STATISTICS IN TRANSITION-new series, Summer 2012
Vol. 13, No. 2, pp. 311—320

INDEPENDENCE ANALYSIS OF NOMINAL DATA WITH THE USE OF LOG-LINEAR MODELS IN R

Justyna Brzezińska[1]

ABSTRACT

Log-linear models are used to analyze the relationship between two or more categorical (e.g. nominal or ordinal) variables. The term log-linear derives from the fact that one can, through logarithmic transformations, restate the problem of analyzing multi-way frequency tables in terms that are very similar to ANOVA. Specifically, one may think of the multi-way frequency table to reflect various main effects and interaction effects that add together in a linear fashion to bring about the observed table of frequencies. There are several types of models between dependence and independence: homogenous association, partial association, conditional association and null model. Expected cell frequencies are obtained with the use of iterative proportional fitting algorithm (IPF) [Deming, Stephen 1940]. The next step is to derive model coefficients for single variables as well as for interaction parameter and the most useful tool for interpreting model parameter is odds and odds ratio. Log-linear models are available in **R** software with the use of `loglm` function in `MASS` library and `glm` function in `stats` library. In this paper log-linear analysis will be presented with the use of available packages on empirical datasets in economic area.

Key words: Log-linear models, cross-tabulation, qualitative data, independence analysis of nominal data.

1. Introduction

Frequency counts of categorical variables are probably the most frequently encountered variables of research. Categorical data analysis has a long history. First papers were published in 1900 [Pearson, Yule] and were focused on independence analysis for categorical variables in two-way tables. Later on it was developed into multi-way contingency tables [Haberman 1974, Bishop, Fienberg Holland 1975]. In the middle of the twentieth century log-linear analysis for nominal [Bishop, Fienberg, Holland 1975, Knoke, Burke 1980, Christensen 1997] and ordinal data [Ishii-Kunts 1994] developed successfully. Nowadays, with the

[1] University of Economics in Katowice.

use of speed computers and professional software, advanced techniques are used for visualizing data structure [Friendly 2000, Zeileis, Mayer, Hornik 2006].

Cross-classification tables are used to answer questions such as whether the relation exists between categorical variables or whether the relation between variable is different for different groups of subject. The usual method for analyzing cross-classified tables, no matter how many variables are considered, is to test relations between variables taken one pair at a time. Most often this is accomplished using the chi-square, Yule`s, Cramer`s and Pearson`s coefficient, however, for higher-way tables more general and wider method should be used. A more general strategy for the analysis of cross-classified categorical data involves testing several models, including not only the model of independence but also models that represent various types of association and interactions included. This strategy is called log-linear analysis and it is one of the most interesting tool for categorical data analysis. We do not differentiate between dependent and independent variable and all response variables are treated as factors. They can be used as an explorative model-building fashion to find the most parsimonious model that describes the data best, as well as for hypothesis testing with simultaneous testing of all possible factors combination.

This paper is concerned with the independence type in log-linear analysis and its application in economic research. It explains the appropriate use of the analysis, describes briefly the calculations involved, and finally illustrates applications of the method based on empirical set of data with the use of **R**.

2. Independence analysis and log-linear models for nominal data

The analysis of cross-classified categorical data has occupied a prominent place in statistics, but most of techniques were associated with the analysis of two-dimensional contingency tables and the calculation of chi-square or related statistics. During last years, the availability of high-speed computers has led to major development in the analysis of multidimensional tables. However, the excellent guides are already available [Bishop, Fienberg, Holland 1975, Cox 1970, Haberman 1974, Lindsey 1973, Plackett 1974], the analysis of higher dimensional tables has fully developed much later.

Categorical data consists of variable whose values comprise a set of discrete categories. Such data requires different statistical methods from those commonly used for quantitative data. The aim of this paper is to provide theoretical and practical methods designed to reveal patterns of relationship among nominal variable with the use of log-linear analysis. The term log-linear derives from the fact that one can, through logarithmic transformations, restate the problem of analyzing multi-way frequency tables in terms that are very similar to ANOVA. Specifically, one may think of the multi-way frequency table to reflect various main effects and interaction effects that add together in a linear fashion to bring about the observed table of frequencies. Log-linear models are used for modelling

cell counts in multi-way contingency tables and it allows one to distinguish several types of association: conditional independence model, joint independence model, homogenous association model, partial independence model, saturated or null model.

