Zero cells problem in the analysis of contingency tables

Justyna Brzezińska[1]

Abstract

Log-linear analysis is a statistical tool used for the independence analysis of categorical data in contingency tables. With this method we can analyze any number of nominal or ordinal variables, we include interactions in the model, we can examine various types of association and the analysis provide a formal model equation. Although the log-linear analysis is a versatile statistical method, there are some limitations in using them due to zero cells. Zero cells in contingency table are of two types: fixed (structural) and sampling zeros. Fixed zeros occur when it is impossible to observe values for certain combinations of the variable. Sampling zeros are due to sampling variations and the relatively small size of the sample when compared with large number of cells. In the paper several options will be presented how to deal with zero cells in the table. All calculations will be conducted in **R**.

Keywords: *log-linear analysis, multi-way contingency tables, zero cells, categorical data analysis.*
JEL Classification: C59

1. Introduction

The analysis of discrete multivariate data, especially in the form of cross-classification, has occupied a prominent place in multivariate statistical analysis. Variety of social, medical, psychological and biological science data come in the form of cross-classified table counts, commonly referred to as contingency tables. Two- or multi-way table gives the observed counts simultaneously for the categories of two- or more categorical variables.

One of the most useful and powerful method for qualitative data is log-linear analysis. This method allows to examine the relationship between categorical data. It includes analysis of multi-way tables where the dimensionality of the table refers to the number of variables. And it is appropriate modeling method concerning multi-way tables including interactions which are useful in identifying. In log-linear analysis the expected value of observation is given by a linear combination of a number of parameters. Maximum likelihood method is used to estimate the parameters, and estimated parameters values may then be used in identifying which variable are of greatest importance in predicting the observed values (Everitt, 1977). This method has a variety of advantages and can give more complex and detailed information about data structure and association type. Log-linear analysis can be used for nominal as well as ordinal variable and it provides a variety of models describing association path.

[1] University of Economics in Katowice, Faculty of Management,
e-mail: justyna.brzezinska@ue.katowice.pl.

Categorical data are usually described in contingency tables (cross-table). The data to be classified in the contingency table can be split into two parts:

1. the fully classified cases where information on all the categories is available (complete tables),
2. the partially classified cases where information on some of the categories is zero (zero cells tables).

Zero cells may cause some problems with further categorical data analysis (Fienberg, 1980, Andersen, 1997, Smirnoff, 2003).

In this paper the analysis of contingency tables with zero cells will be presented. All calculations will be conducted in **R**.

2. Contingency tables containing zero cells.

A contingency table is incomplete if one or more cell have zero count. We distinguish between sampling (random) and structural (fixed) zeros.

Sampling zeros are due to sampling variation and the relatively small size of the sample when compared to the large number of cells. These zeros disappear when we increase the sample size sufficiently. Sampling zeros occur when there is no observation in the cell, i.e. $n_{hj} = 0$, but probabilistically there is a chance of observing this value and the probability of observation in a cell is $\pi_{hj} > 0$. By increasing the sample size we might get $n_{hj} > 0$. Sampling zeros typically correspond to small expected counts, however, so they can indicate that the usual asymptotic approximations for goodness-of-fit tests, tests of significance, etc., might not be valid.

Structural zeros occur when it is impossible to observe values for certain combination of the variable, i.e. $n_{hj} = 0$ and $\pi_{hj} = 0$. Tables with structural zeros are structurally incomplete and they are known as incomplete tables. This case is different from not being able to completely cross-classify all individuals or units. When we deal with table containing structural zeros, it is not allow to fill in the cells with zeros, collapse the table until there are no zeros in the table or quit the analysis.

3. Multi-way frequency analysis for contingency table with zero cells

Correspondence analysis is a method applicable for analyses of contingency tables to analyze the relations between two or more categorical variable (Greenacre, 1984). The method is performed into three steps. The first step is to calculate the categorical profiles (i.e., the

relative frequencies) and masses (marginal proportions). The next step is to compute the chi-square distances between the points and find the *n*-dimensional space that best fits the points (Clausen, 1998). The graphical representation of correspondence analysis is usually presented in perception map. Due to its popularity, the details of the analysis will not be presented in the paper; only application for zero-cell tables will be shown in comparison to log-linear analysis.

