Generalized log odds ratio analysis for the association in two-way contingency table

Luigi D'Ambra, Ida Camminatiello and Pasquale Sarnacchiaro

Abstract. The odds ratio is a measure of association used both for the analysis of a 2×2 contingency table and an $I\times J$ contingency table, where I and J are bigger than 2. Nevertheless, the total number of odds ratios to check grows with I and J and several methods have been developed to summarize them. In the present paper we present a general framework for the analysis of the complete set of log odds ratio. Particularly we propose and connect two different methodologies performed on two different data sets. Moreover starting from these methodologies, we focus our attention on the factorial representation of the log odds ratios.
Keywords: odds ratio**,** log-ratio analysis, factorial representation.

1 Introduction

[1] Luigi D'Ambra, Dipartimento di Economia, Management e Istituzioni – Università degli studi di Napoli Federico II ; email: dambra@unina.it

Ida Camminatiello, Dipartimento di Scienze Economiche e Statistiche – Università degli studi di Napoli Federico II; email: camminat@unina.it

Pasquale Sarnacchiaro, Facoltà di Economia Università Unitelma Sapienza; e-mail: pasquale.sarnacchiaro@unitelma.it

The odds ratio (OR) is one of the main measures of association in 2×2 contingency tables. For an $I \times J$ table the ORs are commonly used to describe the relationship between the row and column variables, in this case the total number of ORs to check is $\left[I\left(I-1\right)\right]/2 \times \left[J\left(J-1\right)\right]/2$. Nevertheless the number of ORs needed to capture the association structure may still be too large to gain insight into the nature of the relationship between the variables. This number can be reduced using spanning or adjacent ORs. Four main alternatives or complementary strategies have been developed to analyse the set of ORs. The first consists in the computation of measures of synthesis (Altham, 1970; Agresti, 1980). The second starts from the construction of the model for frequencies and studies the ORs through the interaction between the row and column variables. In this class there is the log-linear model for two-way contingency table; in fact the saturated log-linear model decomposes the observed logarithms of the cell probabilities in terms of log-linear parameters without imposing any restrictions on the data: $\log p_{ij}^{XY} = \mu_{**}^{XY} + \mu_{i*}^{XY} + \mu_{*j}^{XY} + \mu_{ij}^{XY}$, where μ_{**}^{XY}, μ_{i*}^{XY}, μ_{*j}^{XY} and μ_{ij}^{XY} are the overall effect, the one-variable X effect, the one-variable Y effect and the interaction effect, respectively. The log-linear interaction terms are closely related to the ORs. In fact, the values of the interaction term depend only on the values of the ORs in the table. As in the ANOVA framework, using effect coding and imposing the suitable restrictions, the interaction parameter can be estimated through the mean of the logs of the complete set of ORs. The third solution is the RC(1) association model (Goodman, 1979), which is more parsimonious than the usual log-linear model for the analysis of association in ordered two-way contingency table. The RC(1) association model was extended to the RC(M) association model to decompose the association into M components (Goodman, 1981). RC(M) association model can be interpreted using both an inner product rule and a distance rule (De Rooij and Heiser, 2005). In the second case it is possible to represent graphically, the ORs. The fourth strategy considers a Singular Value Decomposition (SVD) of the matrix containing the basic set of ORs (De Rooij and Anderson, 2007). After computing the standard coordinates for rows and columns, the authors propose to represent the ORs through the projection of the row points onto the vector column points.

After introducing notations in section 2, in section 3 we suggest a general framework for studying the OR structure for a two-way contingency table. In particular, our proposal is developed in two ways which are strongly related: the first one starts from an un-centred generalized principal component analysis of the OR data matrix and shows how, changing the weights system, it is possible to compute interesting measures of synthesis for the complete set of ORs. The second one proceeds from the logarithm transformation of the two way contingency table and performs an un-weighted and weighted SVD. The advantage of these methods is in the graphical displays; in fact it is possible to carry out a direct and un-direct factorial representation of the ORs. Moreover we show how the inertias of both methods are the same. We also give particular attention to the contact point of our proposal with unweighted (Aitchinson, 1990) and weighted (Greenacre, 2009) log-ratio analysis (LRA) and SVD of the log odds ratio structures (De Rooij and Anderson, 2007). This general framework has been developed for all the types of odds ratio (i.e. cumulation, continuation and global) and will be presented in later research.

