STUDY OF INFLUENCE IN LOGLINEAR MODELS OF DIRECT ESTIMATION

Rosana CEPEDA[1]
Claudia MARINELLI[1]
Nélida WINZER[2]

- *ABSTRACT: In order to perform influence studies on contingency tables in which a loglinear model of direct estimation fits adequately, the diagnosis measurements proposed by Andersen (1992) for an independent model can be generalized in a simple way. These generalizations allow to determine how much influence each cell has in the formulated model. Nevertheless these measurements do not give specific information about the interaction terms of the model. In this work we present an analysis complementary to that of Andersen that permits to explain the influence on the interaction terms. More specifically, by means of the definition of influence curve in Canonical Correlation Analysis for categorical data, it is possible to decide which cell has layer influence on the inertia in each marginal table, hence, achieving in this way, a more exhaustive study. The advantages of the proposed methodology are discussed through its application to different data sets.*

- *KEYWORDS: Influence curve; loglinear model; inertia; canonical correlation.*

1 Introduction

When a contingency table is analyzed by means of a Correspondence Analysis (CA), a Canonical Correlation Analysis (CCA) or a Loglinear Model (LLM) is important to determine if the model or the conclusion of the analysis has been influenced by a cell value. In Regression Analysis, Cook's distance, the standardized residuals and the leverage value, constitute the main set of diagnostic measures. For other models, including those for contingency tables, these measures also work if they are appropriately adapted. By omitting one term in the log-likelihood function, Cook's distance (Cook and Weisberg, 1982) is usually derived, namely the term corresponding to an observation under study. This procedure works when the terms in the log-likelihood function are independent. In a Loglinear Model under multinomial distribution, the observed frequencies and the corresponding terms in the log-likelihood function are not independent, but this disadvantage can be corrected by replacing the cell probabilities by conditional probabilities. On this basis, Andersen (1992) studied the influence for a

[1] Instituto Multidisciplinario de Ecosistemas y Desarrollo Sustentable, UNCPBA, Pinto 399, B7000GHG Tandil,Buenos Aires, Argentina. E-mail: *rcepeda@exa.unicen.edu.ar / cmarine@exa.unicen.edu.ar*
[2] Departamento de Matemática, UNS, Alem 1253, B8003JSM Bahía Blanca, Buenos Aires, Argentina. E-mail: *nwinzer@criba.edu.ar*

two-way contingency table, fitting a Loglinear Model of independence. Here the diagnosis measures for models of direct estimation on three way contingency tables are calculated.

It is also useful to analyze the influence on each term of the adjusted loglinear model. For the interaction terms it is possible to use the relationships between the Loglinear model analysis and other methodologies of the Multivariate analysis (van der Heijden et al., 1989; Goodman, 1986).

Tanaka (1989) and Romanazzi (1992) derived influence functions related to those multivariate analysis using the singular value decomposition.

In this work, the influence functions for Canonical Correlation Analysis are used to study the influence of a cell value on each interaction term in a loglinear model. Both approaches are applied to two data sets. One of them consists of survey data collected from freshmen students of Universidad Nacional del Sur, to analyze the association between alcohol consumption, age and sex. In the second set the frequencies of two species of cormorants, regions where they live and presence of adenovirus were registered to study the association between the illness and the two other conditions[3].

2 Diagnostic measures in loglinear models

Let n_{ijk} , $i=1,\ldots,\ I;\ \ j=1,\ldots\ J$; and $k=1,\ldots,\ K$ be the observations of a Multinomial distribution with cell probability π_{ijk} , depending on P parameters, $\pi_{ijk} = \pi_{ijk}\ (\theta_1,\ldots,\ \theta_P)$. Since the Fisher scoring estimation method (Aitchison and Silvey, 1958) can be considered an iterative weighed square minimum procedure (McCullagh and Nelder, 1989) it is possible to write:

$$\theta^{(n+1)} = (D'VD)^{-1} D'V\left[V^{-1}\left(x - n\pi^{(n)}\right) + D\theta^{(n)}\right],$$

where,

- x = (Nijk)'= (xz)' $1 \le z \le IJK$;
- $\pi^{(n)} = (\pi_{ijk}^{(n)})$ is the cell probability evaluated at the parameter values obtained in the n-th step;
- D is a matrix of dimension (IxJxK)xP), whose elements are $d_{zp} = \partial(\ln \pi_z)/\partial\theta_p$;
- V is a diagonal matrix of dimension IxJxK of expected frequencies m_{ijk} ($M_{ijk} = n\pi_{ijk}^{(n)}$).