Log-linear analysis is a standard tool to analyze path of association between nominal or ordinal 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 related to several types of association, depending on a number of variables and interactions included. Models are built with the hierarchy principle saying 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.

The goodness of fit of a log-linear model for two-way table is tested using the Pearson`s chi-square statistic or the likelihood ratio statistic:

$$G^2 = 2\sum_{h=1}^{H}\sum_{j=1}^{J} n_{hj} \ln\left(\frac{n_{hj}}{m_{hj}}\right). \tag{1}$$

Therefore, larger G^2 values indicate that the model does not fit the data well and thus, such model should be rejected. In order to find the best model from a set of possible models, additional measures should considered (determination coefficients, information criteria). It is also advisable to compute G^2/df where value close to 1 indicate the model that fits well.

The Akaike Information Criterion AIC 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 (Akaike, 1973):

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

where: df is the residual degrees of freedom.

Another information measurement is Bayesian information criterion (Raftery, 1986):

$$BIC = G^2 - df \cdot \ln n\,, \tag{3}$$

where n – total sample size.

The model that minimizes AIC and BIC will be chosen. A rule of thumb to determine the degrees of freedom for the table without zeros is $df =$ number of cells – number of free

parameters. In order to test the goodness-of-fit of a model that uses an observed set of marginal totals with at least one zero entry, we must reduce the degrees of freedom associated with the statistic. The reason for this is that if an observed margin entry is zero, both expected and the observed entries for all cells is known to be perfect once it is observed that the marginal entry is zero. As a result, we must delete those degrees of freedom associated with the fit of the zero cell values. A general formula for computing degrees of freedom in cases where some of the margins fitted contain sampling zeros is as follow (Fienberg, 1980):

$$df = (T_e - Z_e) - (T_p - Z_p), \tag{4}$$

where: T_e – number of cells in table that are being fitted, T_p – number of parameters fitted by model, Z_e – number of cells containing zero estimated expected values, Z_p – number of parameters that cannot be estimated because of zero marginal totals.

Log-linear analysis is a widely known statistical method for analyse categorical data in contingency tables. Although the log-linear models are versatile statistical models, there are some limitations in using them. The limitations are largely due to zero cells that may arise in contingency table. The consequence of the zero cells problem can be twofold. First, we cannot include many variables in the analysis. This is related to the second consequence: eventual collapsing of the variables in order to avoid zeros cells in the table, which may distort the process being modeled and may result in a loss of some valuable data. Also odds and odds ratios are undefined with zeros in the denominator (Ishii-Kuntz, 1994).

Other than collapsing variable categories, several options are available for analyzing table with zero cells:

1. add a small value (0.5 is frequently suggested) to every cell in the table when fitting the saturated model (Goodman, 1970),
2. add a small quantity (such as 0.2) only to zero cells (Evers and Namboodiri, 1977),
3. add the value $\frac{1}{r}$ to zero cells, where r equals the number of response categories (Grizzle et al., 1969),
4. arbitrarily define zero divided by zero to be zero (Fienberg, 1980),
5. increase the sample size sufficiently to remove all zeros cells (Knoke and Burke, 1980),
6. replace sampling zeros by 0.1×10^{-8}, or a smaller number and then check results against those obtained without such an adjustment (Clogg and Eliason, 1988).

Technically sampling and structural zeros are treated in the same way. The reason is that in any test statistic, a term corresponding to a cell with zero count will cancel out. Only for the

saturated model is it necessary that the table is complete with no zeros (Smirnoff, 2003). Researchers need to consider carefully the limitations of log-linear models in order to analyze the categorical data in the table effectively.

The comparison of these approaches in the log-linear analysis for contingency table containing zeros will be presented.

