Comparison of Log-Linear Analysis and Correspondence Analysis in Two-Way Contingency Tables: A Medical Application

Ersin Öğüş, Ayşe Canan Yazıcı

Department of Biostatistics, Medical Faculty of Başkent University, Ankara, Turkey

ABSTRACT

Objective: The aim of the study was to examine log-linear analysis and correspondence analysis by an application with regard to advantages and disadvantages of the two methods.

Material and Method: In this study, we used an artificial data set, which is expanded without changing its nature, from a study conducted in Medical Faculty of Baskent University, Infectious Disease and Clinical Microbiology Department. Relations and interactions between variables and among subcategories have been investigated with log linear and correspondence analysis.

Results: It has been shown that when the assumptions of log-linear analysis were not available, better understanding of the data set could be obtained by correspondence analysis (CA). However, conclusions about the data may not be generalized in a confidence interval to the population by CA. In addition, in log-linear analysis, it is possible to make inferences about the population on the basis of sample data. The other important result obtained from the study was that some of the relations between the same variable categories could be detected by CA, which is not possible with log-linear analysis.

Conclusion: As a result, we suggest the complementary use of log-linear analysis and correspondence analysis for detailed results.

Key Words: Correspondence analysis, log-linear analysis, inertia, saturated models, unsaturated model

Received: *25.09.2009* ***Accepted:*** *08.12.2009*

Introduction

In some clinical researches, data sets are discrete or converted into categorical form after collected. In the case of categorical data, relations between two variables are looked for. Traditional approaches to categorical data are relied on the chi-square test. Chi-square test is performed for a two-way contingency table. On the other hand, chi- square is insufficient when you have more than two qualitative variables, and it also tests the independence of the variables. When you have more than two, it cannot detect the varying relations and interactions between the variables.

Log-linear analysis is an extension of the two-way contingency table where the conditional relations between two or more discrete variables are analyzed by taking the natural logarithm of the cell frequencies within a contingency table. It is a goodness-of-fit test that allows one to test all the effects (the main effects, the association effects and the interaction effects) at the same time (1, 2).

Correspondence analysis (CA) is a multivariate method for exploring cross-tabular data by converting such tables into graphical displays, called 'maps', and related numerical statistics. It is primarily intended to reveal features in the data rather than to confirm or reject hypotheses about the underlying processes which generate the data (3, 4). Primary difference between CA and most other categorical data analysis techniques is using of models. In log-linear analysis, a distribution is assumed under which the data is collected, a model for the data is hypothesized, estimations are made under the assumption that this model is true, and then these estimates are compared with the observed frequencies to evaluate the model. In correspondence analysis, there is no assumption about the distribution and model hypothesis, but a generalized singular value decomposition of the data obtained to study the structure in the data (4, 5).

The aim of this study is to compare and discuss advantages and disadvantages of log-linear analysis and correspondence analysis to investigate if there is a combined approach to contingency tables for a best understanding of the data set by these techniques.

Materials and Methods

In this study, the data set, derived from a study conducted in Medical Faculty of Baskent University, Infection Disease and Clinical Microbiology Department was used to gen-

This study was presented as an oral presentation and published as an abstract in the abstract book of the 7th National Congress of Biostatistics abstracts book. (Application of Log-Linear Analysis and Correspondence Analysis as a complementary method in two-way contingency table. 7th National Congress of Biostatistics, 29 September - 2 October, 2004, Mersin).

Address for Correspondence: *Dr. Ayşe Canan Yazıcı, Department of Biostatistics, Medical Faculty of Başkent University, Ankara, Turkey*
Phone: +90 312 212 68 68 E-mail: aysecanan@yahoo.com

erate an artificial data set by MINITAB 14.0 Statistical Software Package. The aim was to increase the sample size without changing the singular value, row and column mass and dimension scores. One of the variables was stool type, which is produced by macroscopic examination, while the other was parasite type which is produced by microscopic examination with trichrom staining method in 2x2 table. Relations and interactions between variables and among sub-categories have been investigated with log linear and correspondence analysis. All statistical calculations were performed by MINITAB 14.0, NCSS 2004 and SPSS 11.5 statistical softwares (6-8).

