A C T A U N I V E R S I T A T I S L O D Z I E N S I S
FOLIA OECONOMICA 285, 2013

*Justyna Brzezińska**

MODEL SELECTION METHODS IN LOG-LINEAR ANALYSIS

Abstract. The main objective of the study is to examine model selection methods in log-linear analysis. Log-linear analysis is a tool for independence analysis of qualitative data. Cell counts are Poisson distributed and all variables are treated as response. This method allows to analyze any number of variables in a multi-way contingency table. In log-linear analysis we model cell counts, where expected cell frequencies are functions of parameters representing characteristics of the categorical variables and their relationships with each other (interaction).

The purpose of this paper is the presentation and comparison of model election criteria. The most popular statistics are chi-square test, likelihood ratio test and information criteria (*AIC* [Akaike 1973] and *BIC* [Raftery 1986]) but also Aitkin [Aitkin 1978] method for high dimensional tables.

Key words: log-linear analysis, contingency table, model selection methods.

I. LOG-LINEAR ANALYSIS – INTRODUCTION

The analysis of discrete cross-classified multivariate data has occupied an important place in statistics since the beginning of XX century. Although several important papers were published, the development and the use of methods for the analysis of cross-classified data had to await the general availability of computer software and statistical packages. Today it is possible to analyze large datasets of cross-classified data and to focus on data itself. In this paper log-linear analysis and its model selection criteria are presented.

Log-linear models are a standard tool to analyze structures of dependency in multi-way contingency tables. The criteria to be analyzed are the expected cell frequencies as a function of all the variables in the survey. There are several types of log-linear models depending on number of variables and interactions included. Saturated model for a three-way $H \times J \times K$ ($h = 1,2,...,H$, $j = 1,2,...,J, k = 1,2,...,K$) table includes all the possible effects in multiplicative form for three variables is given as:

* Ph. D. student at the University of Economics in Katowice.

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

where m_{hjk} - expected cell counts for the contingency table.

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

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

where: λ represents an overall effect or a constant, λ_h^X, λ_j^Y, λ_k^Z represents the effect of the row, column and layer variable *X*, *Y*, *Z*, λ_{hj}^{XY}, λ_{hk}^{XZ}, λ_{jk}^{YZ} represents the interaction between two variables XY, XZ, YZ, λ_{hjk}^{XYZ} is an interaction term between XYZ.

Saturated model reproduces perfectly the observed cell frequencies through the theoretical frequencies and 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 have to be considered. A rule of thumb to determine the degrees of freedom is *df* = number of cells – number of free parameters [Agresti 2002]. The starting point is saturated model. Thus, the aim of a researcher is to find 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 variable.

As the number of dimensions of a multidimensional table increases, so does the number of possible models. The model selection methods considered are the stepwise procedures (forward selection and backward elimination). Significance of test statistics is measured by their *p*-value and a test statistic fails to achieve a predetermined minimum level of significance α if $p > \alpha$ and it maintains that level of significance if $p < \alpha$. A value for α error lies between 0.1 and 0.35 [Bishop et al., 1975]. When the null hypothesis is rejected, the result is said to be statistically significant.

A unique set of ML estimates for every cell can be derived from the sufficient statistics alone with the use of iterative proportional fitting [Deming, Stephen 1940].

II. MODEL SELECTION METHODS IN LOG-LINEAR ANALYSIS

The main goal of log-linear analysis is to find the smallest model that fits the data. The overall goodness-of-fit of a model is assessed by comparing the expected frequencies to the observed cell frequencies for each model. The goodness of fit of a log-linear model is usually tested using either the Pearson chi-square test statistic or the likelihood ratio statistic:

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

where n_{hjk} – observed cell counts a for three-way table.

Therefore, larger G^2 values indicate that the model does not fit the data well and thus, the model should be rejected. This strategy is the opposite of the usual chi-square test of independence, where we seek to reject the null hypothesis of no association. But in trying to find the best fitting log-linear model to describe cross-table, we hope to accept the hypothesized model, hence we want to find a low G^2 value relative to *df* [Knoke, Burke 1980]. The likelihood ratio can also be used to compare an overall model within a smaller, nested model (i.e. saturated model with one interaction or main effect dropped to assess the importance of that term). The equation is $\Delta G^2 = G_2^2 - G_1^2$ with: $\Delta df = df_2 - df_1$, where 2 is nested model, 1 is the higher parameterized model, df_1 and df_2 are degrees of freedom for model 1 and model 2.

Degrees of freedom. If the ΔG^2 comparison statistic is not significant, then the nested model is not significantly worse than the saturated model.

In order to find the best model from a set of possible models, additional measurements should be considered. Akaike information criterion [Akaike 1973] refers to the information contained in a statistical model according to equation:

$$AIC = G^2 - 2df \ . \tag{4}$$

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

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

where n – total sample size.