2 Notations and odds ratio definition

Let $\mathbf{N} = (n_{ij})$ be a two-way contingency table that cross-classifies n units according to I row and J column categories of X and Y variables, respectively. The matrix of proportions is denoted by $\mathbf{P} = n^{-1}\mathbf{N}$ with general term p_{ij}. The marginal relative frequencies of the i-th row and j-th column of $\mathbf{P}$ are $p_{i\bullet}$ and $p_{\bullet j}$. Let the vector $\mathbf{r}$ have elements $p_{i\bullet}$ for i = 1, 2, . . . , I, and the diagonal matrix $\mathbf{D}_r$ have diagonal elements the coordinates of $\mathbf{r}$. Similarly, $\mathbf{c}$ is the vector of $p_{\bullet j}$ values for j = 1, 2, . . . , J, with $\mathbf{D}_c$ being the diagonal matrix of these values.

The association between X and Y can be described by the complete set of ORs, composed by $[I(I-1)]/2 \times [J(J-1)]/2$ ORs

$$OR_{ii'jj'} = \frac{n_{ij}n_{i'j'}}{n_{i'j}n_{ij'}} \quad 1 < i < I, \quad 1 < j < J$$

Let $\mathbf{G}$ be a two-way table of dimension $\tilde{I} \times \tilde{J}$ containing the complete set of ORs, where $\tilde{I} = I(I-1)/2$ and $\tilde{J} = J(J-1)/2$.
A measure of synthesis of these ORs is the Altham's index (Altham, 1970). For the complete set of ORs, it can be computed as follow:

$$\overline{OR} = \left(\frac{\sum_{i} \sum_{j} \left[\log OR_{ii'jj'}\right]^{Q}}{\tilde{I}\tilde{J}} \right)^{1/Q} \qquad Q \geq 1$$

The complete set is redundant and there exist basic sets of $(I-1)\times(J-1)$ ORs that capture all the information about the association between the variables: the *local ORs* and the *spanning ORs*.

3 Generalized log odds ratio analysis

Let $\mathbf{C}$=L($\mathbf{G}$) be a two-way table of dimension $\tilde{I} \times \tilde{J}$ containing the complete set of log ORs, in this table the rows (resp. columns) are formed by all pairs of categories of X (resp. Y). Let $\mathbf{B}$ and $\mathbf{D}$ be two square diagonal matrices of dimensions $\tilde{I}$ and $\tilde{J}$ respectively with general terms $1/\tilde{I}$ and $1/\tilde{J}$. Performing an Un-centred Generalized Principal Component Analysis (UGPCA) of $\mathbf{C}$ with weight matrices $\mathbf{B}$ and $\mathbf{D}$ we obtain a factorial plan in which we represent both pairs of categories of X and Y. Moreover,

we have $tr(\mathbf{B}^{1/2}\mathbf{C}\mathbf{D}\mathbf{C}^T\mathbf{B}^{1/2})$ is equal to the square of Altham's measure for $Q=2$, that is:

$$tr(\mathbf{B}^{1/2}\mathbf{C}\mathbf{D}\mathbf{C}^T\mathbf{B}^{1/2}) = \left(\frac{1}{\widetilde{I}\widetilde{J}}\sum_{i}\sum_{j}\left[\log OR_{ii'jj'}\right]\right)$$

This method does not take into account the weight structures of the rows and columns. Let $\widetilde{\mathbf{B}}$ and $\widetilde{\mathbf{D}}$ be two square diagonal matrices of dimensions $\widetilde{I}$ and $\widetilde{J}$ respectively with general terms $p_{i\bullet}p_{i'\bullet}$ and $p_{\bullet j}p_{\bullet j'}$. Performing an UGPCA of **C** with matrices $\widetilde{\mathbf{B}}$ and $\widetilde{\mathbf{D}}$ we get a weighted analysis of the log OR matrix (WUGPCA). In this case we show that:

$$tr(\widetilde{\mathbf{B}}^{1/2}\mathbf{C}\widetilde{\mathbf{D}}\mathbf{C}^T\widetilde{\mathbf{B}}^{1/2}) = \sum_{i<i'}\sum\sum_{j<j'}\sum p_{i\bullet}p_{i'\bullet}p_{\bullet j}p_{\bullet j'}\left[\log OR_{ii'jj'}\right]$$