The leverage values are the diagonal elements h_{ijk} of the matrix $H = V^{1/2} D(D'VD)^{-1} D'V^{1/2}$, and the standardized residuals are obtained as:

$$r_{ijk} = \frac{\left(n_{ijk} - n\hat{\pi}_{ijk}\right)}{\sqrt{n\hat{\pi}_{ijk}\left(1 - \hat{\pi}_{ijk} - \hat{h}_{ijk}\right)}},$$

with asymptotic normal distribution (Haberman, 1973).

The analogous of Cook´s distance for a multinomial distribution is:

[3] Sanitary Evaluation of South Atlantic Project, Chubut, CENPAT-CONICET. Field Veterinary Program of the Wildlife Conservation Society.

$$DC(z) = \frac{1}{P}\left(\hat{\theta} - \hat{\theta}(z)\right)'(D'VD)\left(\hat{\theta} - \hat{\theta}(z)\right),$$

where $\hat{\theta}$ is the vector of maximum likelihood estimates based on all cell counts, and $\hat{\theta}(z)$ is the vector of parameters estimates obtained from the conditional multinomial distribution given the absence of observations in the $z = ijk$ cell (Andersen 1992). More exactly:

$$DC(z) = \frac{1}{P}\frac{\left(n_{ijk} - n\hat{\pi}_{ijk}\right)^2 \hat{h}_{ijk}}{n\hat{\pi}_{ijk}\left(1 - \hat{\pi}_{ijk} - \hat{h}_{ijk}\right)^2}.$$

These diagnostic measures can be adapted to the loglinear model for three way tables that have no need for iterative estimation methods (Agresti, 1990) i.e. independence models (X,Y, Z), joint independence (XY, Z) and conditional independence (XY, XZ) (Table 1). From the multinomial log-likelihood functions (conditioned to fixed n), the expected frecuencies, ($\hat{m}_{ijk}$) are calculated and replaced in the leverages, residuals and Cook´s distances expressions Table 2, where π_{hkl} are cell probabilities for three categorical variables, m_{hkl} are respective expected frequencies, λ_h^H are deviation of mean, and λ_{hk}^{HK} are the first order association parameter.

Table 1: Cell Probability, model expression and parameters of Loglinear Models of direct estimation

Model	Cell Probability	Loglinear Model	Parameters
Independence (X,Y,Z)	$\pi_{ijk} = \pi_{i..}\pi_{.j.}\pi_{..k}$ $\forall i, j, k$	$\log m_{ijk} =$ $\mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z$	I+J+ K-3
Joint independence (XY,Z)	$\pi_{ijk} = \pi_{..k}\pi_{ij.}$ $\forall i, j, k$	$\log m_{ijk} =$ $\mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY}$	I(J-1)+(K-1)
Conditional independence (XY, XZ)	$\pi_{ijk} = \pi_{i.k}\,\pi_{ij.} / \pi_{i..}$ $\forall i, j, k$	$\log m_{ijk} =$ $\mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ}$	I(J-1)+I (K-1)

Since the estimated parameters vector, is asymptotically normal, it is possible to use a χ^2 distribution with p degrees of freedom to generate confidence ellipsoids to evaluate Cook´s distance (Johnson and Wichern, 1992).

Table 2: Estimated expected frequencies, leverages, residuals and Cook Distances for direct estimation loglinear models