The model that minimizes AIC and BIC will be chosen.

In log-linear models G^2 plays similar role to that of SSE (error sum of squares) in regression analysis. If X_0 indicates the smallest model and X indicates the log-linear model of interest, we define:

$$R^2 = \frac{G^2(X_0) - G^2(X)}{G^2(X_0)}, \tag{6}$$

where $G^2(X)$ and $G^2(X_0)$ are the likelihood ratio test statistics for mail and smallest model. As in standard regression, as well as log-linear analysis R^2 cannot be used to compare models that have different number of degrees of freedom (the larger models have larger R^2). To compare the R^2 measures of various models it is necessary to adjust them by degrees of freedom according to:

$$R^2_{Adj} = 1 - \frac{G^2(X)/(q-r)}{G^2(X_0)/(q-r_0)}, \tag{7}$$

where: q denotes number of cells in contingency table, r and r_0 are degrees of freedom for the model X and X_0. A large value of R^2_{Adj} indicates that the model X fits well.

Aitkin [1978, 1979] suggested a model selection method for testing every intermediate between *s* and *s*–1 factor model. To test the need for *s*-factor effect, we reject the null hypothesis of no *s*-factor effect if:

$$G^2_{s-1} - G^2_s > \chi^2(1-\gamma_s, d_{s-1} - d_s), \tag{8}$$

where: G^2_s is the likelihood ratio test statistic with d_s degrees of freedom for testing all *s*-factor model against the saturated model. This is a test for the adequacy of the *s*–1 – factor model. Aitkin then identifies the smallest value of *s* for which the all *s*-factor effects models adequately fits the data. In (8) γ_s is chosen to satisfy:

$$1 - \gamma_s = (1-\alpha)^{\binom{t}{s}}, \tag{9}$$

or for some value α to choose $\alpha = \gamma_2 = ... = \gamma_t$ [Christensen 1997]. (9) is the probability of not rejecting any of the *s*-factor tests. This is test for the adequacy of all the *s-1*-factor model. Aitkin then identifies the smallest value of *s* for which the all *s*-factor effects model adequately fits the data. The model with the largest *s* value for which (8) is fulfilled will be chosen. Aitkin [1979] suggests that it is reasonable to pick up an α level that yields a γ between 0.25 and 0.5 [Christensen 1997].

III. APPLICATION IN R

Log-linear analysis is available in **R** with the use of `loglm` function in `MASS` library. It allows to build model with any number of variable and any set of interactions. Parameters are estimated with the iterative-proportional fitting algorithm (IPF) [Deming, Stephen 1940] and data in contingency table are Poisson distributed with no distinguish between dependent and independent variable.

Consider multi-way contingency table for `Chile` dataset (`library(car)`) with four categorical variables: `region` (C, Central; M, Metropolitan Santiago area; N, North; S, South; SA, city of Santiago), `sex` (F, female; M, male), `education` (P, Primary; PS, Post-secondary; S, Secondary) and `vote` (A, will abstain; N, will vote no (against Pinochet); U, undecided; Y, will vote yes (for Pinochet)) for 2700 respondents. All *s*-factors models were built and likelihood statistic G^2, information criteria and R^2 were computed with corresponding *p*-value.

Table 1. Model selection statistics

Model	G^2	*df*	*p*	*AIC*	*BIC*	R^2	R^2_{adj}
$[RSEV]$	0	0	1	0	0	1	1
$[RSE][RSV][REV][SEV]$	18.372	24	0.785	–29.628	–171.253	0.956	0.995
$[RS][RE][RV][SE][SV][EV]$	73.404	74	0.498	–74.596	–511.271	0.823	0.958
$[R][S][E][V]$	414.210	109	0.000	196.210	-447.000	0	0

Source: own calculations in **R**.

The comparison of all statistics shows, that the best model is $[RSE][RSV][REV][SEV]$ model. For this model the difference between *df* and

G^2, as well as *AIC* and *BIC* are the smallest with large R^2. This means that for this model the difference between observed and expected cells is the smallest meaning the model fits the data well.

Another way to find the best fitting model is ANOVA method where difference between models are tested with corresponding p-value. We test the null hypothesis that $\Delta G^2 = 0$.

Model fits the data well when its p-value exceeds 0.2 and when the difference between deviance (ΔG^2) and df is relatively small.

```
LR tests for hierarchical log-linear models

Model 1:
 ~region + sex + education + vote
Model 2:
 . ~ sex + region + education
Model 3:
 . ~ sex + region + education

           Deviance  df Delta(Dev) Delta(df) P(> Delta(Dev)
Model 1   414.20949 109
Model 2    76.07504  74  338.13445        35        0.00000
Model 3    18.37152  24   57.70352        50        0.21188
Saturated   0.00000   0   18.37152        24        0.78462
```

It is clear that model that fits the data is model 3 (3-factors model) $[RSE][RSV][REV][SEV]$. Its deviance is close enough to the deviance of the saturated model and its p-value exceeds 0.2. It proves the result above obtained with the use of chi-square statistics, information criteria and R^2.

