ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 162, 2002

*Agnieszka Rossa**

ON THE BACKWARD SELECTION PROCEDURE FOR GRAPHICAL LOG-LINEAR MODELS – MONTE CARLO RESULTS

ABSTRACT. The analysis of categorical data by means of log-linear models is one of the most useful statistical tools available, particularly in the social and medical sciences, thus in all the sciences where we deal with collection of large amounts of qualitative data. They are also widely applied in expert systems (see L a u r i t z e n and S p i e g e l h a l t e r (1988), M a t z k e v i c h and A b r a m s o n (1995)).

Qualitative data are often analysed by cross-classifying two variables at a time only, i.e. examining all the two way marginal tables of the underlying multidimensional table. It is well known that this approach may often produce misleading results. The analysis of multidimensional contingency tables by means of log-linear models allows to avoid most of such problems. However, the number of possible log-linear model for multidimensional tables is so large that one must use some form of stepwise selection strategy to chose a model, which fits to the data and satisfies some additional conditions. In the paper some statistical properties of the backward selection procedure by means of Monte Carlo methods are studied.

Key words: Graphical log-linear models, model fitting procedure, Monte Carlo study.

I. HIERARCHICAL LOG-LINEAR MODEL FOR 5-WAY CONTINGENCY TABLE

In the paper we consider 5-way table $\{n_{ijklr}\}$ formed for 5 categorical variables, say, A, B, C, D and E. A log-linear model specifies a linear relation between the expected cell counts m_{ijklr} and some unknown parameters, for example

$$\log m_{ijklr} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_l^D + \lambda_r^E + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{lr}^{DE},$$

where the parameters (called interactions) satisfy certain conditions, e. g.

* Dr., Institute of Econometrics and Statistics, University of Łódź.

$$\sum_i \lambda_i^A = \sum_j \lambda_j^B = \ldots = \sum_r \lambda_r^E = \sum_i \lambda_{ij}^{AB} = \sum_j \lambda_{ij}^{AB} = \ldots = \sum_l \lambda_{lr}^{DE} = \sum_r \lambda_{lr}^{DE} = 0.$$

In practice one considers only hierarchical log-linear models, i.e. the ones in which presence of some interaction term, for example λ_{ij}^{AB} implies presence of all terms marginal to it, here λ_i^A and λ_j^B. It is easy to see that each such model contains a set of terms which are not marginal to any other terms in the model, and this set defines the model: this set is called the generating class. It the above model it is λ_{ij}^{AB}, λ_{ik}^{AC} and λ_{lr}^{DE}. This enables us to write the model in the alternative notation $\{AB\}\{AC\}\{DE\}$, that is by direct specification of the generating class. It can be shown that the model can be interpreted as saying that two-dimensional variable $\{DE\}$ is independent on variables $\{ABC\}$ and that variables B and C are conditionally independent given A. We can use Dawid's notation for these relations is: $\{DE\} \perp \{ABC\}$ and $B \perp C | A$ (D a w i d 1979).

II. GRAPHICAL LOG-LINEAR MODEL

Consider an undirected graph, that is a set of vertices and edges. The graph can be associated with a log-linear model. For example, the graph associated with the model $\{AB\}\{AC\}\{DE\}$ is given in Fig. 1.

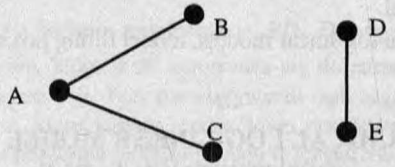

Fig. 1. Interaction graph for 5-way contingency table

Vertices in this graph correspond to main effects and edges correspond to the two-factor interactions present in the model. Such a graph is called interaction graph (D a r r o c h, L a u r i t z e n et al. 1980).

We call a set of vertices complete if all possible edges between the vertices in the set are in the graph. For example in Fig. 2 $\{ABC\}$, $\{ACE\}$, $\{ABD\}$, $\{ADE\}$ are complete, whereas $\{BDE\}$, $\{CDE\}$ are not complete since the edges BE and DC are not in the graph.

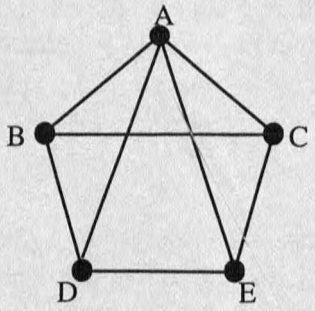

Fig. 2. Interaction graph for 5-way contingency table

We define a clique – a maximal complete set, i.e. a complete set which cannot be extended to a larger complete set by the addition of more vertices. The log-
-linear model is called a graphical model if its generating class is the set of all cliques of the corresponding graph. Graphical models can be understood purely in terms of independence and conditional independence relationships. That B and C are conditionally independent given A or $\{DE\}$ is independent on $\{ABC\}$ can be read directly off the graph (see Fig. 1). Thus, the attractive feature of graphical models is that they are easy to interpret. In many applications these relationships may be easily understood in terms of causality.

