VISUALIZATION OF THE SIGNIFICANT EXPLICATIVE CATEGORIES USING CATANOVA METHOD AND NON-SIMMETRICAL CORRESPONDENCE ANALYSIS FOR EVALUATION OF PASSENGER SATISFACTION

Ida CAMMINATIELLO[1]
PhD, Researcher,
University "Federico II" of Naples, Italy

E-mail: camminat@unina.it

Luigi D'AMBRA[2]
PhD, University Professor,
University "Federico II" of Naples, Italy

E-mail: dambra@unina.it

Abstract: *ANalysis Of VAriance (ANOVA) is a method to decompose the total variation of the observations into sum of variations due to different factors and the residual component. When the data are nominal, the usual approach of considering the total variation in response variable as measure of dispersion about the mean is not well defined. Light and Margolin (1971) proposed CATegorical ANalysis Of VAriance (CATANOVA), to analyze the categorical data. Onukogu (1985) extended the CATANOVA method to two-way classified nominal data. The components (sums of squares) are, however, not orthogonal. Singh (1996) developed a CATANOVA procedure that gives orthogonal sums of squares and defined test statistics and their asymptotic null distributions. In order to study which exploratory categories are influential factors for the response variable we propose to apply Non-Symmetrical Correspondence Analysis (D'Ambra and Lauro, 1989) on significant components. Finally, we illustrate the analysis numerically, with a practical example.*

Key words: *ANOVA; CATANOVA; Non-Symmetrical Correspondence; Passenger Satisfaction*

1. Model

Many authors analyzed categorical data taking their bearings from quantitative statistics. Some of this methods require transformation of the data before analysis, others (Light and Margolin, 1971; Onokogu, 1985) do not. We start form Onukogu's approach.

Let *A*, *B*, *C*, be the two explicative variables and the response respectively. Let *i=1,2,...,I* index the categories of *C*, *j=1,2,...,J* index the categories of *A* and *k=1,2,...,K* index the categories of *B*. Let n the number of units. Denote by *N* the three-way contingence table and by n_{ijk} the joint frequency. Let $n_{.jk} = \sum_{i=1}^{I} n_{ijk}$.

Onukogu developed the following linear model for analysis of data from three-way contingency table.

$$E\left(n_{ijk}/n_{.jk}\right)=\mu_i+\tau_{ij}+\beta_{ik}+\gamma_{ijk} \tag{1}$$

where μ_i, τ_{ij}, β_{ik} and γ_{ijk} represent the constant, *j*-th *A* effect, *k*-th *B* effect and their interaction for the *i*-th response, respectively.

Under model 2, the null (H_0) and alternative (H_1) hypotheses for testing the *A* effect, *B* effect and their interaction effect are defined as

$$H_{0A}:\tau_{ij}=0, \qquad H_{1A}:\tau_{ij}\neq 0$$

$$H_{0B}:\beta_{ik}=0, \qquad H_{1B}:\beta_{ik}\neq 0$$

and

$$H_{0A}:\gamma_{ijk}=0, \qquad H_{1A}:\gamma_{ijk}\neq 0 \tag{2}$$

respectively.

The purpose of CATANOVA is to obtain Sums of Squares (SS) and tests for these hypotheses.

2. Sums of Squares decomposition for categorical data

One of hurdles to be cleared in any analysis of variance concerns the definition and computation of Sums of Squares (SS). When the data are nominal, the usual approach of considering the total variation in response variable as measure of dispersion about the mean is not well defined. One way out is to introduce the analysis of variance in vector notation and in terms of projectors.

Let **A**, **B** and **C** be the binary indicator matrix related to the complete disjunctive coding of variables *A*, *B*, and *C* respectively. Let $\mathbf{Z}_{AB}$ be the indicator matrix that represent the interaction effect between *A* and *B*. The contingence table can be constructed as

$$\mathbf{N}=\mathbf{C}^T\mathbf{Z}_{AB} \tag{3}$$

Denote by $\Re_J$ the subspace generated by the columns of **A** and $\Re_J^{\perp}$ its orthocomplement subspace. The projection operators on $\Re_J$ and $\Re_J^{\perp}$ are constructed as

$$\mathbf{P}_A=\mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T, \qquad \mathbf{P}_A^{\perp}=\mathbf{I}_N-\mathbf{P}_A \tag{4}$$

In the same way we define:

$$\begin{aligned} \mathbf{P}_B&=\mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{A}^T, & \mathbf{P}_B^{\perp}&=\mathbf{I}_N-\mathbf{P}_B \\ \mathbf{P}_{AB}&=\mathbf{Z}_{AB}(\mathbf{Z}_{AB}^T\mathbf{Z}_{AB})^{-1}\mathbf{Z}_{AB}^T, & \mathbf{P}_{AB}^{\perp}&=\mathbf{I}_N-\mathbf{P}_{AB} \\ \mathbf{P}_M&=\mathbf{1}_N(\mathbf{1}_N^T\mathbf{1}_N)^{-1}\mathbf{1}_N^T, & \mathbf{P}_M^{\perp}&=\mathbf{I}_N-\mathbf{P}_M \end{aligned} \tag{5}$$