	(X, Y, Z)	**(XY, Z)**	**(XY, XZ)**
$\hat{m}_{ijk}$	$\frac{\left(n_{i..}n_{.j.}n_{..k}\right)}{n^2}$	$\frac{n_{ij.}n_{..k}}{n}$	$\frac{n_{ij.}n_{i..k}}{n_{i..}}$
$\hat{h}_{ijk}$	$\frac{n_{i..}n_{.j.}}{n^2}+\frac{n_{i..}n_{..k}}{n^2}+\frac{n_{.j.}n_{..k}}{n^2}-3\frac{n_{i..}n_{.j.}n_{..k}}{n^3}$	$\frac{n_{..k}}{n}+\frac{n_{ij.}}{n}-2\frac{n_{..k}n_{ij.}}{n^2}$	$\frac{n_{ij.}}{n_{i..}}+\frac{n_{i.k}}{n_{i..}}-\frac{n_{ij.}n_{i.k}}{n_{i..}}\left(\frac{1}{n}+\frac{1}{n_{i..}}\right)$
r_{ijk}	$\frac{n^2n_{ijk}-n_{i..}n_{.j.}n_{..k}}{\sqrt{n_{i..}n_{.j.}n_{..k}\left(n^2-n_{i..}n_{.j.}-n_{i..}n_{..k}-n_{.j.}n_{..k}+2\frac{n_{i..}n_{.j.}n_{..k}}{n}\right)}}$	$\frac{n_{ijk}-\frac{n_{ij.}n_{..k}}{n}}{\sqrt{\frac{n_{ij.}n_{..k}}{n}\left(1-\frac{n_{ij.}}{n}\right)\left(1-\frac{n_{..k}}{n}\right)}}$	$\frac{n_{ijk}-\frac{n_{ij.}n_{i.k}}{n_{i..}}}{\sqrt{\frac{n_{ij.}n_{i.k}}{n_{i..}}\left(1-\frac{n_{ij.}}{n_{i..}}\right)\left(1-\frac{n_{i.k}}{n_{i..}}\right)}}$
DC	$\frac{1}{P}\frac{r_{ijk}^2\hat{h}_{ijk}}{(1-\frac{n_{i..}n_{.j.}n_{..k}}{n^3}-\hat{h}_{ijk})}$	$\frac{1}{P}\frac{r_{ijk}^2\hat{h}_{ijk}}{(1-\frac{n_{ij.}n_{..k}}{n}-\hat{h}_{ijk})}$	$\frac{1}{P}\frac{r_{ijk}^2\hat{h}_{ijk}}{(1-\frac{n_{ij.}n_{i.k}}{nn_{i..}}-\hat{h}_{ijk})}$

3 Influences on the model parameters

The obtained diagnosis measures allow us to calculate the cell influence on the adjustment of the model, but not on each parameter. In order to measure the influence on the association terms the relation between Correspondence Analysis, Loglinear Models and Canonical Correlation will be used.

Loglinear Models and Correspondence Analysis can be used as complementary techniques in the analysis of a two-way contingency table, and such approach can be extended without problems to more general contingency tables (Greenacre, 1984, van der Heijden et al., 1985, 1989). More specifically, the Correspondence Analysis solution can be interpreted in terms of the difference between two specific Loglinear models (Appendix A).

For example, under an LLM model (XY, YZ), it is possible to obtain the interaction terms, $\lambda_{ij}^{XY}, \lambda_{ik}^{XZ}$, with a correspondence analysis made on each one of the two-way marginal tables: $\Pi^{k[ij]}$ (X and Y in interaction) and $\Pi^{j[ik]}$ (X and Z in interaction).

On the other hand, Canonical Correlation Analysis for categorical data is related to Correspondence Analysis (Goodman, 1985): main inertias obtained from AC for the two-way table (XY) agrees with the squared canonical correlations resulting in the (X, Y) Canonical Correlation Analysis (Appendix B). So, it is sufficient to study the influence for

the squared canonical correlation according to each marginal table to analyze influence on the marginal association terms.

The theoretical influence curve, *CI(x,T,F)* of a statistical functional *T* with distribution function *F*, is defined as:

$$\lim_{\varepsilon \to 0} \frac{T\{(1-\varepsilon)F + \varepsilon\delta_x\} - T(F)}{\varepsilon} = CI_{T,F}(x),$$

for all x where the limit exists. This expression is the first order term of the Taylor series (Hampel, 1974) for *T*. Actually, *F* is unknown and can be estimated with the empirical distribution $\hat{F}$, and it is denoted as the empirical influence curve (EIC).

Several authors (Tanaka, 1989, Romanazzi, 1992) have analyzed the trustworthiness of the results of those multivariate analyses associated with a generalized singular value decomposition by means of the influence function. The Canonical Correlation analysis is a special case of the generalized eigenvalue problem,

$$(A - \gamma_s B)u_s = 0,\ u_s' B u_s = 1,$$

where, $B = \Sigma_{11}$, is a definite positive, real and symmetric matrix pxp, and $A = \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ is a real and symmetric matrix pxp. Assuming that the eigenvalues of interest, $\gamma_s, s = 1,..,k \le p$, are simple, the influence function for eigenvalues can be calculated using the influence functions for *A*, $A^{(1)} = CI_{A,F}(x)$, and *B*, $B^{(1)} = CI_{B,F}(x)$, as

$$CI(x,\gamma_s) = \gamma_s^{(1)} = u_s'(A^{(1)} - \gamma_s B^{(1)})u_s,\ s = 1,..,k \le p.$$