Log-linear Analysis

The primary aim of log-linear analysis is to involve fitting the models to the observed frequencies in the cross-tabulation of categoric variables. The models are a goodness of fit test that allows you to test all the effects (the main effects, the association effects, and the interaction effects) at the same time.

The following model refers to the traditional chi-square test where two variables, each with two levels (2x2 table), are evaluated to see if a relation exists between the variables. $f$ij is the expected frequency of the cell *ij* of the contingency tables.

$$\text{Ln}\ (f_{ij}) = \mu + \lambda i^{I} + \lambda j^{J} + \lambda ij^{IJ} \qquad (1)$$

Ln (f *ij*) = is the log of the expected cell frequencies

μ = is the overall mean of the natural log of the expected frequencies

λ = terms each represent "effects" which the variables have on the cell frequencies

I and J = the variables

i and j = refer to the categories within the variables

λi^{I} = the main effect for variable I

λj^{J} = the main effect for variable J

λij^{IJ} = the interaction effect for variables I and J

This is the called as saturated model, because it includes all possible one-way and two-way effects. Given that the saturated model has the same amount of cells in the contingency table as its effects, the expected cell frequencies will always match the observed frequencies exactly, with no degrees of freedom remaining (2). In a 2 x 2 table, there are four cells, and in a saturated model involving two variables, there are four effects (μ, λi^{I}, λj^{J}, λij^{IJ}). Therefore, the expected cell frequencies will exactly match the observed frequencies. By setting some of the effect parameters to zero we can find a more general model that will isolate the effects that are demonstrating the data patterns best. This general model is called the unsaturated model. Unsaturated model is titled as the Independence Model because it lacks an interaction effect parameter between I and J. This model holds that the variables are unassociated:

$$\text{Ln}(f_{ij}) = \mu + \lambda_i^{I} + \lambda_j^{J} \qquad (2)$$

Log-linear analysis has some limitations because of the assumptions of the method; one of them is of adequate sample size. With log-linear models, one needs to have at least 5 times the number of cases of cells in one's data. If one does not have the required amount of cases, either the sample size needs to be increased or one or more of the variables eliminated. Another way adds to each cell correction term 0.5.

Another limitation is the size of expected frequencies. For all two-way relations, the expected cell frequencies should be greater than one. Upon failing to meet this requirement, the Type I error rate usually does not increase, but the power can be reduced to the point where analysis of the data is worthless. If low expected frequencies are encountered, the following could be done:

1. Accept the reduced power for testing the effects associated with low expected frequencies.
2. Collapse categories for variables with more than two levels, which means that one could combine two categories to make a "new" variable. However, if this is done, relations between the variables can be lost, resulting in a complete reduction in power for testing those relations. Therefore, nothing has been gained.
3. Delete variables to reduce the number of cells, but in doing so one must be careful not to delete variables that are associated with any other variables.
4. Add a constant to each cell (.5 is typical). This is not recommended because power will drop, and Type I error rate only improves minimally (1, 2, 9)

Correspondence Analysis

Correspondence Analysis investigates the relations between the variables in two or more ways cross tables. It was first proposed for analyzing two-way contingency tables. Data in each cell of the table are frequencies. These frequencies are positive integers or zeros. The primary goal of correspondence analysis is to transform a table of numerical information into a graphical display, facilitating the interpretation of this information (3 ,10, 11).

The general way to compute the solution for CA constitutes, firstly, transforming the contingency table into a table of contributions to the chi-square statistic after fitting a null model to the table. Secondly, applying the singular value decomposition (SVD), which is a generalization of the eigenvalue decomposition (EVD), to that table, and computing the eigenvalues and eigenvectors. Then, further matrix manipulations led? to the tables are required for plotting in ordination space as coordinate points of maps.

