{m}$. For this purpose, usually two chi-square statistics are used: the likelihood-ratio statistic and the Pearson statistic. For a three-way table, the Pearson chi-square statistic equals

$$X^2 = \sum_{abc} \frac{\left(n_{abc} - \hat{m}_{abc}\right)^2}{\hat{m}_{abc}},$$

and the likelihood-ratio chi-square statistic is

$$L^2 \;=\; 2\sum_{abc} n_{abc}\log\left(\frac{n_{abc}}{\hat{m}_{abc}}\right). \tag{16}$$

The number of degrees of freedom for a particular model is

$$df \;=\; \text{number of cells} - \text{number of independent } u \text{ parameters.}$$

Both chi-square statistics have asymptotic, or large sample, chi-square distributions when the postulated model is true. In the case of small sample sizes and sparse tables, the chi-square approximation will generally be poor. Koehler [18] showed that X^2 is valid with smaller sample sizes and sparser tables than L^2 and that the distribution of L^2 is usually poor when the sample size divided by the number of cells is less than 5. Therefore, when sparse tables are analyzed, it is best to use both chi-square statistics together. When X^2 and L^2 have almost the same value, it is more likely that both chi-square approximations are good. Otherwise, at least one of the two approximations is poor.[2]

The likelihood-ratio chi-square statistic is actually a conditional test for the significance of the difference in the value of the log-likelihood function for two nested models. Two models are nested when the restricted model has to be obtained by only linearly restricting some parameters of the unrestricted model. Thus, the likelihood-ratio statistic can be used to test the significance of the additional free parameters in the unrestricted model, given that the unrestricted model is true in the population. Assuming multinomial sampling, L^2 can be written more generally as

$$\begin{aligned} L^2_{(r|u)} &= \left(-2\,\log\mathcal{L}_{(r)}\right) - \left(-2\,\log\mathcal{L}_{(u)}\right) \\ &= 2\,n_{abc}\log\hat{\pi}_{abc(u)} - 2\,n_{abc}\log\hat{\pi}_{abc(r)} \\ &= 2\,n_{abc}\log\left(\frac{\hat{m}_{abc(u)}}{\hat{m}_{abc(r)}}\right), \end{aligned}$$

where the subscript (u) refers to the unrestricted model and the subscript (r) to the restricted model. Note that in Equation 16, a particular model

[2] An alternative approach is based estimating the sampling distributions of the statistics concerned rather than using their asymptotic distributions. This can be done by bootstrap methods (Langeheine, Pannekoek, and Van de Pol, 1996). These computationally intensive methods are becoming more and more applicable as computers become faster.

is tested against the completely unrestricted model, the saturated model. Therefore, in Equation 16, the estimated expected frequency in the numerator is the observed frequency n_{abc}. The $L^2_{(r|u)}$ statistic has a large sample chi-square distribution if the restricted model is approximately true. The approximation of the chi-square distribution may be good for conditional L^2 tests between non-saturated models even if the test against the saturated model is problematic, as in sparse tables. The number of degrees of freedom in conditional tests equals the number of parameters which are fixed in the restricted model compared to the unrestricted model. The $L^2_{(r|u)}$ statistic can also be computed from the unconditional L^2 values of two nested models,

$$L^2_{(r|u)} = L^2_{(r)} - L^2_{(u)} ,$$

with

$$df_{(r|u)} = df_{(r)} - df_{(u)} .$$

Another approach to model selection is based on information theory. The aim is not to detect the true model but the model that provides the most information about the real world. The best known information criteria are the Akaike [1] information criterion (AIC) and the Schwarz [26] or Bayesian information criterion (BIC). These two measures, which can be used to compare both nested and non-nested models, are usually defined as

$$AIC = L^2 - 2\,df . \tag{17}$$

$$BIC = L^2 - \log N\,df . \tag{18}$$

5.3 Software

Software for log-linear analysis is readily available. Major statistical packages such as SAS and SPSS have modules for log-linear analysis that can be used for estimating hierarchical and general log-linear models, log-rate models, and logit models. Special software is required for estimating log-multiplicative models, log-linear path models, and latent class models. The command language based ℓ_{EM} program developed by Vermunt [27][28] can deal with any of the models discussed in this article, as well as combinations of these. Vermunt and Magidson's [29] Windows based Latent GOLD can deal with certain types of log-linear models, logit models, and latent class models, as well as combinations of logit and latent class models.