In a three-way table with nominal variable X, Y and Z, several types of potential independence can be presented, for example homogenous association, joint independence, conditional independence, partial independence, etc. Saturated model includes all the possible effects in multiplicative form for three variables given as:

$$m_{hjk} = \eta \tau_h^X \tau_j^Y \tau_k^Z \tau_{hj}^{XY} \tau_{hk}^{XZ} \tau_{jk}^{YZ} \tau_{hjk}^{XYZ} . \quad (1)$$

By taking the natural logarithms we have additive equation given as:

$$\log\left(m_{hjk}\right) = \lambda + \lambda_h^X + \lambda_j^Y + \lambda_k^Z + \lambda_{hj}^{XY} + \lambda_{hk}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{hjk}^{XYZ} . \quad (2)$$

Saturated model reproduces perfectly the observed cell frequencies through the theoretical frequencies but such model is meaningless since the aim is to find a more parsimonious model with less parameters. In order to find the best model from a set of possible models, some additional measures can also be considered. Fitting a log-linear model is a process of deciding which association terms are significantly different from zero. These terms are included in the model that is used to explain the observed frequencies. Terms that are excludes from the model go into the residual or error term, which reflects the overall badness-of-fit of the model [Friendly 2000]. The goal of the analysis is to find a small model (with fewer association terms) that nonetheless achieves a reasonable fit.

A rule of thumb to determine the degrees of freedom is df = number of cells – number of free parameters [Agresti 2002]. With the use of backward elimination method the starting point is saturated model. Thus, the aim of a researcher is to find a reduced model. A reduced model is a more parsimonious model with fewer parameters and thus fewer dependencies and effects. The hierarchy principle reveals that a parameter of lower order cannot be removed when there is still a parameter of higher order that concerns at least one of the same variables.

The most useful statistic used to test the goodness of fit for log-linear model is the likelihood ratio statistic [Christensen 1997]:

$$G^2 = 2\sum_{h=1}^{H}\sum_{j=1}^{J}\sum_{k=1}^{K} n_{hjk} \ln\left(\frac{n_{hjk}}{m_{hjk}}\right). \quad (3)$$

Therefore, larger G^2 values indicate that the model does not fit the data well and thus the model should be rejected.

Akaike information criterion [Akaike 1973] refers to the information contained in a statistical model according to equation:

$$AIC = G^2 - 2df . \quad (4)$$

Another information criterion is Bayesian information criterion [Raftery 1986]:

$$BIC = G^2 - df \cdot \ln n . \tag{5}$$

The model that minimizes AIC and BIC will be chosen.

The likelihood ratio can also be used to compare an overall model within a smaller, nested model. The equation is as follows:

$$\Delta G^2 = G_1^2 - G_2^2 , \tag{6}$$

with: $\Delta df = df_2 - df_2$ degrees of freedom. If the ΔG^2 comparison statistic is not significant, then the nested model (1) is not significantly worse than the saturated model (2). There are several models which appear to provide an adequate fit, therefore, the more parsimonious (nested) model will be chosen. Models are getting reduced through the hierarchy principle. The hierarchy principle reveals that a parameter of lower order cannot be removed when there is still a parameter of higher order that concerns at least one of the higher orders [Knoke, Burke 1980]. We obtain MLEs for elementary cells under any hierarchical model by iterative proportional fitting of the sufficient configuration [Bishop, Fienberg, Holland 1975]. An acceptable model is the one whose expected cell frequencies do not significantly differ from the observed data. Although the model has been described for three-dimensional tables, the extension to higher dimensional tables is straightforward.

3. Application in R

Empirical example presented in this paper is based on housing market dataset on Copenhagen housing conditions (library(MASS), data(housing)) with sample size of 1681. Multi-way table contains four categorical data [Cox, Snell 1984, Madsen 1976] (Table 1).

Table 1. Variables in the analysis

Variable	Factor levels
Sat (S)	Satisfaction of householders with their present housing circumstances (High, Medium. Low)
Infl (I)	Perceived degree of influence householders have on the management of the property (High, Medium, Low)
Type (T)	Type of rental accommodation (Tower, Atrium, Apartment, Terrace)
Cont (C)	Contact residents are afforded with other residents (Low, High)

*Source: **R** software, library(MASS), data(housing).*

We will use the loglm() function in the library(MASS) to fit log-linear models. From a different perspective equivalent models can also be fit as a generalized linear models with the use of glm(…,family=poisson) function in stats package.