III. MODEL SELECTION PROCEDURE

In the paper we consider the problem of selecting graphical log-linear models for tables of counts cross-classified by 5 categorical variables and collected under multinomial sampling scheme. There is a number of graphical log-linear models for 5-way contingency table and it is necessary to use an exploratory procedure to select a model which fits the data (see also B e n e d e t t i, B r o w n 1978). Restricting model selection to graphical models has several practical consequences. Firstly model selection is easier, because the number of models under consideration is reduced. Secondly, graphical models characterise conditional independence relationships.

Most methods proposed for model selection consist of two phases: choice of an initial base model, and stepwise improvement from the base model. The stepwise improvement from the base model can involve both the backward selection, i.e. removal of non-significant interaction terms, and the forward selection, addition significant terms. The test statistics used usually in such selection procedures is either Pearson's χ^2 statistic or likelihood ratio statistic. In the paper

the backward selection procedure is considered with the goodness-of-fit statistics χ^2 employed. The most common Pearson's χ^2 is defined as

$$\chi^2 = \sum_i \frac{(O_i - m_i)^2}{m_i},$$

where O_i is the observed counts in the i-th cell of the table and m_i is the expected cell count under hypothesis. It is well known that when m_i are not small the χ^2 statistic is distributed approximately as chi-squared variable. But when the table under analysis is large and sparse, with many zeros both in the body of the table and in the marginal totals, the distribution of the test statistics does not follow their prescribed asymptotic form. There is wide difference of opinion how small the m_i can be without invalidating the chi-squared approximation. The aim of this paper is analysis of properties of the backward selection test procedure especially for large and sparse contingency tables. Results of Monte Carlo experiments are presented in the next section.

IV. SIMULATION RESULTS FOR 5-WAY CONTINGENCY TABLES

In order to generate sample of size N from multinomial distribution, random numbers were generated from the unit interval (0, 1). These observations were next inserted into s subintervals associated with s cells of the contingency table. The length of each subinterval of the unit interval was equal to the assumed cell probability. It was assumed that the probabilities reflect a known pattern of association for 5 variables and the underlying graphical log-linear model. Observations falling in the respective intervals were then enumerated and these counts was entered in the resulting contingency table. The sampling process was repeated 10 000 times for various underlying models, and various sample sizes. Each final model obtained in the testing process that did not fit the underlying model effected the level of incorrect fitting. Typical results of the simulations obtained for four various 5-way log-linear models are presented in Table 1.

Table 1

Fraction of models selected in the backward selection procedure that do not fit the underlying model, sample sizes N (10 000 replications for each combination)

Underlying model	Number of cells in the 5-way tables	Fraction of models that do not fit the underlying model				
		N=100	N=300	N=500	N=1000	N=3000
{*ABC*} {*DE*}	$s = 32$	0.523	0.207	0.057	0.052	0.051
{*AB*} {*BC*} {*DE*}		0.521	0.201	0.056	0.051	0.050
{*AB*} {*C*} {*DE*}		0.503	0.189	0.054	0.050	0.050
{*AB*} {*C*} {*D*} {*E*}		0.480	0.169	0.051	0.050	0.050

V. SUMMARY AND CONCLUSIONS

In the paper some statistical properties of a model fitting procedure developed for log-linear models were studied. From practical standpoint, many computing facilities already have selection routines for log-linear contingency table analysis and the backward selection procedure is one of the most popular.

In the paper small Monte Carlo simulations were performed and some properties of the backward selection test procedure were studied. Typical results obtained for 5-way contingency tables indicate that for sparse contingency tables the selection procedure leads to a substantially large fraction of models that differ from the underlying correct one.

REFERENCES

Benedetti J. K., Brown M. B. (1978), *Strategies for the Selection of Log-linear Models*, „Biometrics", **34**.

Dawid A. P. (1979), *Conditional Independence in Statistical Theory*, „Journal of the Royal Statistical Society", ser. B, **41**.

Darroch J. N., Lauritzen S. L., Speed T. P. (1980), *Markov Fields and Log-linear Interaction Models for Contingency Tables*, „Ann, Stat.", **8**.

Lauritzen S. L., Spiegelhalter D. J. (1988), *Local Computations with Probabilities on Graphical Structures and Their Applications to Expert Systems*, „Journal of the Royal Statistical Society", ser. B, **50**, 157–224.

Matzkevich I., Abramson B. (1995), *Decision Analytic Networks and Artificial Intelligence*, „Management Science" **41**, 1–22.

Agnieszka Rossa

BADANIE WŁASNOŚCI PROCEDURY SELEKCJI „WSTECZ" DLA GRAFICZNYCH MODELI LOGARYTMO-LINIOWYCH – ANALIZA MONTE CARLO

W pracy przedstawione są wyniki analizy Monte Carlo przeprowadzonej na podstawie 5-wymiarowych tablic kontyngencyjnych. Celem analizy jest oszacowanie frakcji graficznych modeli logarytmo-liniowych poprawnie wybranych przez tzw. procedurę selekcji wstecz.