Assume that F= I x J contingency table:

I and J denote the categorical variables which have been collected for N objects or individuals. f_{ij} give the frequencies with which row category i occurs together with column category j. The row and column totals of the F matrix of frequencies are called the row total marginals and column total marginals, respectively (12-14).

$$F_{(ixj)}=[f_{ij}],\ f_{ij} \geq 0\ (i=1, 2,......,I),\ (j=1, 2,.....,J) \qquad (3)$$

$$f_{i+} = \sum_{j=1}^{J} f_{ij}, \quad f_{+j} = \sum_{i=1}^{I} f_{ij}, \quad f_{++} = \sum_{i=1}^{I}\sum_{j=1}^{J} f_{ij}$$

The goal of the analysis is to represent the entries in the table of frequencies in terms of the distances between individual rows and columns in a low dimensional space. The aim is to construct a low-dimensional joint map of objects and categories in Euclidean space R^p. Hence, chi-square distances can be computed between rows as well as between columns. If we consider chi-square distances between rows, these distances are computed on the profiles of the rows of a matrix, where the profile of row i is the vector of conditional proportions f_{ij}/f_{i+}. The Chi-square distances between rows i and *i'* of table F is given by

$$d^2(i,i') = n\sum_{j=1}^{J}\frac{(f_{ij}/f_{i+} - f_{i'j}/f_{i'+})^2}{f_{+j}} \tag{4}$$

Equation (4) denotes that $d^2(i,i')$ is a measure for the difference between the profiles of rows i and *i'*. It shows that since the entries of the table are corrected for the row marginals, proportional rows obtain zero distances. Remaning squared differences between entries are weighted heavily if the corresponding column marginals are small, while these differences do not contribute much to the chi-square distances if the column marginals are large. The configuration of I row points is located in a Euclidean space at dimension I-1. In a similar manner, Chi-square distances can be defined between columns of the crosstable (13, 15, 16).

The CA solution can be found by SVD solution. In order to derive the coordinates X of the row categories of table F in the new Euclidean space, we consider the SVD of the matrix of the observed frequencies minus the expected frequencies corrected for row and column marginals.

Let D_r and D_c be the diagonal matrices whose diagonal entries are respectively the marginal row proportions f_{i+} and column proportion f_{+j}, where it is assumed that $f_{i+}>0$ and $f_{ij}>0$.

$E=D_r tt$ and $E=D_r tt'D_c$, where t is a unit vector and elements are $e_{ij}=f_{i+}f_{+j}$.

SVD of the matrix;

$$D_r^{-1/2}(F-E)D_c^{-1/2} = U\Lambda V' \tag{5}$$

Elements of this matrix have the value $(p_{ij}-e_{ij})/e_{ij}^{1/2}$.

$$U'U = I \tag{6}$$

$$V'V = I \tag{7}$$

Λ is a diagonal matrix with singular values λ_α in descending order. The row and column scores are normalized as follows:

$$R = D_r^{-1/2}U \tag{8}$$

$$C = D_c^{-1/2}V \tag{9}$$

The relation between the row and column points is specified by the "transition formula"

$$\tilde{R} = D_r^{-1}FC \tag{10}$$

$$\tilde{C} = D_c^{-1}F'R \tag{11}$$

Using $\tilde{R}$ as coordinates for row points and $\tilde{C}$ as coordinates for column points, distances between row points and distances between column points are chi-square distances.

The measure of the dispersion of the profiles in multidimensional space is called inertia and computed as χ^2/N. The higher the inertia, the more spread out they are (3, 5, 15, 17-19).

Results

Relations between stool type and parasite type were analyzed by log-linear analysis and correspondence analysis. Contingency table of the variables was given in Table 1.