6 An application

Consider the four-way cross-tabulation presented in Table 1 containing data taken from four annual waves (1977-1980) of the National Youth Survey [9]. The table reports information on marijuana use of 237 respondents who were age 14 in 1977. The variable of interest is an ordinal variable measuring marijuana use in the past year. It has the three levels "never" (1), "no more than once a month" (2), and "more than once a month" (3). We will denote these four time-specific measures by A, B, C, and D, respectively.

[INSERT TABLE 1 ABOUT HERE]

Several types of log-linear models are of interest for this data set. First, we might wish to investigate the overall dependence structure of these repeated responses, for example, whether it is possible to describe the data by a hierarchical log-linear model containing only the two-way associations between consecutive time points; that is, by a first-order Markov structure. Second, we might want to investigate whether it is possible to simplify the model by making use of the ordinal nature of the variables using uniform or RC association structures. Third, latent class analysis could be used to determine whether it is possible to explain the associations by assuming that there is a small number of groups of children with similar developments in marijuana use.

Table 2 reports the L^2 values for the estimated models. Because the asymptotic p values are unreliable when analyzing sparse frequency tables such as the one we have here, we estimated the p values by means of 1000 parametric bootstrapping replications. The analysis was performed with the Latent GOLD program.

[INSERT TABLE 1 ABOUT HERE]

The high L^2 value obtained with Model 1 – the independence model $\{A, B, C, D\}$ – indicates that there is a strong dependence between the 4 time-specific measures. Model 2 is the model with all 2-variable associations: $\{AB, AC, AD, BC, BD, CD\}$. As can be seen from its p value, it fits very well, which indicates that higher-order interactions are not needed. The Markov model containing only associations between adjacent time points – Model 3: $\{AB, BC, CD\}$ – seems to be too restrictive for this data set. It

turns out that we need to include one additional term; that is, the association between the second and fourth time point, yielding $\{AB, BC, BD, CD\}$ (Model 4).

Model 5 has the same structure as Model 4, with the only difference that the two-variable terms are assumed to be uniform associations (see Equation 5). This means that each two-way association contains only one instead of four independent parameters. These "ordinal" constraints seems to be too restrictive for this data set.

Models 6 and 7 are latent class models or, equivalently, log-linear models of the form $\{XA, XB, XC, XD\}$, where X is a latent variable with either two or three categories. The fit measures indicate that the associations between the time points can be explained by the existence of three types of trajectories of marijuana use.

Based on the comparison of the goodness-of-fit measures for the various models, as well as their AIC values that also take into account parsimony, one can conclude that Model 4 is the preferred one. The three-class, however, yields a somewhat simpler explanation for the associations between the time-specific responses.

References

[1] Akaike, H. (1987). Factor analysis and AIC. *Psychometrika*, 52, 317-332.

[2] Bishop, R.J., Fienberg, S.E., and Holland, P.W. (1975). *Discrete multivariate analysis: theory and practice*. Cambridge, Mass.: MIT Press.

[3] Agresti, A. (1990). *Categorical data analysis*. New York: Wiley, second edition 2002.

[4] Agresti, A. (1993). Computing conditional maximum likelihood estimates for generalized Rasch models using simple loglinear models with diagonal parameters. *Scandinavian Journal of Statistics*, 20, 63-71.

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

[6] Clogg, C.C., and Eliason, S.R. (1987). Some common problems in log-linear analysis. *Sociological Methods and Research*, 16, 8-14.

[7] Clogg, C.C., and Shihadeh, E.S. (1994). *Statistical models for ordinal data*. Thousand Oakes, CA: Sage Publications.

[8] Darroch, J.N., and Ratcliff, D. (1972). Generalized iterative scaling for log-linear models. *The Annals of Mathematical Statistics*, 43, 1470-1480.

[9] Elliot, D.S., Huizinga, D., and Menard, S. (1989). *Multiple problem youth: delinquence, substance use, and mental health problems*. New York: Springer-Verlag.

[10] Evers, M., and Namboodiri, N.K. (1978). On the design matrix strategy in the analysis of categorical data. K.F. Schuessler (ed.), *Sociological Methodology 1979*, 86-111. San Fransisco: Jossey Bass.

[11] Goodman, L.A. (1972). A modified multiple regression approach for the analysis of dichotomous variables. *American Sociological Review*, 37, 28-46.