therefore the WUGPCA can be seen as a decomposition of a synthesis measure of the Log ORs. This last one is a weighted version of Altham's measure.
UGPCA and WUGPCA allow to representing pairs of categories but don't permit to visualise the single categories. Starting from the logarithm of the matrix **N**, called L(**N**), we perform a SVD of the following double-centered matrix

$$\mathbf{Z}^U = (IJ)^{-1/2}(\mathbf{I}-(1/I)\mathbf{1}\mathbf{1}^T)L(\mathbf{N})(I-(1/J)\mathbf{1}\mathbf{1}^T)$$

where **1** denotes a vector of ones of appropriate order in each case. It is the unweighted LRA (Aitchinson, 1990).
The unweighted LRA is linked with UGPCA, in fact the inertia of $\mathbf{Z}^U$ is equal to $tr(\mathbf{B}^{1/2}\mathbf{C}\mathbf{D}\mathbf{C}^T\mathbf{B}^{1/2})$ and therefore to Altman measure. For improving this method, we can introduce a weighting system. In many situations, in the absence of additional information, the row and column margins of the original data table provide an excellent default weighting system. Here we choose the same masses $p_{i\bullet}$ and $p_{\bullet j}$ as in correspondence analysis. Then the matrix that we analyse is double-centered respect to its weighted row and column average

$$\mathbf{Z} = \mathbf{D}_r^{1/2}(\mathbf{I}-\mathbf{1}\mathbf{r}^T)L(\mathbf{N})(\mathbf{I}-\mathbf{1}\mathbf{c}^T)^T\mathbf{D}_c^{1/2} = \mathbf{D}_r^{1/2}\mathbf{A}\mathbf{D}_c^{1/2}$$

Performing a SVD of **Z** we obtain the weighted LRA (Greenacre, 2009), particularly

$$\mathbf{Z} = \mathbf{U}\boldsymbol{\Lambda}\ \mathbf{V}^T = \sum_{m=1}^{M}\mathbf{u}_m\lambda_m\mathbf{v}_m^T \quad \text{where } M = rank(\mathbf{Z})$$

with $\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I}$ where $\mathbf{u}_m$ is the m-th column of **U**, $\mathbf{v}_m$ is the m-th column of **V** and the singular values down the diagonal of $\boldsymbol{\Lambda}$ are in descending order $\lambda_1 > \lambda_2 > .... > \lambda_M$. The standard and principal coordinates for rows and columns are computed as follow

$$\mathbf{F} = \mathbf{D}_r^{-1/2}\mathbf{U}\boldsymbol{\Lambda}$$
$$\mathbf{G} = \mathbf{D}_c^{-1/2}\mathbf{V}\boldsymbol{\Lambda}$$
$$\widetilde{\mathbf{F}} = \mathbf{D}_r^{-1/2}\mathbf{U}$$
$$\widetilde{\mathbf{G}} = \mathbf{D}_c^{-1/2}\mathbf{V}$$

The weighted LRA is linked with WUGPCA; as a matter of fact, the total variance of Z is equal to $tr(\widetilde{\mathbf{B}}^{1/2}\mathbf{C}\widetilde{\mathbf{D}}\mathbf{C}^T\widetilde{\mathbf{B}}^{1/2})$, moreover this quantity can be evaluated in term of log ORs

$$tr\left(\mathbf{D}_r^{1/2}\mathbf{A}\mathbf{D}_c\mathbf{A}^T\mathbf{D}_r^{1/2}\right) = \sum_{m=1}^{M}\lambda_m^2 = \sum_{i<i'}\sum\sum_{j<j'}\sum\sum p_{i\bullet}p_{i'\bullet}p_{\bullet j}p_{\bullet j'}\left[\log OR_{ii'jj'}\right]^2$$

We verified empirically the row and column coordinates obtained by this analysis presents the following two important properties which characterize the RC(M) models.

