Analysis of emigration in Poland with the use log-linear models

Justyna Brzezińska[1]

Abstract

Log-linear analysis is a statistical method used for the independence analysis of qualitative data in contingency table. 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 table. Log-linear models, where interaction terms are included, enable to examine various types of association (conditional model, complete independence model, homogenous association model, saturated model, null model). In log-linear analysis we model cell counts in a contingency table in terms of associations among variables and marginal frequencies. For testing the goodness of fit the chi-square test, likelihood ratio test and information criteria *AIC* [3] and *BIC* [12] are used. Visualizing with the use of several plots designed for categorical data will be presented to enhance the interpretation. The data presented in this paper come from Demographic Yearbook of Poland 2012 and include report on emigration in Poland in 2011. All calculations will be conducted in **R**.

Keywords*: log-linear models, categorical data analysis, emigration in Poland.*

JEL Classification: C53, C10, C25

AMS Classification: 97K80

1. Introduction

Log-linear analysis is a standard tool to analyze path of association between nominal variables in a multi-way contingency table. The criteria to be analyzed are the expected cell frequencies m_{hjk} represented as a function of all variables in the survey. There are several types of log-linear models depending on a number of variables and interactions included [1], [2], [5], [6], [8], [11]. The most complex model for a three-way table containing all possible effects and interactions is given as:

$$\log\left(m_{hjk}\right) = \lambda + \lambda_h^X + \lambda_j^Y + \lambda_k^Z + \lambda_{hj}^{XY} + \lambda_{hk}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{hjk}^{XYZ} \qquad (1)$$

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

[1] University of Economics in Katowice, Faculty of Management, 1-go Maja 50, 40-136 Katowice, Poland, justyna.brzezinska@ue.katowice.pl.

To obtain the maximum likelihood estimates of m_{hjk} for a particular model we can use three sampling distributions: Poisson, multinomial and product multinomial, however the parameter estimates for all sampling schemes are the same [4]. Saturated model reproduces perfectly the observed cell frequencies through the theoretical frequencies and such model is meaningless since the aim of the analysis is to find a more parsimonious model with less parameters. In order to find the best model from a set of possible models, additional measures should considered. A rule of thumb to determine the degrees of freedom is $df =$ number of cells – number of free parameters.

For a three-way table, a model that allows association between variables is the saturated model (1), where restrictions must be imposed for estimation: $\sum_{h=1}^{H}\lambda_h^X = \sum_{j=1}^{J}\lambda_j^Y = \sum_{k=1}^{K}\lambda_k^Z = 0$, $\sum_{h=1}^{H}\lambda_{hj}^{XY} = \sum_{j=1}^{J}\lambda_{hj}^{XY} = \sum_{h=1}^{H}\lambda_{hk}^{XZ} = \sum_{k=1}^{K}\lambda_{hk}^{XZ} = \sum_{j=1}^{J}\lambda_{jk}^{YZ} = \sum_{k=1}^{K}\lambda_{jk}^{YZ} = 0$, $\sum_{h=1}^{H}\lambda_{hjk}^{XYZ} = \sum_{j=1}^{J}\lambda_{hjk}^{XYZ} = \sum_{k=1}^{K}\lambda_{hjk}^{XYZ} = 0$. There are $H-1$ linearly independent λ_h^X row parameters, $J-1$ linearly independent λ_j^Y column parameters, $K-1$ linearly independent λ_k^Z layer parameters, $(H-1)(J-1)$ linearly independent λ_{hj}^{XY} association parameters, $(H-1)(K-1)$ linearly independent λ_{hk}^{XZ} association parameters, $(J-1)(K-1)$ linearly independent λ_{jk}^{YZ} association parameters and finally $(H-1)(J-1)(K-1)$ linearly independent λ_{hjk}^{XYZ} association parameters. This is an example of hierarchical models and it means that that the model must contain all lower-order terms contained within any high-order term in the model. The aim of log-linear analysis is to find a reduced 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. In a contingency table with more variables there are hundreds of models that can be considered. For a three-way table there are 19 possible models. Dozens may provide acceptable levels of fit at conventional levels of significance. There are some processes in all specification searches.