Suppose for a four-dimensional table $H \times J \times K \times L$ with the total sample size N and that the cell count for $hjkl-$cell is n_{hjkl}. Given the large number of possible hierarchical models that can be fit to multidimensional table, it is reasonable to ask whether a systematic approach to model selection is possible. Many different approaches has been proposed, but none of them entirely satisfactory [Fienberg 1980]. In this paper stepwise procedures will be presented with a significance level 0.15. First, following the hierarchy principle and starting from saturated model $[SITC]$, the goodness of fit statistic G^2 with corresponding $p-$value are computed for one-, two- and three-ordered level of interaction (Table 2).

Table 2. Goodness of fit statistics

Model	G^2	df	$p-$value
S.I.T.C	295.352	63	0,000
SI.ST.SC.IT.IC.TC	43.9518	40	0.308
SIT.SIC.STC.ITC	5.94433	12	0.919

Source: own calculations in ***R****.*

Concerning $\alpha-$level of 0.15 only models $[SI][ST][SC][IT][IC][TC]$ and $[SIT][SIC][STC][ITC]$ (and the saturated model) fit the data. However, the likelihood ratio can be used for comparing two nested models against each other. Comparison of nested models can be done using ANOVA method.

```
Deviance df Delta(Dev) Delta(df) P(> Delta(Dev)
Model 1   43.951777 40
Model 2    5.944334 12  38.007443        28        0.09826
Saturated  0.000000  0   5.944334        12        0.91886
```

According to the output, the only model that fits data well is model `Model 2` called model of conditional independence $[SIT][SIC][STC][ITC]$. Its deviance is close enough to the deviance for the saturated model to give a non-significant ($p > 0.15$) $p-$value. However, we would prefer a simpler model than all three-way interactions. It is possible to fit all-pairwise models with the addition of one three-way interaction $[SIT]$, $[SIC]$, $[STC]$ and finally $[ITC]$.

Table 3. Goodness of fit statistics

Model	G^2	df	$p-$value
SI.ST.SC.IT.IC.TC.SIT	22.132	82	0.775
SI.ST.SC.IT.IC.TC.SIC	42.286	36	0.218
SI.ST.SC.IT.IC.TC.STC	31.783	34	0.577
SI.ST.SC.IT.IC.TC.ITC	38.662	34	0.267

Source: own calculations in ***R****.*

For all possible models with two-way interaction plus one three way interaction $p-$value is non-significant, and those models are simpler than three-way interactions model, so it is very difficult to choose the best fitting model. The comparison of goodness of fit statistics for all good models will be done again with the use of ANOVA method.

```
           Deviance df Delta(Dev) Delta(df) P(> Delta(Dev)
Model 1   43.951777 40
Model 2   42.285862 36   1.665915         4        0.79690
Model 3   31.783060 34  10.502802         2        0.00524
Model 4   38.662216 34  -6.879155         0        1.00000
Model 5   22.131810 28  16.530406         6        0.01117
Model 6    5.944334 12  16.187476        16        0.43995
Saturated  0.000000  0   5.944334        12        0.91886
```

Concerning $\alpha-$level of 0.15 and following the hierarchy principle, the difference in fit of model 1 and model 2 is non-significant (P(> Delta(Dev)= 0.79690); moreover Delta(df)=4 is close enough to Delta(df)= 1.6659 so model 1 $[SI][ST][SC][IT][IC][TC]$ as mode parsimonious, nested model should be the appropriate model fitting best. It means that this model includes the effect of single variables, as well as all possible to-way interactions on cell counts. It also means that none of three-way interaction $[SIT]$, $[SIC]$, $[STC]$ or $[ITC]$ have influence on cell counts. Satisfaction of householders with their present housing circumstances (S) depends on perceived degree of influence householders have on the management of the property (I), type of rental accommodation (T) and contact residents are afforded with other residents (C) individually, what is proved by interactions $[SI][ST][SC]$. Furthermore, perceived degree of influence householders have on the management of the property (I) depends on type of rental accommodation (T) and contact residents are afforded with other residents (C) individually: $[IT][IC]$. Finally, the type of rental accommodation (T) depends on contact residents are afforded with other residents (C): $[TC]$. The model equation for relationship $[SI][ST][SC][IT][IC][TC]$ can be described as:

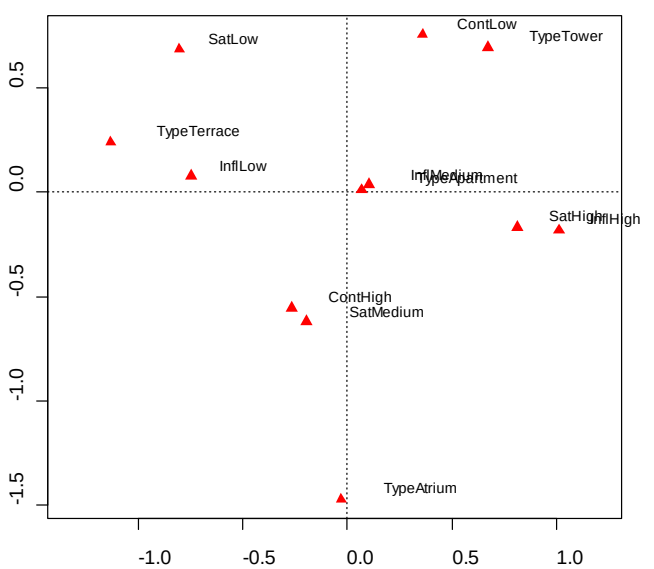
$$\log\left(m_{hjkl}\right)=\lambda+\lambda_h^S+\lambda_j^I+\lambda_k^T+\lambda_l^C+\lambda_{hj}^{SI}+\lambda_{hk}^{ST}+\lambda_{hl}^{SC}+\lambda_{jk}^{IT}+\lambda_{jl}^{IC}+\lambda_{kl}^{TC}. \qquad (7)$$

Sometimes it is convenient to interpret only the model equation and possible interactions, as with many variables included it might be difficult to interpret every single parameter for every cell. Parameters give information on each cell, if modelled cell is greater or smaller than empirical cell.

Another method widely used for categorical data analysis is correspondence analysis (CA) for two categorical variables and multiple correspondence analysis (MCA) for several variables. The method is particularly helpful in analyzing cross-tabular data in the form of numerical frequencies, and results in an elegant but simple graphical display which permits more rapid interpretation and understanding of data [Greenacre 1993]. Correspondence analysis is a statistical method for picturing the associations between the levels of a two- or multi-way contingency table. The observed association is done by cell frequencies, and typical inferential aspect is the study of whether certain levels of one characteristic are associated with some levels of another. Correspondence analysis is a geometric technique for displaying the rows and columns of a contingency table as a points in low-dimensional space. The main goal of the method is a global view of the data that is useful for interpretation with the use of perception map.

Symmetric map of the data housing, using the multiple correspondence analysis (MCA) in `ca` package is presented below.

Figure 1. Perception map: multiple correspondence analysis (Inertia=0.024)

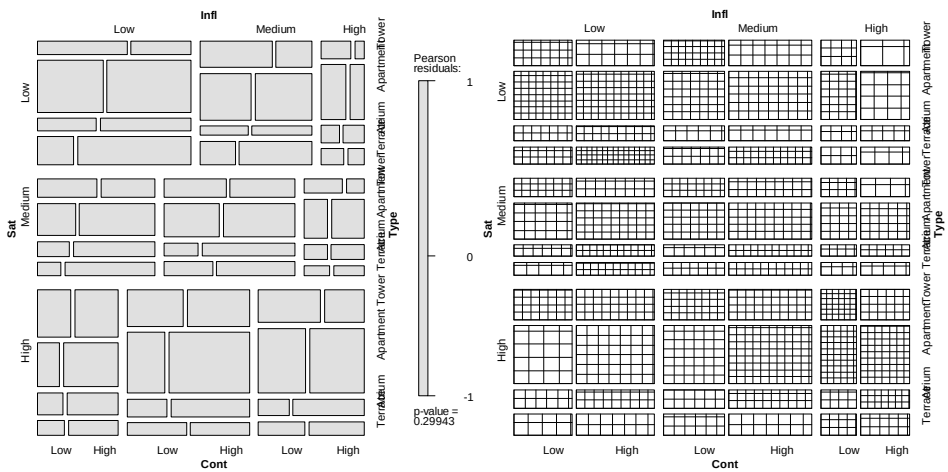


*Source: own calculations in **R**.*

Total inertia is 0.024, so there is no relationship between variables. It is also seen that it would be quite difficult to group the points into well separated groups or clusters. Eigenvalue for first and second axis are: 59% and 8.2%, and it means that over 67.2% of the association between satisfaction of householders with their present housing circumstances (S), influence householders have on the management of the property (I), type of rental accommodation (T) and contact residents are afforded with other residents (C) can be represented well in two dimensions. This perception map might be helpful, when it is difficult to see any type of relationship, especially for multi-way tables. However, in this case log-linear analysis provided much more detailed information on data structure and pattern of the relationship.

