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 269, 2012

Justyna Brzezińska[*]

HIERARCHICAL LOG-LINEAR MODELS FOR CONTINGENCY TABLES

Abstract. Log-linear models are widely used for qualitative data in multidimensional contingency tables. Hierarchical log-linear models are models that include all lower-order terms composed from variables contained in a higher-order model term. The starting point is a saturated model, then homogenous associations, conditional independence and complete independence. There are several statistics that help to choose the best model. The first is the likelihood ratio approach, next is AIC and BIC information criteria. In **R** software there is `loglm()` function in MASS library and glm in stats library. The first approach is presented in this paper.
Key words: log-linear models, hierarchical log-linear models, AIC, BIC.

I. INTRODUCTION

Log-linear analysis is a widely used tool for modeling qualitative data in contingency table. Log-linear models provide a powerful tool for teasing out the relationships among the variables in multi-way contingency tables. In this paper log-linear analysis for contingency tables is presented. Log-linear analysis is technique that makes no distinction between dependent and independent variables and it is used to examine relationship among categorical variables. The standard approach is hierarchical modeling, where a set of possible model is chosen by regarding fit criteria. There are two approaches called stepwise procedure in model selection: stepwise selection and backward elimination. In log-linear analysis expected values of the observations are given by a linear combination of a number parameters. Maximum likelihood method is used to estimate the parameters, and estimated parameter values may be used in identifying which variable are of great importance in predicting the observed values.

II. CONTINGENCY TABLE

The problem of interaction between variables was developed by Bartlet [1935], Roy and Kastenbaum [1956], Darroch [1962], Birch [1963] and

[*] Msc, Department of Statistics, The Karol Adamiecki University of Economics, Katowice.

Goodman [1970]. A widely test used for testing the independence model is the Pearson chi-square test or likelihood ratio defined as:

$$G^2 = 2\sum_{h=1}^{H}\sum_{j=1}^{J} n_{hj} \ln\left(\frac{n_{hj}}{\hat{n}_{hj}}\right), \tag{1}$$

Statistical independence between row and column variable is:

$$p_{hj} = p_{h.} \cdot p_{.j}, \tag{2}$$

where:

$$p_{h.} = \sum_{j=1}^{J} p_{hj} = \frac{n_{h.}}{n_{..}}, \; p_{.j} = \sum_{h=1}^{H} p_{hj} = \frac{n_{.j}}{n_{..}}.$$

The frequencies equal:

$$\hat{n}_{hj} = n \cdot p_{hj} = n \cdot p_{h.} \cdot p_{.j}. \tag{3}$$

Depending on which marginal frequencies are fixed from the begin of the study and hence, which marginal frequencies are random, it is essential to distinguish between the distributions of the cell frequencies in the table. There are three possible survey distributions in contingency table (Mair [2006]): multinomial, product-multinomial and Poisson distribution, but the most frequent for hierarchical log-linear models is Poisson.

III. ODDS AND ODDS-RATIO

Odds are the ratios of the probability of an event occurring to the probability of the event not occurring. Odds ratio is defined as (Agresti [2002]):

$$\theta = \frac{\omega_1}{\omega_2} = \frac{n_{11}/n_{21}}{n_{12}/n_{22}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}. \tag{4}$$

There is another function of the odds-ratio called Q Yulle`s statistic (Knoke, Burke [1980]):

$$Q = \frac{n_{11}n_{22} - n_{12}n_{21}}{n_{11}n_{22} + n_{12}n_{21}} \tag{5}$$

ranges from [-1,1], with „0” indicating no relationship between variables.

IV. OVERVIEW OF MODELS FOR TWO-WAY CONTINGENCY TABLE

There are several types of log-linear models for two-way contingency table. Saturated model includes all the possible effects to explain every single expected cell frequency is: $\log(\hat{n}_{hj}) = \log(\eta \tau_h^X \tau_j^Y \tau_{hj}^{XY}) = \mu + \lambda_h^X + \lambda_j^Y + \lambda_{hj}^{XY}$, where: λ represents an overall effect or a constant, λ_h^X represents the main or marginal effect of the row variable X, λ_j^Y represents the main or marginal effect of the column variable Y.