The goodness of fit of a log-linear model in three-way table is tested using the likelihood ratio statistic [6]:

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

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

In the usual chi-square test of independence, we seek to reject the null hypothesis of no association between the variables; hence we hope to find a large χ^2 or G^2 value relative to df. But in trying to find the best fitting-model in a multi-way table, hope to accept the hypothesized model; hence, we want to find a low G^2 value relative to df. An acceptable model is one whose expected cell frequencies do not differ from the observed data.

The Akaike Information Criterion AIC [3] is based on information theory, but a heuristic way to think about it is a criterion that seeks a model that has a good fit to the truth but few parameters. The chosen model is the one that minimizes the Kullback-Leibler distance between the model and the truth. Akaike information criterion refers to the information contained in a statistical model according to the equation:

$$AIC = G^2 - 2df \tag{3}$$

where df is the residual degrees of freedom. The model that minimizes AIC 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 the coefficient of determination [6]:

$$R^2 = \frac{G^2(X_0) - G^2(X)}{G^2(X_0)} \tag{4}$$

where $G^2(X)$ and $G^2(X_0)$ are the likelihood ratio test statistics for base 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 [6]:

$$R^2_{Adj} = 1 - \frac{G^2(X)/(q-r)}{G^2(X_0)/(q-r_0)} = 1 - \frac{q-r_0}{q-r}\left(1 - R^2\right) \tag{5}$$

where: q denotes number of cells in the 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.

2. Analysis of emigration in Poland with the use of log linear models

The history of migration in Poland is characterized largely by emigration. Until the end of the 20th century, emigration took place both in large waves and in continual yearly movements. The opening up of borders and restoration of the freedom of travel and the creation of business and employment opportunities in Poland were of crucial importance in the transformation of migration trends in the 1990s. Poland's participation in European networks and institutions dealing with migration matters has prompted Poland to modernize her border control infrastructure and to increase the numbers of better trained border guards. Reports from the Central Statistical Office of Poland show the rate of migration in Poland increases every year.

The dataset presented in the paper comes from the Central Statistical Office of Poland and present migrations structure in 2011 in Poland in term of three variables: Region (R) (Austria, Belarus, France, Spain, Ireland, Netherlands, Germany, Norway, Sweden, United Kingdom, Italy, Asia, Africa, North and Central America, South America, Australia and Oceania), Area (A) (Urban, Rural) and Sex (S) (Male, Female). Sample size is 19157 respondents.

All possible log-linear models for a three-way table were tested and goodness-of-fit statistics (G^2, AIC, R^2, R^2_{Adj}) were computed. Goodman`s bracket notation is used to express the corresponding model equation with the highest order according to the hierarchy principle. [10]. Log-linear analysis is available in **R** with the use of `loglm {MASS}` function.

Model	G^2	df	p	AIC	R^2	R^2_{Adj}
$[RAS]$	0	0	1	0	1.000	1
$[RA][RS][AS]$	21.726	15	0.115	-8.274	0.980	0.980
$[RA][RS]$	32.071	16	0.010	0.071	0.971	0.971
$[RA][AS]$	210.994	30	0	150.994	0.810	0.810
$[RS][AS]$	927.677	30	0	867.677	0.164	0.165
$[RA][S]$	212.853	31	0	150.853	0.808	0.808
$[RS][A]$	929.536	31	0	867.536	0.163	0.163
$[AS][R]$	1108.459	45	0	1018.459	0.002	0.001
$[R][A][S]$	1110.318	46	0	1018.318	0.000	0

Table 1 Goodness-of-fit statistics for three-way table.