$$OR_{ii'jj'} = \exp\left(\sum_{m=1}^{M}\lambda_m(\widetilde{f}_{im}-\widetilde{f}_{i'm})(\widetilde{g}_{jm}-\widetilde{g}_{j'm})\right) \quad (1)$$

$$\theta_{ii'jj'} = \exp\left(\frac{1}{2}d^2\left(\mathbf{f}_{i'},\mathbf{g}_j\right)+\frac{1}{2}d^2\left(\mathbf{f}_i,\mathbf{g}_{j'}\right)-\frac{1}{2}d^2\left(\mathbf{f}_i,\mathbf{g}_j\right)-\frac{1}{2}d^2\left(\mathbf{f}_{i'},\mathbf{g}_{j'}\right)\right) \quad (2)$$

Where $d^2\left(\mathbf{f}_i,\mathbf{g}_j\right)$ is the squared Euclidean distance between the points with coordinates $\mathbf{f}_i$ and $\mathbf{g}_j$.

This means the factorial representation of weighted LRA can be explained both in term of inner product rule (type I), both as distance rule (type II). In type I representation the interpretation is through the inner products, i.e. the association equals the length of a row vector times the length of a column vector times the cosine of the angle between the two vectors. In type II representation the OR can be expressed in term of squared Euclidean distances between the row and column point coordinates.

Thanks to these properties it is possible to visualize both the categories and the ORs, particularly the log-ORs can be represented as a combination of the corresponding categories (un-direct factorial representation) or as a point (direct factorial representation). In this last case pointing out in which quadrant of the factorial plan the OR is placed we can interpret the association.

Unfortunately, these important properties do not work for un-weighted LRA, so the graphical display resulting from the $\mathbf{Z}^U$ cannot be be interpreted using a distance rule or an inner product rule.

In order to compute a synthesis measure of the complete set of ORs, we can calculate the OR mean applying the previous formula, 1

$$Me\left(\theta_{ii'jj'}\right) = \sum_{i=1}^{I}\sum_{j=1}^{J}\exp\left(\frac{1}{2}\left[\left(\mathbf{f}_{i'}-\mathbf{g}_j\right)^2 p_{i'j}+\left(\mathbf{f}_i-\mathbf{g}_{j'}\right)^2 p_{ij'}-\left(\mathbf{f}_i-\mathbf{g}_j\right)^2 p_{ij}-\left(\mathbf{f}_{i'}-\mathbf{g}_{j'}\right)^2 p_{i'j'}\right]\right)$$

In conclusion we have presented a general framework for studying the log-OR structure for a two-way contingency table. Great attention has been given to the factorial representation. Particularly, we have shown that performing the SVD of log-OR matrix

(WUGPCA) and weighted LRA of the logarithm of original two-way contingency table we decompose the same inertia that represents a synthesis measure for the ORs. The same property has been shown for unweighted LRA and UGPCA. Afterwards, through the factorial graphical display we have a direct and un-direct factorial representation of the odds ratio for weighted LRA. This factorial representation has been obtained by the RC(M) model (De Rooij and Heiser, 2005) with similar results. However, in RC(M) model the number of components to be retained has to be chosen before to perform the analysis, while in our proposal this choice can be made after the analysis according to different criteria. For lack of space the case study with the relative direct and un-direct representation of the ORs will be presented during the conference.

References

Agresti, A.: Generalized odds ratios for ordinal data. Biometrics, 36, 59-67 (1980).

Aitchinson, J.: Relative variation diagrams for describing patterns of compositional variability. Mathematical geology, 22, 487-511 (1990).

Altham, P. M. E.: The measurement of association of rows and columns for an r x s contingency table. Journal of the Royal Statistical Society, B, 32, 63-73 (1970).

De Rooij, M., Anderson, C. J.: Visualizing, Summarizing, and Comparing Odds Ratio Structures. European Journal of Research Methods for the Behavioral and Social Sciences, 3(4), 139-148 (2007) doi: 10.1027/1614-2241.3.4.139.

De Rooij, M., Heiser, W.J.: Graphical representations and odds ratios in a distance-association model for the analysis of cross-classified data. Psychometrika, 70 (1), 99-122 (2005).

Goodman, L. A.: Simple models for the analysis of association in cross classifications having ordered categories. Journal of the American Statistical Association, 74, 537-552 (1979)..

Goodman, L. A.: Association Models and Canonical Correlation in the Analysis of Cross-Classification Having Ordered Categories. Journal of the American Statistical Association, 76, 320-334 (1981).

Greenacre, M.: Power transformations in correspondence analysis. CSDA, 53 (8), 3107-3116 (2009).