4. Application in R.

Data come from the Central Statistical Office of Poland from the Local Data Bank and show number of deadly injured in accidents at work in the first three quarters 2013. Two-dimensional contingency table for 2 variables: *Voivodeships* (1. Dolnośląskie, 2. Kujawsko-pomorskie, 3. Lubelskie, 4. Lubuskie, 5. Łódzkie, 6. Małopolskie, 7. Mazowieckie, 8. Opolskie, 9. Podkarpackie, 10. Podlaskie, 11. Pomorskie, 12. Śląskie, 13. Świętokrzyskie, 14. Warmińsko-mazurskie, 15. Wielkopolskie, 16. Zachodnio-pomorskie) and *Cause of the accident* (1. Electricity, 2. Explosion, fire, 3. Ignition, 4. Slipping in the fall or collapsing of the material factor, 5. Slipping, falling of the person) is analyzed. Out of 80 cells (16×5), 48 cells contain zeros. The sample size is 55.

Now we compare some options of transformation for dealing with zero cells in log-linear analysis. As the saturated model should be build only for non zero cells, this model will not be analyzed. The effects of zero cell action for the independence model $[Voivodeship][Cause]$ are presented in Table 1.

Adjustment	G^2	df	G^2/df	AIC	BIC
No adjustment for zeros	47.380	60	0.790	-72.620	-193.060
$n \to n + 0.5$	21.331	60	0.356	-98.669	-219.109
$n = 0 \to n + 0.2$	21.331	60	0.356	-98.669	-219.109
$n = 0 \to n = 0.1 \times 10^{-8}$	21.331	60	0.356	-98.669	-219.109

Table 1 Adjustment and goodness of fit criteria for two-way zero-cells table. Source: own calculations in **R** based on the data from the Central Statistical Office (www.stat.gov.pl).

The goodness of fit statistics for no adjustment seems to be the best for independence model. The comparison for other transformations show that the result for conducted adjustments is the same and no significant differences are seen.

The second example is based on data Titanic{datasets} summarized in four-way table. Data provides information on the fate of passengers on the fatal maiden voyage of the ocean liner Titanic summarized according to variables: *class*, *sex*, *age* and *survival*. Out of 32 cells ($4\times2\times2\times2$), 8 cells are zeros. The sample size is 2201.

The effects of zero cell action for the independence model $[Class][Sex][Age][Survived]$ are summarized in Table 2.

Adjustment	G^2	df	G^2/df	AIC	BIC
No adjustment for zeros	1243.663	25	49.747	1193.663	1051.246
$n \to n+0.5$	1216.387	25	48.655	1166.387	1023.970
$n=0 \to n+0.2$	1229.594	25	49.184	1179.594	1037.177
$n=0 \to n=0.1\times10^{-8}$	1243.663	25	49.747	1193.663	1051.246

Table 2 Adjustment and goodness of fit criteria for [Class][Sex][Age][Survived] model. Source: own calculations in **R**.

Table 2 shows that the best transformation for zero cells is adding to all cells 0.5. The likelihood criteria, as well as information criteria, are minimum for this adjustment, and G^2/df is closest to 1. This results show that the best fit is obtained for the adjustment where we add 0.5 to each cell in the table.

As the interaction between *Class* and *Survived* seems to be interesting, the second model tested is the conditional association model $[ClassSurvived][Sex][Age]$. The goodness of fit statistics are summarized in Table 3.

Adjustment	G^2	df	G^2/df	AIC	BIC
No adjustment for zeros	1062.762	22	48.307	1018.762	893.435
$n \to n+0.5$	1036.113	22	47.096	992.113	866.786
$n=0 \to n+0.2$	1049.425	22	47.701	1005.425	880.098
$n=0 \to n=0.1\times10^{-8}$	1062.762	22	48.307	1018.762	893.435

Table 3 Adjustment and goodness of fit criteria for [Class Survived][Sex][Age] model. Source: own calculations in **R**.

The result is the same as for the previous model. The best fit is obtained for adjustment where 0.5 is added to each cell.