Parameters estimates for the homogenous association model							
$\lambda^{R}_{Austria}$	0.3108	$\lambda^{RA}_{Belarus,Urban}$	-0.2070	$\lambda^{RA}_{Africa,Urban}$	0.1806	$\lambda^{RS}_{Norway,Male}$	0.1616
$\lambda^{R}_{Belarus}$	-0.1136	$\lambda^{RA}_{Belarus,Rural}$	0.2070	$\lambda^{RA}_{Africa,Rural}$	-0.1807	$\lambda^{RS}_{Norway,Female}$	-0.1616
λ^{R}_{France}	-0.1047	$\lambda^{RA}_{France,Urban}$	0.0245	$\lambda^{RA}_{NCAm,Urban}$	0.0906	$\lambda^{RS}_{Sweden,Male}$	-0.0036
λ^{R}_{Spain}	-0.3674	$\lambda^{RA}_{France,Rural}$	-0.0245	$\lambda^{RA}_{NCAm,Rural}$	-0.0906	$\lambda^{RS}_{Sweden,Female}$	0.0036
$\lambda^{R}_{Ireland}$	0.3999	$\lambda^{RA}_{Spain,Urban}$	0.1143	$\lambda^{RA}_{SAm,Urban}$	-0.3115	$\lambda^{RS}_{UK,Male}$	0.0718
λ^{R}_{Neth}	0.8489	$\lambda^{RA}_{Spain,Rural}$	-0.1143	$\lambda^{RA}_{SAm,Rural}$	0.3115	$\lambda^{RS}_{UK,Female}$	-0.0718
$\lambda^{R}_{Germany}$	3.0867	$\lambda^{RA}_{Ireland,Urban}$	0.2345	$\lambda^{RA}_{Australia,Urban}$	0.0159	$\lambda^{RS}_{Italy,Male}$	-0.3773
λ^{R}_{Norway}	-0.4169	$\lambda^{RA}_{Ireland,Rural}$	-0.2345	$\lambda^{RA}_{Australia,Rural}$	-0.0159	$\lambda^{RS}_{Italy,Female}$	0.3773
λ^{R}_{Sweden}	-0.1571	$\lambda^{RA}_{Neth,Urban}$	-0.1208	$\lambda^{RS}_{Austria,Male}$	-0.0589	$\lambda^{RS}_{Asia,Male}$	0.2029
λ^{R}_{UK}	0.1490	$\lambda^{RA}_{Neth,Rural}$	0.1208	$\lambda^{RS}_{Austria,Female}$	0.0589	$\lambda^{RS}_{Asia,Female}$	-0.2029
λ^{R}_{Italy}	-1.9358	$\lambda^{RA}_{Germany,Urban}$	-0.3408	$\lambda^{RS}_{Belarus,Male}$	-0.0019	$\lambda^{RS}_{Africa,Male}$	0.1463
λ^{R}_{Asia}	-1.9358	$\lambda^{RA}_{Germany,Rural}$	0.3408	$\lambda^{RS}_{Belarus,Female}$	0.0019	$\lambda^{RS}_{Africa,Female}$	-0.1463
λ^{R}_{Africa}	-2.2186	$\lambda^{RA}_{Norway,Urban}$	0.0384	$\lambda^{RS}_{France,Male}$	-0.0486	$\lambda^{RS}_{NCAm,Male}$	-0.0064
λ^{R}_{NCAm}	1.6926	$\lambda^{RA}_{Norway,Rural}$	-0.0384	$\lambda^{RS}_{France,Female}$	0.0486	$\lambda^{RS}_{NCAm,Female}$	0.0064
λ^{R}_{SAm}	-2.6109	$\lambda^{RA}_{Sweden,Urban}$	-0.0169	$\lambda^{RS}_{Spain,Male}$	0.0189	$\lambda^{RS}_{SAm,Male}$	-0.2089
$\lambda^{R}_{Australia}$	-0.7717	$\lambda^{RA}_{Sweden,Rural}$	0.0169	$\lambda^{RS}_{Spain,Female}$	-0.0189	$\lambda^{RS}_{SAm,Female}$	0.2089
λ^{A}_{Urban}	0.6512	$\lambda^{RA}_{UK,Urban}$	0.2673	$\lambda^{RS}_{Ireland,Male}$	0.1485	$\lambda^{RS}_{Australia,Male}$	-0.0880
λ^{A}_{Rural}	-0.6512	$\lambda^{RA}_{UK,Rural}$	-0.2673	$\lambda^{RS}_{Ireland,Female}$	-0.1485	$\lambda^{RS}_{Australia,Female}$	0.0880
λ^{S}_{Male}	-0.0473	$\lambda^{RA}_{Italy,Urban}$	-0.0220	$\lambda^{RS}_{Neth,Male}$	0.1331	$\lambda^{AS}_{Urban,Male}$	-0.0277
λ^{S}_{Female}	0.0473	$\lambda^{RA}_{Italy,Rural}$	0.0220	$\lambda^{RS}_{Neth,Female}$	-0.1331	$\lambda^{AS}_{Urban,Female}$	0.0277
$\lambda^{RA}_{Austria,Urban}$	-0.4194	$\lambda^{RA}_{Asia,Urban}$	0.4721	$\lambda^{RS}_{Germany,Male}$	-0.0896	$\lambda^{AS}_{Rural,Male}$	-0.0277
$\lambda^{RA}_{Austria,Rural}$	0.4194	$\lambda^{RA}_{Asia,Rural}$	-0.4721	$\lambda^{RS}_{Germany,Female}$	0.0896	$\lambda^{AS}_{Rural,Female}$	0.0277
λ	4.4287						

Table 2 Parameters estimates for the $[RA][RS][AS]$ model.