In the next step Aitkin method is presented where all *s* and *s-1*-factors models are compared with the use of chi-square statistic.

Table 2. Aitkin`s model selection

s	$s-1$ vs. s	$G^2_{s-1} - G^2_s$	Δdf	$\chi^2(1-\gamma_s, df_{s-1} - df_s)$
3	3 versus 4	18.372	24	$\chi^2(0.95,24) = 13.848$
2	2 versus 3	55.032	50	$\chi^2(0.815,50) = 40.983$
1	1 versus 2	340.806	35	$\chi^2(0.265,35) = 29.397$

Source: own calculations in **R**.

The largest value of s for which $G^2_{s-1} - G^2_s > \chi^2\left(1-\gamma_s, d_{s-1} - d_s\right)$ is $s = 4$. According to Aitkin`s criteria the saturated model $[RSEV]$ fits the data, however this model is useless because it includes all possible effects and interaction between variables. In this example Aithin`s methods shows that best fitted model is saturated mode. Classical criteria showed previously showed that 3-factors model fits the data well and this model should be chosen as the best because the aim of log-linear analysis is the smallest model that fits the data.

Sometimes it is not so clear which model to choose, then the best is to stick to one method (e.g. information criteria or one of chi-square statistic). In this example two methods give the same result, Aitkin`s method gives saturated model as the best. Sometimes this method can give smaller model.

IV. CONCLUSION REMARKS

Log-linear models are concerned with the analysis of cross-classified data and they allow to analyze the relationship between two or more categorical variable. It allows to examine relationship between categorical data with interactions and several types of independence are considered: independence model, saturated model, homogeneous association, joint independence and partial independence. As the number of variables increases, number of possible interactions creases dramatically as well. Several models are built and with the use of model selection criteria the best model is chosen (model with the fewest parameters that fits data well). The analysis can easily be extended t tables with more variables, however as well as independence and odds ratio relationship become more complex.

Model selection criteria presented in this paper are: chi-square and likelihood ratio, information criteria (*AIC, BIC*), R^2 coeficient, ANOVA and Aitkin`s method for higher-dimensional tables. They may give smaller model than with the use of stepwise method. However, it is always advisable to use particular model selection criteria to choose model that fits data well, but as well as information and structure that goes with particular model (model that is easy to interpret).

"*The analysis of the data does not end with finding an appropriate type model; that is just an important first step*" [Christensen 1997].

REFERENCES

Agresti A. (2002), *Categorical data analysis*, Wiley & Sons, Hoboken, New Jersey.
Aitkin M. (1978), *The analysis of unbalanced cross-classification*s, Journal of the Royal Statistical Society, Series B, 141, 195–23.
Aitkin M. (1979), *A simultaneous test procedure for contingency tables*, Applied Statistics, 28, 233–242.
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.
Deming W., Stephan F. (1940), *On a least squares adjustment of a sampled frequency table when the expected marginal totals are known*, Annals of Mathematical Statistics, 11, 427–444.
Knoke D., Burke P. J. (1980), *Log–linear models,* Quantitative Applications in the Social Science, No 20, Sage University Papers, Sage Publications, Newbury Park, London, New Delhi.
Raftery A. E. (1986), *Choosing models for cross-classification*, Amer. Sociol. Rev. 51, 145–146.

Justyna Brzezińska

KRYTERIA WYBORU MODELU W ANALIZIE LOGARYTMICZNO-LINIOWEJ

Analiza logarytmiczno-liniowa jest metodą przeznaczoną do badania zależności pomiędzy zmiennymi niemetrycznymi w tablicy kontyngencji. Zmienne o rozkładzie Poissona traktowane są jako zmienne objaśniane. Metoda ta pozwala na analizę dowolnej liczby zmiennych, a także na uwzględnienie interakcji zachodzących pomiędzy nimi. W analizie logarytmiczno-liniowej modelowane są liczebności w poszczególnych komórkach tablicy, przy czym liczebności oczekiwane są funkcją parametrów reprezentujących zmienne dyskretne oraz relacje między nimi.

Celem niniejszego artykułu jest prezentacja i porównanie kryteriów wyboru modelu w analizie logarytmiczno-liniowej. Podstawowymi kryteriami wyboru modelu są statystyka chi-kwadrat oraz iloraz wiarygodności oraz kryteria informacyjne *AIC* i *BIC*. W niniejszym artykule zaprezentowana zostanie także metoda Aitkina, która przeznaczona jest do porównywania jakości dopasowania modeli o dużej liczbie zmiennych.