As the interaction between *Class*, *Sex* and *Survived* seems to be interesting, the next model tested is the conditional association model $[ClassSurvivedSex][Age]$. The goodness of fit statistics are summarized in Table 4.

Adjustment	G^2	df	G^2/df	AIC	BIC
No adjustment for zeros	225.338	15	15.023	195.338	109.888
$n \to n + 0.5$	210.616	15	14.041	180.616	95.166
$n = 0 \to n + 0.2$	214.707	15	14.314	184.707	99.257
$n = 0 \to n = 0.1 \times 10^{-8}$	225.338	15	15.023	195.338	109.888

Table 4 Adjustment and goodness of fit criteria for [Class Survived][Age] model. Source: own calculations in **R**.

Table 4 shows similar result: the best fit is obtained for adjustment where we add 0.5 to each cell. The results presented in table 2, 3 and 4 prove that the best transformation of zero cell in multi-way table is obtained when we add 0.5 for each cell. The other adjustments are not significantly worse and the differenced in the ft is not big.

The differences between presented adjustments for two—and four-way table are not significantly different and the results obtained may not be generalized, as some of them show only very little improvement. However for four-way table adding 0.5 to each cell in the table leads to the best fitting model.

Conclusions

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 in the table as a function of all the variables in the survey. The analysis of a such table may be troublesome, when some cells are zeros. For log-linear models, most of the derivations of expected frequencies and other quantities assume $n_{hj} > 0$, however in the research we may have tables containing zeros. Zero frequencies may occur in contingency table for two reasons: sampling and structural zeros. To avoid this problem, some action is needed.

In this paper some solutions for zero-cells frequencies are presented. For two- and four-way table different solutions are compared with the use of likelihood statistic and information criteria (*AIC* , *BIC*). This analysis may be helpful in looking for the best solution in the analysis of high-dimensional tables.

Acknowledgements

The project is funded by a grant from the National Science Centre based on the decision DEC-2012/05/N/HS4/00174.

Refferences

Akaike, H. (1973). *Information theory and an extension of the maximum likelihood principle*. In Petrow B. N., & Czaki F. (Eds.), *Proceedings of the 2nd International Symposium on Information,* Budapest: Akademiai Kiado.

Andersen, E. B. (1997). *Introduction to the statistical analysis of categorical data*, Springer New York.

Clausen, S. E. (1998). *Applied correspondence analysis. An introduction*, Sage Publications, Thousand Oaks.

Clogg, C. C., & Eliason, S. R. (1988). Some common problems in log-linear analysis, Long J. S. (Eds.) *Common Problems/Propers Solutions*, 226-257, Newbury Park, CA: Sage.

Everitt, B. S. (1977). *The analysis of contingency tables*, 2nd edition, Chapman & Hall/CRC.

Evers, M., & Namboodiri, N. K. (1977). A Monte Carlo assessment of the stability of log-linear estimates in small samples. *Proceedings of the American Statistical Association, Social Statistics Section*. Washington, DC: American Statistical Association.

Fienberg, S. E. (1980). *The analysis of cross-classified categorical data*. MIT Press, Cambridge.

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

Greenacre, M. J. (1984). *Theory and Applications of Correspondence Analysis*, Academic Press, London.

Grizzle, J. E., Starmer, C. F., & Koch, G. C. (1969). Analysis of categorical data by linear models. *Biometrics*, 26, 489-504.

Ishii-Kunts, M. (1994). *Ordinal log-linear models*. Sage University Paper Series on Quantitative Applications in the Social Science, series no. 07-097, Beverly Hills and London Sage.

Knoke, D., & Burke, P. J. (1980). *Log–linear model.* Sage University Paper Series on Quantitative Applications in the Social Science, series no. 07-020, Beverly Hills and London Sage.

Raftery, A. E. (1986). Choosing models for cross-classification. *American Sociological Review*, *51,* 145-146.

Smirnoff, J. S. (2003). *Analyzing categorical data.* Springer Texts in Statistics, Springer.