Using a criterion of $\alpha = 0.1$, only model the homogenous association model $[RA][RS][AS]$ (and of course the saturated model $[RAS]$) fits the data ($p = 0.115 > \alpha$). The difference between the likelihood-ratio for the saturated and the homogenous association model is $\Delta G^2 = 21.726 - 0 = 21.725$ with $\Delta df = 15 - 0 = 15$.

Thus, we have no evidences to reject the null hypothesis and conclude that the difference between observed and expected cell counts between two compared log-linear models do not differ significantly and we choose a simpler model which is $[RA][RS][AS]$.Other models do not provide an adequate fit to the data. Also Akaike information criterion AIC [3], as well as R^2 and R^2_{Adj} coefficients indicate the homogenous association model as best fitting model. In this model any interaction between two variables is permitted. The iterative proportional fitting [7] process generates maximum likelihood estimates of the expected cell frequencies for a hierarchical model. The parameters of the model are obtained with the use of `coef()` function for particular model.

The magnitude of lambda effects is measured as a departure from the value of 0. Values of parameters that are positive show that there will be more than the average number of cases expected in the cell, while if the lambdas negative indicate that there will be fewer than the average number of cases expected in that cell. When log-linear model is very complicated, it is impossible to interpret single parameter and the model equation is searched to present the path of association between variables. Also visualizing methods can provide detailed information on the data structure in multi-way contingency tables (e.g. mosaic plots, sieve plots, double-decker plots, multiple correspondence analyses etc.).

There are many techniques and methods for visualizing categorical data in contingency table which are defined as a recursive generalization of barcharts [9]. Visualization brings out the departure of an observed table from the expected table in a graphical way. Mosaic plot is one of the most popular and useful method for log-linear modelling which generalizes readily to multi-way tables. Friendly [9] extended the use of the mosaic plots for fitting log-linear models. A mosaic represents each cell of the table by a rectangle (or tile) whose area is proportional to the cell count. The mosaic is constructed by dividing a unit square vertically by one variable, then horizontally by the other. Further variables are introduced by recursively subdividing each tile by the conditional proportions of the categories of the next variable in each cell, alternating on the vertical and horizontal dimensions of the display. The mosaic plot shows that the difference between observed and expected cell counts are small and all

Pearson`s residuals ( $d_{hj} = \frac{n_{hj} - m_{hj}}{\sqrt{m_{hj}}}$ ) are $\left| d_{hj} \right| < 2$ and the fit of the model is good. Visualizing of log-linear models in mosaic plot in **R** is available with the use of `mosaic{vcd}` function.