According to the log-linear analysis results, partial relations were significant (Table 2).

Significant parameter estimations were given in Table 3. Mucous and watery categories of the stool type variable and giardia category of the parasites type variable were statistically significant on the relations. There were negative interactions between mucous and giardia, watery and giardia, bloody mucous and giardia categories.

According to Table 4, of the 410 researchers 69% has classified as no parasites group, 16.1% has giardia parasites group, 7.6% has entamoeba group, and 7.3% has blastocyst parasites group.

Table 1. Stool type x parasites type contingency table

Stool Type	Parasites Type				
	No	Giardia	Entamoeba	Blastocyst	Total
Mucous	54	10	18	17	99
Watery	177	11	6	10	204
Bloody and mucous	46	0	6	2	54
Foamy	7	45	1	0	53
Total	284	66	31	29	410

Table 2. Partial relations

Effect	df	χ^2_L	p
Stool type	3	134.28	0.000
Parasites type	3	371.88	0.000
Stool type x parasites type	9	205.84	0.000

Table 3. Row profiles ($f_{ij}/f_{i.}$)

Parameter	Estimation	Z
Mucous	3.5553	2.48*
Watery	3.0445	2.10*
Giardia	4.5109	3.17**
Mucous x Giardia	-5.0217	-3.41***
Watery x Giardia	-4.4199	-2.98**
Bloody and mucous x Giardia	-6.1203	-2.91**

*: p<0.05, **: p<0.01, ***: p<0.001

Table 5 shows the decomposition of inertia with respect to the three principal axes. Each axis accounts for a part of the inertia, which is expressed as a percentage and graphically displayed in the form of a histogram. The first two dimensions account for 99% of inertia. The sum of the principal inertias is 0.6385.

Figure 1 shows the plot of categories to understand the relations between variables. The relations between stool type and parasites type were observed clearly on the first and second principal axis. Foamy category of the diarhea type variable and Giardia category of the parasites type were settled together on the second axis. It means that Giardia parasites cause to foamy diarrhea type. There was a negative correlation between Giardia parasite type and three stool types, mucous, watery, bloody and mucous categories, because they were settled on the opposite sides of the second axis. Blastocyte parasite category and entamoeba categories were located nearly in the plot. It means that there was similarity between the distribution of blastocyte parasite category and entamoeba categories. Two of them can cause to the mucous diarrhea type, since they are located near than other categories. The distance from the origin to each category point approximates the importance of that category point. Three categories that "no", "watery", "bloody and mucous" have no intense effect on the relations, because they were settled closely to the origin.

Table 4. Row profiles ($f_{ij}/f_{i.}$)

Stool Type	Parasites Type				
	No	Giardia	Entamoeba	Blastocyst	Total
Mucous	0.545	0.101	0.182	0.172	1.000
Watery	0.863	0.054	0.029	0.054	1.000
Bloody and mucous	0.852	0.000	0.111	0.037	1.000
Foamy	0.132	0.849	0.019	0.000	1.000
Column mass	0.690	0.161	0.076	0.073	1.000

Table 5. Contingency table results

Axis	Inertia	Proportion	Cumulative	Histogram
1	0.5276	0.8264	0.8264	**************** **************
2	0.1044	0.1636	0.9900	*****
3	0.0064	0.0100	1.0000	
Total	0.6385			