The influence curve for the squared canonical correlations ($\gamma_s = \rho_s^2$), can be written as:

$$CI(x,\rho_s^2) = 2\rho_s\alpha_s\beta_s' - \rho_s^2\alpha_s^2 - \rho_s^2\beta_s^2,$$

under the conditions

$$u_j'\Sigma_{11}u_i = \delta_{ji} = v_j'\Sigma_{22}v_i,$$

and the corresponding canonical vectors,

$$\alpha_s = u_s'(x - \mu_1) \text{ y } \beta_s = v_s'(y - \mu_2).$$

The equation of the influence curve for the k-th canonical correlation allows to evaluate the influence each cell has on the k-th main inertia in CA. These inertias results from CA applied to the two-way marginal table containing both variables included in the interaction. Therefore, the influence curve on the corresponding canonical correlation, determines the influence on the association term in MLL. This methodology has to be applied in so many marginal tables as one-order significant association terms have been found in the model.

4 Applications

4.1 A survey on frenchmen students at UNS (National University in the South). The 1162 answers are registered in Table 3

The conditional independence Log-linear model with sex - age and sex - consumption association terms fit the data. The cells that exert greater influence on this adjustment correspond to women of 19 and 20 years who drink a little (Table 5, CD (z) = 1.601 and CD (z) =2.375 respectively). Both cells have high residuals and high leverage values, hence the influence could be due to a joint effect of these two components (Table 5, columns of h_{ijk} and r_{ijk}). For the association term sex - age (XY), 20 years-old women, exert the greatest influence (EIC=0.1421, Table 9), whereas for the association between sex and consumption (XZ) women who drink Medium are those that have the greatest influence (EIC=0.28, Table 7). It is observed that the cells corresponding to the different classifications according to consumption and age for men have anomalous values neither in residuals nor in the leverage values. On the other hand, they do not exert influence on the general model nor on the significant terms of interaction, although sex is significant for the model (Table 4). Finally, for the association terms, the cells with greater contribution to inertia agree with the category that has the greatest influence on the corresponding interaction term.

Table 3: Surveys Data: Sex, Consumption and age registered for 1,162 frenchmen students to the National University in the South. Argentina

Sex	Consumption	Age			
		17	18	19	20
Men (M)	Little (L)	60	144	87	73
	Medium (Me)	62	84	51	34
	Much (Mu)	6	13	4	6
Women (W)	Little (L)	111	181	69	61
	Medium (Me)	30	48	27	6
	Much (Mu)	2	1	1	1

Table 4: Partial and marginal test for a loglinear model applied to the data in Table 3

Effects	Partial Association			Marginal Association		
	df	X2	P	Df	X2	P
Consumption	2	866.77	0			
Age	3	153.97	0			
Sex	1	6.53	0.0106			
Consumption–Age	6	13.28	0.0388	6	9.46	0.1492
Consumption- Sex	2	65.97	0	2	62.14	0
Age- Sex	3	18.79	0.0003	3	14.97	0.0018
Consumption-Age-Sex	6	10.35	0.1106			

Table 5: Diagnostic measures calculated for a loglinear model applied to the data in Table 3

Cell	CD (z)	h_{ijk}	r_{ijk}	Cell	CD (z)	h_{ijk}	r_{ijk}
M17L	1.588	0.605	-2.949	W17L	0.036	0.745	-0.28
M18L	0.079	0.623	0.5698	W18L	0.009	0.721	0.125
M19L	0.123	0.607	0.807	W19L	1.601	0.758	-1.932
M20L	0.394	0.602	1.4936	W20L	2.375	0.766	2.417
M17Me	0.825	0.459	3.0008	W17Me	0.001	0.392	0.1197
M18Me	0.110	0.537	-0.888	W18Me	0.002	0.505	0.1177
M19Me	0.009	0.468	-0.31	W19Me	0.192	0.332	1.9363
M20Me	0.247	0.448	-1.69	W20Me	0.282	0.295	-2.5745
M17Mu	0.000	0.237	0.023	W17Mu	0.017	0.272	0.6825
M18Mu	0.034	0.405	0.703	W18Mu	0.081	0.431	-1.0331
M19Mu	0.049	0.258	-1.18	W19Mu	0.000	0.187	0.1151
M20Mu	0.004	0.215	0.37	W20Mu	0.004	0.134	0.4976

Table 6: Inertia Contribution (XZ) for a loglinear model applied to survey data

Consumption	Sex		Total
	Men	Women	
Little	7.99361	9.27140	17.26500
Medium	12.204	14.15572	26.36049
Much	6.31973	7.32995	13.64968
Total	26.5181	30.75707	57.27518