[12] Goodman, L.A. (1973). The analysis of multidimensional contingenty tables when some variables are posterior to others: a modified path analysis approach. *Biometrika*, 60, 179-192.

[13] 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.

[14] Haberman, S.J. (1978). *Analysis of qualitative data, Vol. 1, Introduction topics*. New York, San Francisco, London: Academic Press.

[15] Haberman, S.J. (1979). *Analysis of qualitative data, Vol 2, New developments*. New York: Academic Press.

[16] Hagenaars, J.A. (1990). *Categorical longitudinal data - loglinear analysis of panel, trend and cohort data.*. Newbury Park: Sage.

[17] Kelderman, H. (1984). Log-linear Rasch model tests. *Psychometrika*, 49, 223-245.

[18] Koehler, K.J. (1986). Goodness-of-fit statistics for log-linear models in sparse contingency tables. *Journal of the American Statistical Association*, 81, 483-493.

[19] Laird, N., and Oliver, D. (1981). Covariance analysis of censored survival data using log-linear analysis techniques. *Journal of the American Statistical Association*, 76, 231-240.

[20] Langeheine, R., Pannekoek, J., and Van de Pol, F. (1996). Bootstrapping goodness-of-fit measures in categorical data analysis. *Sociological Methods and Research*, 24, 492-516.

[21] Lazarsfeld, P.F. (1950). The logical and mathematical foundation of latent structure analysis. S.A. Stouffer et al. (eds.), *Measurement and Prediction*, 362-412. Princeton, NJ: Princeton University Press.

[22] Luijkx, R. (1994). *Comparative loglinear analyses of social mobility and heterogamy*. Tilburg: Tilburg University Press.

[23] McCullagh, P., and Nelder, J.A. (1983). *Generalized linear models*. London: Chapman & Hall, second edition 1989.

[24] Mellenbergh, G.J., and Vijn, P. (1981). The Rasch model as a log-linear. *Applied Psychological Measurement*, 5, 369-376.

[25] Rindskopf, D. (1990). Nonstandard loglinear models. *Psychological Bulletin*, 108, 150-162.

[26] Schwarz, G. (1978). Estimating the dimensions of a model. *Annals of Statistics*, 6, 461-464.

[27] Vermunt J.K. (1997). Log-linear models for event history histories. *Advanced Quantitative Techniques in the Social Sciences Series, Volume 8*. Thousand Oakes: Sage Publications.

[28] Vermunt, J.K. (1997) *LEM: A general program for the analysis of categorical data. User's manual*. Tilburg University, The Netherlands.

[29] Vermunt, J.K., Magidson, J. (2000). Latent GOLD 2.0 User's Guide. Belmont, MA: Statistical Innovations Inc.

[30] Wermuth, N., and Lauritzen, S.L. (1983). Graphical and recursive models for contingency tables. *Biometrika*, 70, 537-552.

[31] Xie, Yu (1992). The log-multiplicative layer effects model for comparing mobility tables. *American Sociological Review*, 57, 380-395.

Table 1: Data on marijuana use in the past year taken from four yearly waves of the National Youth Survey (1977-1980)

		1979 (C)								
		1			2			3		
1977	1978	1980 (D)			1980 (D)			1980 (D)		
(A)	(B)	1	2	3	1	2	3	1	2	3
1	1	115	18	7	6	6	1	2	1	5
1	2	2	2	1	5	10	2	0	0	6
1	3	0	1	0	0	1	0	0	0	4
2	1	1	3	0	1	0	0	0	0	0
2	2	2	1	1	2	1	0	0	0	3
2	3	0	1	0	0	1	1	0	2	7
3	1	0	0	0	0	0	0	0	0	1
3	2	1	0	0	0	1	0	0	0	1
3	2	0	0	0	0	2	1	1	1	6

Table 2: Goodness-of-fit statistics for the estimated models for the data in table 1

Model		L^2	*df*	$\widehat{p}$	AIC
1.	Independence	403.3	72	.00	259.3
2.	All two-variable terms	36.9	48	.12	-59.1
3.	First-order Markov	58.7	60	.05	-61.3
4.	Model 3 + $\lambda_{j\ell}^{BD}$	41.6	56	.30	-70.4
5.	Model 4 with uniform associations	83.6	68	.00	-52.4
6.	Two-class latent class model	126.6	63	.00	-51.0
7.	Three-class latent class model	57.0	54	.12	-56.2