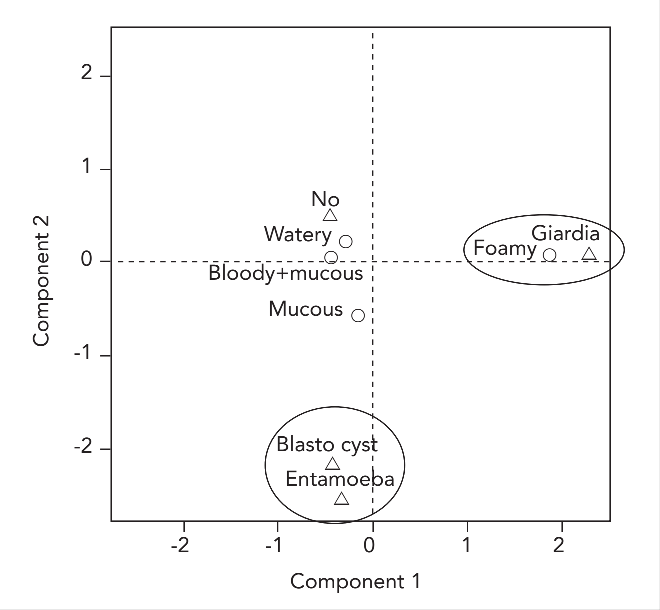

Figure 1. Correspondence analysis graphic (Asymmetric map)

Discussion

In literature, Van der Heijden and Worsley (1988) proposed to use loglinear analysis to detect interactions in a multiway contingency table, and to explore the form of these interactions with correspondence analysis, they showed that how the results can be used for confirmation (20).

Van der Heijden, de Falguerolles A, and de Leeuw J. A (1989) used loglinear analysis and CA for the decomposition of contingency tables and reported that there were cases in which these two techniques can be used complementary to each other. More specifically, they showed that often CA can be viewed as providing a decomposition of the difference between two matrices, each following a specific loglinear model. Therefore, in these cases the CA solution can be interpreted in terms of the difference between these loglinear models (5).

Moser EB (1989), used a different algorithm for correspondence analysis, through the analysis of several biological examples, to supplement the log-linear models approach to the analysis of contingency tables, both in the model identification and model interpretation stages of analysis. A simple two-way contingency table of tumor data is analyzed using correspondence analysis. This example emphasises the relationships between the parameters of the log-linear model for the table and the graphical correspondence analysis results. The technqiue is also applied to a three-way table of survey data concerning ulcer patients to demonstrate applications of simple correspondence analysis to higher dimensional tables with fixed margins (21).

Panogiotakos et al. (2004) presented a combined approach to the contingency tables analysis using correspondence analysis and log-linear models by applying both methodologies to an epidemiological database which include 848 individuals regarding coronary heart disease risk factors. They reported that applying CA can reduce the interaction parameters that are necessary for the classical log-linear models. Beyond the better understanding of the structure of the data the computational time may be significantly reduced (19).

In our study, we looked for relations and similarities between stool type and parasites type variables. Then, we saw that the log-linear analysis and correspondence analysis had given us very detailed information about the relations between the variables. We could detect the relations and interactions between the variables and variable categories by log-linear analysis, and observe the results about the relations between the variables and the similarities between their own categories by the correspondence analysis map. By the application of the log-linear analysis and correspondence analysis, various aspects of the data can be studied. We have visualized the associations and similarities between the investigated parameters set. However, some relations that we could not see by log-linear analysis have been seen by correspondence analysis. For example, according to the log-linear analysis results, there was not any significant relation between giardia and foamy stool types because of the small frequency of the cell. However, on the correspondence analysis map we saw the relation between these two categories. It means that, when the assumptions of the log-linear analysis are not available, correspondence analysis, which has no assumptions, could be useful. When the sample size is inadequate, the model is not fitting the data set, or the model selection is unsuccessful, using log-linear analysis is not correct. Another difference between two methods is that, we could not obtain the similarities between the same variable categories by the log-linear analysis, which we could obtain by the correspondence analysis. For example, we could observe the relations between blastocyst and entamoeba categories of the parasites type variable on the map. It was observed that the better and easily understanding of the structure of the data can be obtained by the graphic. As mentioned, we could not obtain a probability value about the hypothesis by correspondence analysis. So, conclusions about the data may not be generalized to the population, but in log-linear analysis it is possible to make inferences about the population on the basis of the sample data. It would be more useful if two methods are used in complementary way: a probability value about the hypothesis and interaction between the categories can be obtained by the log-linear analysis, and then relations and similarities can be observed on a correspondence analysis map.