Table 7: Empirical Influence Curve calculated for XZ association term model applied to survey data

Consumption	Sex	
	Men	Women
Little	0	-0.0493
Medium	-0.0337	0.2843
Much	-0.0156	0.1846

Table 8: Inertia Contribution (XY) for a loglinear model applied to survey data

Age	Sex		Total
	Men	Women	
17	2.111236	2.448720	4.55996
18	0.562652	0.652593	1.21525
19	1.452967	1.685226	3.13819
20	2.569039	2.979703	5.54874
Total	6.695894	7.766242	14.46214

Table 9: Empirical Influence Curve calculated for XY association term in model applied to survey data

Age	Sex	
	Men	Women
17	0	-0.0124
18	-0.0008	0.0432
19	-0.0052	0.1262
20	-0.0065	0.1421

4.2 Southern Argentine Cormorants. Information of 96 cormorants were registered: two species, Imperial (Phalacrocorax atriceps) or Rock (Phalacrocorax magellanicus); two regions, Northern Gulf (NG) or Northern Patagonian (NP), in presence or not of adenovirus.

The conditional dependence loglinear model (species - disease) fits the data (Table 11). The cells that exert greater influence on this adjustment correspond to imperial cormorants that were negative to the presence of adenovirus (CD (z) =2.7429, Table 12). However, the other species, rock cormorant, negative to adenovirus, exerts greater influence on the significant term of interaction between the species and the disease (EIC=0.4526, Table 12). It is also observed that the greatest contribution to inertia agrees with the category that the greatest influence has on the interaction term (Table 13). Both analyses agree as to the category negative of the disease, this fact points out that this classification has a high weight in the whole model.

Table 10: Cormorant Data. Species, Disease and Region registered for 96 cormorants in Southern Argentina

Species	Region	Positive Disease	Negative Disease
Imperial	NG	37	9
	NP	7	3
Rock	NG	15	19
	NP	3	3

Table 11: Marginal and Partial Associations tests for a loglinear model applied to data in Table 10

Effect	*FD*	χ^2 Parcial Contrib.	Parcial Contrib. P	Marg. Contrib. χ^2	Marg. Contrib. P
Species	1	2.57105	0.108	2.57105	0.108846
Disease	1	7.94581	0.0048	7.94581	0.004823
Region	1	44.35073	0	44.35073	0
Species-disease	1	11.20106	0	11.11772	0.000856
Species- region	1	0.17107	0.6792	0.08774	0.767067
Disease- region	1	0.16217	0.687171	0.07885	0.778869

Table 12: Diagnostic Measures calculated for a loglinear model applied to cormorant data

Species	Region	Residuals		Leverages		CD (z)		EIC	
		Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg
Imperial	NG	0.183	-0.828	0.528	0.75	0.016	1.646	0	-0.120
	NP	-0.183	0.828	0.472	0.25	0.069	2.743	0	-0.120
Rock	NG	0	0.434	0.708	0.681	0	0.224	-0.119	0.453
	NP	0	-0.434	0.292	0.319	0	0.526	-0.119	0.453

Table 13: Inertia Contribution calculated for loglinear model applied to cormorant data

Category	Positive Disease	Negative Disease	Total
Imperial- NG	1.7897	3.2635	5.053
Imperial –NP	0.0454	0.0828	0.1283
Rock – NG	2.205013	4.0209	6.226
Rock – NP	0.197581	0.360294	0.55787
Total	4.237703	7.727576	11.96528

As usually happens in Regression analysis, cells with high residual and leverage values usually have considerable Cook's Distance. But there are cells that show high residuals or high leverages and are not influential in the model (cells M17Me and W20Me, W17L and W18L in Table 5).

In the given examples, high values of EIC are frequently present, but not always associated with categories with high inertias. A counterexample can be seen in the Rock-NP-Neg category, (Tables 12 and 13): EIC = 0.4526 (maximum value) and Inertia = 0.36 (maximum = 4.02)). A category with high inertia is a reasonable good candidate to be influential in an association term, and a high leverage point is usually bound to be influential in a classical linear regression, although more accurate measurements (such as Cook's distance for the regression) do not always confirm that this is the case.

Therefore, none of the calculated diagnostic measurements by itself allows a clear view of the cells or categories that are influential in the model or in the estimation of the association parameters. Consequently, it would be advisable to calculate all the measurements proposed in this work.