V. TESTING AND GOODNESS-OF-FIT

In addition, the use of the model selection criteria will be discussed. The main goal 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 can be tested using either the Pearson chi-square test statistic or the likelihood ratio statistic (1). In order to find the best model from a set of possible models, additional measurements should be considered. Akaike information criteria (Akaike [1973]) refers to the information contained in a statistical model according to equation:

$$AIC = G^2 - 2df \,. \tag{6}$$

Another information criteria is Bayesian Information Criteria (Raftery [1986]):

$$BIC = G^2 - df \cdot \ln n \,. \tag{7}$$

Significance of test statistics is measured by their *p*-value. 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 proposed 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. In this paper α-error is set to be 0.2.

VI. APPLICATION IN R

This data frame contains the responses of 237 students at the University of Adelaide to a number of questions (Venables, W. N., Ripley, B. D. [1999]). Data is available in `library(MASS), data(survey).` Log-linear analysis with three categorical variables: `Sex ("Male", "Female")`, W. Hnd (`"Right", "Left"`), `Exer` (`"Freq", "Some", "None"`). Log-linear analysis can be used with the use of loglm function.

```
>  print(model.no.interaction)
Call:
loglm(formula = ~Sex + W.Hnd + Exer, data =
contingency.table,
    fit = T, param = T)

Statistics:
                      X^2 df  P(> X^2)
Likelihood Ratio  9.713843  7 0.2053780
Pearson          10.066693  7 0.1848254
```

For model with no interaction the likelihood ratio is P(>X^2)=0.205 what means, that model is fitted well and we can select this model and final model. In the next step models containing pairs of interaction will be tested.

```
> model.no.interaction.plusSW <-
update(model.no.interaction, .~. + Sex:W.Hnd,
data=contingency.table)
>  print(model.no.interaction.plusSW)
Call:
loglm(formula = . ~ Sex + W.Hnd + Exer + Sex:W.Hnd, data =
contingency.table,
    fit = T, param = T)

Statistics:
                     X^2 df  P(> X^2)
Likelihood Ratio 9.167555  6 0.1643719
Pearson          9.258123  6 0.1595734

>  model.no.interaction.plusSE <-
update(model.no.interaction, .~. + Sex:Exer,
data=contingency.table)
>  print(model.no.interaction.plusSE)
Call:
loglm(formula = . ~ Sex + W.Hnd + Exer + Sex:Exer, data =
contingency.table,
    fit = T, param = T)
```

```
Statistics:
                      X^2 df  P(> X^2)
Likelihood Ratio 3.659013  5 0.5994751
Pearson          4.097142  5 0.5355164

> model.no.interaction.plusWE <-
update(model.no.interaction, .~. + W.Hnd:Exer,
data=contingency.table)
> print(model.no.interaction.plusWE)
Call:
loglm(formula = . ~ Sex + W.Hnd + Exer + W.Hnd:Exer, data =
contingency.table,
    fit = T, param = T)

Statistics:
                      X^2 df  P(> X^2)
Likelihood Ratio 8.082703  5 0.1517362
Pearson          8.022221  5 0.1550152
```

Only one model (*model.no.interaction.plusSE*) fits data with p-value greater than 0,20. The next model is build.

```
> print(model.interaction2)
Call:
loglm(formula = . ~ Sex + W.Hnd + Exer + Sex:W.Hnd +
Sex:Exer +
  W.Hnd:Exer, data = contingency.table, fit = T, param = T,
  print = TRUE)

Statistics:
                      X^2 df  P(> X^2)
Likelihood Ratio 1.303964  2 0.5210121
Pearson          1.348596  2 0.5095139
```

For the next model the likelihood ratio is `P(>X^2)=0.521` what means that the second model also fits data and observed and expected cell frequencies do not differ. The next step is to compare all models that fit data and to choose one using likelihood ratio statistic. Each item in the last column (`Delta(Dev)`) compares `Deviance` between the current row and the previous row.

```
             Deviance df Delta(Dev) Delta(df) P(>
Delta(Dev)
   Model 1   9.713843  7
   Model 2   3.659013  5   6.054830        2        0.04844
   Model 3   1.303964  2   2.355049        3        0.50206
   Saturated 0.000000  0   1.303964        2        0.52101
```

Table 1 presents comparison using other statistics (χ^2, G^2, AIC, BIC and R^2).