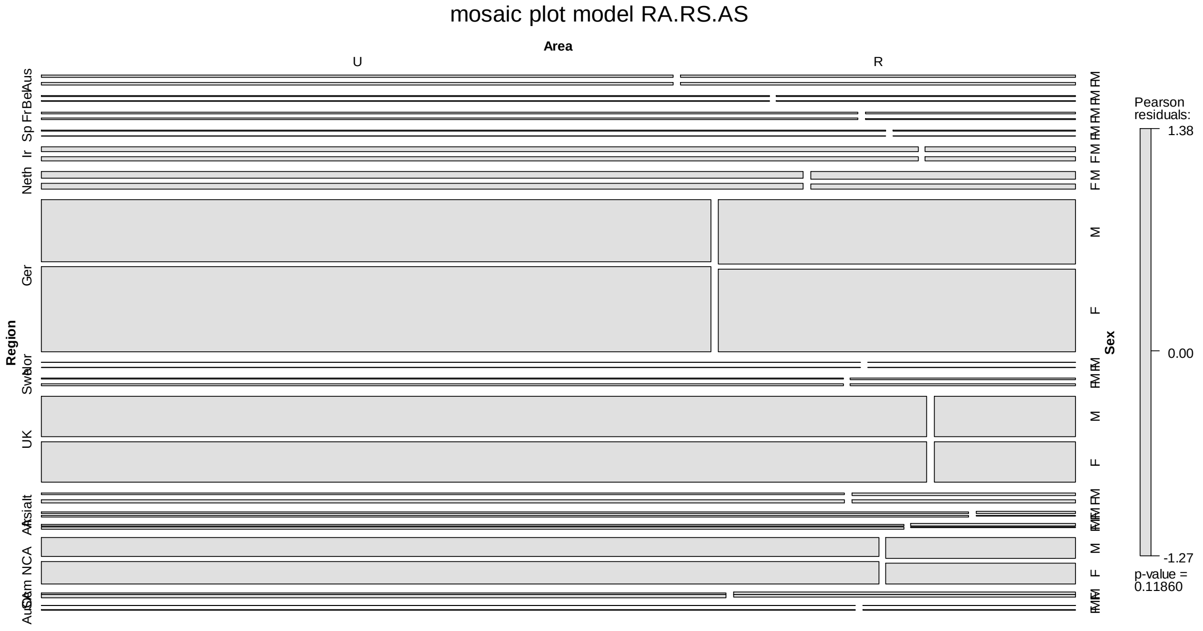

Fig. 1. Mosaic plot for the homogenous association model $[RA][RS][AS]$.

For this model the G^2 is relatively small which indicates a good fit. In this case mosaic plots can be used for testing goodness of fit. The purpose of highlighting cells is to draw attention to the pattern of departures of the data from the assumed model. Positive values indicate cells whose observed frequency is substantially greater than would be found under independence; negative values indicate cells that occur less often than under independence.

3. Conclusions

Log-linear analysis is a method of statistical analysis that is used when all the variables of interest are categorical. There is no dependent variable that can be predicted, instead, cell frequencies are modelled. Log-linear models can 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 for multi-way tables (conditional independence, complete independence, mutual independence). The advantages to be gained from the model-fitting techniques are that they provide a systematic approach to the analysis of multidimensional tables, tests of higher order interactions are available, interactions contrasts and parameters can be presented in terms of odds ratio and a hypothesis-testing or a model building scheme can be used. Both log-linear analysis as well as visualizing categorical data tools is useful and practical tool that helps to analyze independence between categorical data in contingency table. The purpose of this paper is to provide an outline of log-linear models and its application in economic research. The `loglm` package in **R** provides full log-linear analysis. Visualizing tools were also used to enhance the interpretation. Mosaic plot presents the differences between observed and expected cell counts for particular log-linear model with the use of color shaded squares.

In the paper log-linear analysis was presented on the dataset on migrations in Poland in 2011 in term of three variables: Region, Area and Sex. The best fitting-model is the homogenous association model $[RA][RS][AS]$ where the association between pairs Region-Area, Region-Sex and Area-Sex can be observed. This model shows that all possible interactions between variables are important in migration process. Mosaic plot for homogenous association model is also provided to present the data structure.

References

[1] Agresti, A., 2002. Categorical data analysis. Hoboken, New Jersey: Wiley & Sons.

[2] Agresti, A., 2006. An Introduction to Categorical Data Analysis. Hoboken, New Jersey: Wiley & Sons.

[3] Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. Proceedings of the 2nd International Symposium.

[4] Birch, M.W., 1963. The maximum likelihood in three-way contingency tables. Journal of Royal Statistical Society Series B 25, 220-233.

[5] Bishop, Y.M.M., Fienberg E.F., Holland P.W., 1975. Discrete multivariate analysis. Cambridge, Massachusetts: MIT Press.

[6] Christensen R., 1997. Log-Linear Models and Logistic Regression. New York: Springer-Verlag.

[7] 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.

[8] Fienberg, S.E., 1980. The analysis of cross-classified categorical data. Cambridge: MIT Press.

[9] Friendly, M., 2000. Visualizing categorical data. SAS Institute Inc.

[10] Goodman, L.A., 1970. The multivariate analysis of qualitative data: Interaction among multiple classifications. Journal of the American Statistical Association 65, 226-256.

[11] Mair, P., 2006. Interpreting standard and nonstandard log-linear models. Waxmann Verlag.

[12] 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, Series B 48, 249-250.