Conclusion

Diagnostic measures of model deviations and of the influence of particular data sets are used extensively in modern regression analysis. For contingency tables, and more generally for loglinear models, it is not the influence of individual observations which is of interest, but rather the contribution to a lack of fit model, the contribution to a significant association term or to the values of the parameter estimates from a single cell in the table, or a category of a variable must be evaluated. Hence, diagnostics for contingency tables take somewhat different forms.

Through the generalization of the diagnostic measures obtained by Andersen it is possible to know what observations have influence on the model fit. On the other hand, from the multivariate analysis relations it is possible to identify the influence cells for the interaction parameters. The application of both methodologies allows to study the influence on categorical data in a more exhaustive way, since one of them allows to detect the most influential category of each variable over each interaction term, and the other gives information on which of the cells in that category exerts the greatest influence on the model fit.

CEPEDA, R.; MARINELLI, C.; WINZER, N. Estudo de influência em modelo log-linear de estimação direta. *Rev. Mat. Est*., São Paulo, v.25, n.1, p.7-21, 2007.

- RESUMO: Para efetuar análises de influência nas tabelas de contingência onde um modelo log-linear de estimação direta ajusta-se adequadamente, as medidas de diagnóstico propostas por Andersen (1992) para um modelo de independência podem ser generalizadas de uma maneira simples. As mesmas permitem determinar quanto influi cada cela no modelo formulado. Porém, estas medições não fornecem informação específica sobre os termos de interação do modelo. Neste trabalho apresenta-se uma análise complementar àquela de Andersen para poder explicar a influência sobre os termos de interação. Mais especificamente, por meio da definição da curva de influência na Análise da Correlação Canônica para dados categóricos, é possível determinar qual das celas exerce maior influência sobre a inércia em cada tabela marginal, conseguindo assim um estudo mais exaustivo. As vantagens da metodologia proposta são discutidas através da sua aplicação a diferentes conjuntos de dados.
- PALAVRAS-CHAVE: Curva da influência; modelo log-linear; inércia; correlação canônica.

References

AGRESTI, A. *Categorical data analysis.* New York: John Wiley, 1990. 560 p.

AITCHISON J.; SILVEY S. D. Maximum likelihood estimation of parameters subject to restraints. *Ann. Math. Stat.*, Ann Arbor, v.29, p. 813-828, 1958.

ANDERSEN, E. Diagnostics in categorical data analysis. *J. R. Stat. Soc. Ser.B: Methods,* London, v.54, n.3, p.781-791, 1992.

COOK, D.; WEISBERG, S. *Residuals and influence in regression*. New York: Chapman and Hall, 1982. 230p.

ESCOUFIER, Y. L'analyse des tableaux de contingence simples et multiples. *Metron*: *Riv. Int. Stat*., Roma, v.40, n.1-2, p.53-77, 1982.

GIFI, A. *Non linear multivariate analysis*. New York: John Wiley, 1991. 602p.

GOODMAN, L. The analysis of cross-classified data having ordered and/or unordered categories: association models, correlation models and asymmetry models for contingency tables with or without missing entries. *Ann. Stat.,* Washington, v.13, p.10-69, 1985.

GOODMAN, L. Some useful extensions of usual correspondence analysis approach and the usual log-linear models approach in the analysis of contingency tables. *Int. Stat. Rev.,* Vooburg, v.54, n.3, p.243-309, 1986.

GREENACRE, M. J. *Theory and applications of correspondence analysis*. London: Academic Press, 1984. 364p.

HABERMAN, S. J. The analysis of residuals in cross – classification tables. *Biometrics*, Washington, v.29, p.205-220. 1973.

HAMPEL, F. The influence curve and its role in robust estimation. *J. Am. Stat. Assoc.*, Washington, v.69, p.383-393, 1974.

HEIJDEN P.; FALGUEROLLES, A.; LEEUW, J. A Combined approach to contingency tables analysis using correspondence analysis and log-linear analysis. *J. R. Stat. Ser.C: Appl. Stat.*, London, v.38, n.2, p.249-292, 1989.

HEIJDEN, P.; LEEUW, J. Correspondence analysis used complementary to loglinear analysis. *Psychometrika*, Williamsburg, v.50, p.429-447, 1985.

HEIJDEN, P.; MOOIJAART, A.; TAKANE, Y. Correspondence analysis and contingency table models. In: GREENACRE, M.; BLASIUS, J. *Correspondence analysis in the social sciences*. San Diego: Academic Press, 1994. p.79-111.

JOHNSON. R. A.; WICHERN, D. W. *Applied multivariate statistical analysis*. 3th ed. Prentice Hall, 1992. 642p.