Table 1. Goodness-of-fit for tested models with the hierarchy principle

Symbol	Model	χ^2	G^2	df	AIC	BIC	R^2	Δdf
$[S][W][E]$	model.no.interaction	10.067	9.714	7	-4.286	-28.563	0	
$[SE][W]$	model.no.interaction.plusSE	4.097	3.659	5	-6.341	-23.681	0.623	2
$[WE][WS][ES]$	model.interaction2	1.349	1.304	2	-2.696	-9.632	0.866	3
$[SEW]$	saturated model	0	0	0	0	0	1	2

Source: own calculations.

The model that fit data well is model `model.interaction2` ($[WE][WS][ES]$). This model is a model of homogenous association and no graphical result is available. In this model any interaction between two variables is permitted. Its deviance is close enough to the deviance for the saturated model to give the p-value greater than 0.20. Fitted counts for this model are given:

```
, , Exer = Freq
        W.Hnd
Sex          Left   Right
  Female 1.965148 46.0351
  Male   4.034852 60.9649

, , Exer = None
        W.Hnd
Sex         Left     Right
  Female 1.09609  9.903966
  Male   1.90391 11.096034

, , Exer = Some
        W.Hnd
Sex          Left    Right
  Female 3.939492 54.06017
  Male   4.060508 35.93983
```

VII. CONCLUSION REMARKS

Log-linear models are very effective statistical tool for analyzing multiway tables. The procedure using hierarchical models is widely used in marketing, social and psychological research providing information about data structure. Log-linear models have two advantages: they are flexible and interpretable. Log-linear models are extendable for any dimensionality of contingency table.

Interaction parameters are most useful in association interpretation. Log-linear models can be estimated in **R** software with `loglm` and `glm` function but the most popular models are hierarchical.

REFERENCES

Agresti A. (2002), *Categorical Data Analysis*, Wiley & Sons, Hoboken, New Jersey.

Akaike H. (1987), *Factor Analysis and AIC. Psychometrika, 52.*

Bartlet M. S., (1935), *Contingency table interactions,* in: Journal of the Royal Statistical Society, Supplement 2.

Birch M. W. (1963), *Maximum likelihood in three-way contingency tables,* in: Journal of the Statistical Society, Series B, 25.

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.

Darroch J. N. (1963), *Interactions in multi-factor contingency tables,* in: Journal of the Royal Statistical Society, Series B, 24.

Goodman L. (1970), *The Multivariate Analysis of Qualitative Data: Interaction among Multiple Classifications*, in: Journal of the American Statistical Association, 65.

Knoke D., Burke P. J. (1980), *Log–linear Models,* Quantitative Applications in the Social Science", nr 20, Sage University Papers, Sage Publications, Newbury Park, London, New Delhi.

Mair P. (2006), *Interpreting Standard and Nonstandard Log-linear Models*, Waxmann Verlag, Münster.

Raftery A. E. (1986), *A note on Bayesian Factors for log-linear contingency table models with vague prior information,* Journal of the Royal Statistical Society, Ser. B, 48.

Reynolds H. T. (1977), *Analysis of Nominal Data*, Beverly Hills, CA: Sage.

Roy S. N., Kastenbaum M. A. (1956), *On the hypothesis of no interaction in a multiway contingency table*, in: The Annals of Mathematical Statistics, 27.

Venables, W. N., Ripley, B. D. (1999), *Modern Applied Statistics with S-PLUS,* Third Edition, Springer.

Justyna Brzezińska

HIERARCHICZNE MODELE LOGARYTMICZNO-LINIOWE DLA TABLIC KONTYNGENCJI

Hierarchiczne modele logarytmiczno-liniowe służą do analizy struktury zależności zmiennych w postaci tablicy kontyngencji. Modele budowane według zasady hierarchiczności są modelami hierarchicznymi. Do modeli tych zaliczany jest model pełny, model niezależności homogenicznej, model niezależności warunkowej oraz model niezależności całkowitej. Do kryteriów wyboru modelu należą: współczynnik największej wiarygodności, kryterium informacyjne AIC oraz BIC. Analiza logarytmiczno-liniowa w programie **R** możliwa jest dzięki funkcji `loglm()` z pakietu MASS oraz funkcji `glm` z pakietu `stats`.