By using these two methods in a complementary way and interpreting the results from a medical perspective, researchers could obtain inherent associations between the investigated parameters and design their decisions with a more effective way.

Acknowledgement

We would like to extend our appreciation to Assoc. Prof. Dr. Hande Arslan and Dr. Emine Kuru Inci from Infectious Disease and Clinical Microbiology Department of Medical Faculty of Baskent University who gave the data set.

Conflict of Interest

No conflict of interest was declared by the authors.

References

1. Agresti A. An Introduction to Categorical Data Analysis. 1st ed. USA: John Wiley & Sons; 1996;145-50.
2. Knoke D, Burke PJ. Log-Linear Models. 1st ed. USA: Sage Publications; 1980;8-17.
3. Greenacre M, Blasius J. Correspondence Analysis in the Social Sciences. 1st ed. London; Academic Press; 1994;3-111.
4. Hoffman DC, Franke GR. Correspondence Analysis: Graphical Representation of Categorical Data in Marketing Research. Journal of Marketing Research 1986;23:213-27.
5. Van der Heijden PGM, de Falguerolles A, de Leeuw J. A Combined Approach to Contingency Table Analysis Using Correspondence Analysis and Loglinear Analysis. Applied Statistics 1989; 38: 249-92.
6. MINITAB, Version 14.0, MINITAB Inc. 2006.
7. NCSS: Number Cruncher Statistical System, Version 2004. J Hintze, Kaysville, UT, 2003, in Basic and Clinical Biostatistics, Dawson B, Trapp RG, New York; Lange Medical Books/McGraw-Hill; 2004. pp. ix, x.
8. SPSS: Statistical Program for Social Sciences, Version 11.5, Spss Inc. Chicago, Il, 2003.
9. Theus M, Lauer SRW. Visualizing Loglinear Models. Journal of Computational and Graphical Statistics 1999;8:396-412.
10. Andersen EB. Introduction to the Statistical Analysis of Categorical Data. 1st edition. Germany: Springer Verlag 1997;217-24.
11. Yazici, A.C. The application and the Interpretation of the Correspondence Analysis in Discrete Data Obtained from Biological Events. PhD Thesis. Ankara University, Ankara 2003.
12. Hawkins DM. Topics in Applied Multivariate Analysis. 1st edition. USA: Cambridge University Press; 1982;183-206.
13. Escofier B, Pages J. Presentation of Correspondence Analysis and Multiple Correspondence Analysis with the Help of Examples, In: Devillers J and Karcher W, eds. Applied Multivariate Analysis in SAR and Environmental Studies, Netherlands: Kluwer Academic Pub;1991;1-31.
14. Friendly M. Mosaic Displays for Multi-Way Contingency Tables. Jasa 1994;89:190-200.
15. Greenacre M, Hastie T. The Geometric Interpretation of Correspondence Analysis. Journal of the American Statistical Association 1987;82:437-47.
16. Michailidis G, De Leeuw J. The Gifi System of Descriptive Multivariate Analysis. Statistical Science 1998;13:307-36.
17. Gifi A. Nonlinear Multivariate Analysis, New York: John Wiley &Sons; 1990;265-72.
18. Greenacre M. Correspondence Analysis in Practice. 2nd edition. London: Academic Press; 2007;97-144.
19. Panogiotakos DB, Pitsavos C. Interpretation of Epidemiological Data Using Multiple Correspondence Analysis and Log-linear Models. Journal of Data Science 2004;2:75-86.
20. Van der Heijden PGM, Worsley KJ. Comment on "Correspondence Analysis Used Complementary to Loglinear Analysis". Psychometrica 1988;53:287-91.
21. Moser EB. Exploring Contingency Tables with Correspondence Analysis. Bioinformatics 1989;5:183-9.