LANCASTER, H. O. *The chi-squared distribution.* New York: Wiley, 1969. 356p.

McCULLAGH, P.; NELDER, J. *Generalized linear models.* 2nd ed. London: Chapman & Hall, 1989. 511p.

ROMANAZZI, M. Influence in canonical correlation analysis. *Psychometrika,* Williamsburg, v.57, n.2, p.237-259. 1992.

TANAKA, Y. Influence functions related to eigenvalue problems which appear in mathematics. *Commun. Stat.: Theory Methods*. New York, v.18, n.11, p.3991-4010, 1989.

Received in 12.12.2006.

Approved after revised in 09.05.2007.

APPENDIX A

Complementary tools of analysis – Correspondence Analysis (CA) and Loglinear Model (LM)

Let $N_{IxJ} = [n_{ij}]$ be a two way contingency table, $n_{i\bullet} = \sum_{j=1}^{J} n_{ij}$ and $n_{\bullet j} = \sum_{i=1}^{I} n_{ij}$ row and column totals and $n = \sum_{j=1}^{J} n_{\bullet j} = \sum_{i=1}^{I} n_{i\bullet}$ the general total. Let define Π as the associated probability matrix: $\Pi = \frac{1}{n} N$. Let D_r and D_c be diagonal matrices whose elements correspond to the marginal proportion row ($\pi_{i.}$) and column ($\pi_{.j}$) respectively. Let $E = D_r t t' D_c$ be the matrix with elements $e_{ij}=\pi_{i.}\ \pi_{.j}$, where t is a ones vector. From the singular value decomposition (SVD) of the matrix, $D_r^{-1/2}(\Pi - E)D_c^{-1/2}$, whose elements are proportional to the standardized residual, it is possible to obtain, $D_r^{-1/2}(\Pi - E)D_c^{-1/2} = U\Lambda V'$ where, $U'U=I=V'V$ and Λ is a diagonal matrix with singular values ordered in descending form, $\Lambda_{KxK} = diag(\mu_1, \mu_2, ..., \mu_K)$ K= min (I-1, J-1).

Therefore the standardized row and column scores are given from $R = D_r^{-1/2}U$, $C = D_c^{-1/2}V$,where $R' D_r R = I$, $C' D_c C = I$, $t' D_r R = 0$ y $t' D_c C = 0$.

Greenacre (1984), replacing R and C in the SVD decomposition previously mentioned, the reconstruction formula is obtained:

$$\Pi = E + D_r R \Lambda C' D_c \tag{1}$$

The previous expression indicates that the Correspondence Analysis decomposes the P matrix in two parts: one of independence (E) and the other a residual one.

Escoufier (1982) proposes a generalization of the Correspondence Analysis from a redefinition of the reconstruction formula

$$\Pi = G + S_r R \Lambda C' S_c \tag{2}$$

where G is an arbitrary matrix, with the same order as Π , S_r and S_c are diagonal matrices with weights for row and column categories respectively, and R, C and Λ can be derived by the singular value decomposition of $S_r^{-1/2}(\Pi - G)\, S_c^{-1/2}$ instead of $D_r^{-1/2}(\Pi - E)\, D_c^{-1/2}$. This new expression generalizes CA in two ways: a) it is possible to use different models from that of independence (G instead of E) and b) the weights of rows and columns do not have to be defined from marginal row and column.

For each cell ij the reconstitution formula can be rewritten as:

$$\pi_{ij} = \pi_{i.}\pi_{.j}\left(1 + \sum_{k=1}^{K} \mu_k r_{ik} c_{jk}\right) \tag{3}$$

Escoufier (1982) noticed that the sum $\sum_{k=1}^{K} \mu_k r_{ik} c_{jk}$, is small compared with the unit, so by (3):

$$\log \pi_{ij} \approx u + u_{1(i)} + u_{2(j)} + \sum_{k=1}^{K} \mu_k r_{ik} c_{jk} ;$$

where u=0, $u_{1(i)} = \log \pi_{i.}$ y $u_{2(j)} = \log \pi_{.j}$ and the r_{ik} , c_{jk} parameters are standardized by $R'D_r R = I = C'D_cC$ (van der Heijden and de Leeuw (1985); van der Heijden et al (1989); van der Heijden et al (1994)). Thus, CA allows to describe the interaction parameters of the Loglinear Models with a logmultiplicative approximation. For example, in a two-way classification table, the solution of the Correspondence Analysis (decomposition of (Π - E)) can be interpreted in terms of the difference between Loglinear Models (XY) and (X, Y) (Agresti, 1990). This analysis can be generalized without difficulty to multiple tables, P [a] [b] where a and b indicate the subgroups of two groups of variables A and B respectively, with $A \cap B = \varnothing$. Thus CA allows to study the difference between Loglinear Models (AB) and (A, B).