With the use of vcd package it is also possible to present the data structure graphically with the use of mosaic and sieve plot [Friendly 2000].

Figure 2. Mosaic plot for housing data. **Figure 3.** Sieve plot for housing data.



*Source: own calculations in **R**.*

The mosaic plot (Figure 2) and sieve plot (Figure 3) indicate that the model $[SI][ST][SC][IT][IC][TC]$ fits well (residuals in mosaic plot, as well as differences between observed and theoretical counts in sieve plot, are small).

The advantage of log-linear models is that they can be used for any number of categorical variables and there are no limits with the table dimension. Different types of association can be distinguished which gives more detail information about relation between variables and visualization methods can be applied (ca, vcd, vcdExtra).

4. Conclusions

Log-linear analysis is a method of statistical analysis that is used when all the variables of interest are categorical. The method has had wide use in the economic, psychological and social sciences and has several advantages. In log-linear analysis there is no dependent variable that can be predicted, instead cell frequencies are modelled. Log-linear models can easily incorporate more than two categorical variables and are useful for patterns of association analysis. We can also use them to identify different types of independence.

The purpose of this paper was to provide an outline of log-linear models and its application in economic research. The `loglm` package in **R** was used as well as multiple correspondence analysis (`ca`) and visualizing tools for multi-way tables (`vcd`, `vcdExtra`) to facilitate interpretation of results. In the paper the log-linear analysis was presented using dataset Copenhagen housing conditions. The advantages to be gained from the model-fitting techniques are that they provide a systematic approach to the analysis of multidimensional tables and also estimates of the magnitude of effects interest. The method allows for distinguishing different types of association: saturated model, complete (mutual) independence, joint independence, conditional independence or homogeneous association. However, Agresti [2002] says: "*there is no guarantee that either strategy will lead to a meaningful model*". We should also look for a model that is simple to interpret and smoothes rather than overfits the data.

REFERENCES

AGRESTI, A., 2002. Categorical Data Analysis, Wiley & Sons, Hoboken, New Jersey.

AKAIKE, H., 1973. Information theory and an extension of the maximum likelihood principle, in: Proceedings of the 2nd International Symposium on Information, Petrow B. N., Czaki F., Budapest: Akademiai Kiado.

BISHOP, Y. M. M., FIENBERG E. F., HOLLAND P. W., 1975. Discrete Multivariate Analysis, MIT Press, Cambridge, Massachusetts.

CHRISTENSEN, R., 1997. Log-Linear Models and Logistic Regression, Springer–Verlag, New York.

COX, D. R., 1970. Analysis of Binary Data, London, Mathuen.

COX, D. R. and SNELL, E. J., 1984. Applied Statistics, Principles and Examples. Chapman & Hall.

FIENBERG, S., 1980. The analysis of cross-classified categorical data, MIT Press, Cambridge.

FRIENDLY, M., 2000. Visualizing categorical data, SAS Institute Inc.

GREENACRE, M. J., 1993. Correspondence analysis in practice, London Academic Press.

HABERMAN, S. J., 1974. The Analysis of Frequency Data, Chicago, University of Chicago Press.

HORNIK, K., MAYER D., ZEILEIS A., 2006. The strucplot framework: visualizing multi-way contingency tables with vcd, Journal of Statistical Software, 17 (3), 1-48.

ISHII-KUNTS, M., 1994. Ordinal log-linear models, Sage University Papers.

KNOKE, D., BURKE P. J., 1980. Log–linear Models, Quantitative Applications in the Social Science" 20, Sage University Papers, Sage Publications, Newbury Park, London, New Delhi.

LINDSEY, J. K., 1973. Inferences from Sociological Survey Data: A Unified Approach, New York, Elsevier.

MADSEN, M., 1976. Statistical analysis of multiple contingency tables. Two examples. Scand. J. Statist. 3, 97–106.

PEARSON, K., 1900. On a criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling, Philos. Mag. Ser. 5, 50, 157-175.

PLACKETT, R. L., 1974. The Analysis of Categorical Data, London, Griffin.

RAFTERY, A. E., 1986. A note on Bayesan Factors for log-linear contingency table models with vague prior information, Journal of the Royal Statistical Society, Ser. B, 48, 249-250.

YULE, G. U., 1900. On the association of attributes in statistics, Phil. Trans. Ser. Q 194, 257-319.