APPENDIX B

Relation between Correspondence Analysis (CA) and Canonical Correlation Analysis (CCA)

Let A and B be variables whose values are *1,...,I* and *1,....,J* respectively. Let $\pi_{ij}=P(A=i,\ B=j)$ with the condition $\pi_i=P(A=i)>0$ and $\pi_j=P(B=j)>0$. From the Fisher´s identity (Lancaster, 1969), the cell probability can be written like:

$$\pi_{ij} = \pi_{i\bullet}\pi_{\bullet j}(1+\sum_{k=1}^{m}\rho_k x_{ik} y_{jk}), \tag{1}$$

with $1 \le i \le I$ y $1 \le j \le J$, $m = min(I-1, J-1)$, and

$$\sum_{i=1}^{I} x_{ik}\pi_{i\bullet} = 0, \quad \sum_{i=1}^{I} x_{ik}x_{ik'}\pi_{i\bullet} = \begin{cases} 1 & si\ \ k=k' \\ 0 & si\ \ k\neq k' \end{cases}$$
$$\sum_{j=1}^{J} y_{jk}\pi_{\bullet j} = 0, \quad \sum_{j=1}^{J} y_{jk}y_{jk'}\pi_{\bullet j} = \begin{cases} 1 & si\ \ k=k' \\ 0 & si\ \ k\neq k' \end{cases}.$$

where x_{ik} $(1 \le i \le I,\ 1 \le k \le m)$ and y_{jk} $(1 \le j \le J,\ 1 \le k \le m)$ are A and B canonical scores respectively such that canonical variables $X_k = x_{Ak}$ and $Y_k = y_{Bk}$ have zero mean, unit variance and are uncorrelated. Thus, ρ_k represents the canonical correlation between X_k and Y_k ($\sum_{i=1}^{I}\sum_{j=1}^{J} x_{ik}y_{jk}\pi_{ij} = \rho_k$). Formula (1) is the well-known saturated canonical RC model (Gilula and Haberman, 1986). The Correspondence Analysis (CA) can be expressed, (appendix A), as:

$$\pi_{ij} = r_i c_j (1+\sum_{k=1}^{K}\frac{f_{ik}g_{jk}}{\rho_k}).$$

Therefore, CA can be seen as a CCA reparameterisation:

$$x'_{ik} = f_{ik} = \rho_k x_{ik}, \quad y'_{jk} = g_{jk} = \rho_k y_{jk},$$

with the following restrictions:

$$\sum_{i=1}^{I} x'_{ik}\,\pi_{i\bullet} = 0, \quad \sum_{i=1}^{I} x'_{ik}\,x'_{ik'}\,\pi_{i\bullet} = \rho_k^2\,\delta_{kk'},$$
$$\sum_{j=1}^{J} y'_{jk}\,\pi_{\bullet j} = 0, \quad \sum_{j=1}^{J} y'_{jk}\,y'_{jk'}\,\pi_{\bullet j} = \rho_k^2\delta_{kk'}.$$

On the other hand, considering that the correlation between two consecutive scores is zero, we have:

$$\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{\left(\pi_{ij}-\pi_{i\bullet}\pi_{\bullet j}\right)^2}{\left(p_{i\bullet}\ p_{\bullet j}\right)}=\sum_{i=1}^{I}\sum_{j=1}^{J}\pi_{i\bullet}\ \pi_{\bullet j}\left(\sum_{k=1}^{m}\frac{x'_{ik}\ y'_{jk}}{\rho_k}\right)^2$$
$$=\sum_{k=1}^{m}\frac{1}{\rho_k^2}\sum_{i=1}^{I}\left(x'^2_{ik}\ \pi_{i\bullet}\right)\sum_{j=1}^{J}\left(y'^2_{jk}\ \pi_{\bullet j}\right)=\sum_{k=1}^{m}\rho_k^2.$$

Then, CA total inertia agrees with CCA total correlation. It also can be seen, that the goodness of fit statistic is partitioned in m components, ρ_k^2, and each component has two parts: one related to the row contribution and another one related to the columns. The previous deduction can also be made for more than two classification models (Gifi, 1991 and Greenacre, 1984, pp.140-143).