The Total Sum of Squares (TSS), the Between Sum of Squares (BSS) and the Within Sum of Squares (WSS) are constructed as

$$\begin{aligned}TSS &= tr(\mathbf{C}^T\mathbf{P}_M\mathbf{C})\\ BSS &= tr(\mathbf{C}^T(\mathbf{P}_{AB}-\mathbf{P}_M)\mathbf{C})\\ WSS &= tr(\mathbf{C}^T\mathbf{P}_{AB}\mathbf{C})\end{aligned} \tag{6}$$

To study the relationship between the response and the explicative variables, Light e Margolin defined the following directional measure:

$$R^2 = \frac{BSS}{TSS} \tag{7}$$

To test if the measure is significant, they proposed:

$$C_0 = (n-1)(I-1)R^2 \cong \chi^2_{(I-1)(JK-1)} \tag{8}$$

If the dependence relationship between the response and the explicative variables is significant, we proceed to test the different effects. Onukogu defined the following SS for testing different effects:

$$\begin{aligned}SS_A &= tr(\mathbf{C}^T(\mathbf{P}_A-\mathbf{P}_M)\mathbf{C})\\ SS_B &= tr(\mathbf{C}^T(\mathbf{P}_B-\mathbf{P}_M)\mathbf{C})\\ InteractionSS &= tr(\mathbf{C}^T(\mathbf{P}_{AB}-\mathbf{P}_A-\mathbf{P}_B+\mathbf{P}_M)\mathbf{C})\end{aligned} \tag{9}$$

where SS_A, SS_B and $InteractionSS$ are the sum of squares due to the A effect, B effect and their interaction effect.

If there is independence between the explicative variables, TSS can be decomposed as

$$TSS = SS_A + SS_B + InteractionSS + WSS \tag{10}$$

For testing main and interaction effects, Onukogu defined the following tests

$$\begin{aligned}\chi^2_A &= (n-1)(I-1)\frac{SS_A}{TSS} \cong \chi^2_{(J-1)}\\ \chi^2_B &= (n-1)(I-1)\frac{SS_B}{TSS} \cong \chi^2_{(K-1)}\\ \chi^2_{AB} &= (n-1)(I-1)\frac{InteractionSS}{TSS} \cong \chi^2_{(J-1)(K-1)}\end{aligned} \tag{11}$$

If there is association between the explicative variables, the previous components SS are not orthogonal and the decomposition (10) is not true. So Singh (1996) defined the following adjusted SS.

$$SS_{A/B} = tr(\mathbf{C}^T \mathbf{P}_{A/B} \mathbf{C})$$

$$= tr(\mathbf{C}^T (\mathbf{I}_N - \mathbf{P}_B)\mathbf{A}(\mathbf{A}^T (\mathbf{I}_N - \mathbf{P}_B)\mathbf{A})^{-1} \mathbf{A}^T (\mathbf{I}_N - \mathbf{P}_B)\mathbf{C}) \quad (12)$$

$$SS_{B/A} = tr(\mathbf{C}^T \mathbf{P}_{B/A} \mathbf{C})$$
$$= tr(\mathbf{C}^T (\mathbf{I}_N - \mathbf{P}_A)\mathbf{B}(\mathbf{B}^T (\mathbf{I}_N - \mathbf{P}_A)\mathbf{B})^{-1} \mathbf{B}^T (\mathbf{I}_N - \mathbf{P}_A)\mathbf{C}) \quad (13)$$

where $SS_{A/B}$ is the adjusted SS due to A variable effect after eliminating the B effect and $SS_{B/A}$ is the adjusted SS due to B variable effect after eliminating the A effect.

The interaction SS is obtained by subtraction as

$$IntSS = BSS - SS_A - SS_{A/B}$$
$$= BSS - SS_B - SS_{B/A} \quad (14)$$

As consequence *TSS* can be decomposed as

$$TSS = SS_A + SS_{B/A} + IntSS + WSS$$

$$= SS_B + SS_{A/B} + IntSS + WSS \quad (15)$$

For testing main and interaction effects, Singh (1996) defined the following tests

$$C_{01} = (n-1)(I-1)\frac{SS_{A/B}}{TSS} \cong \chi^2_{(I-1)(J-1)}$$
$$C_{02} = (n-1)(I-1)\frac{SS_{B/A}}{TSS} \cong \chi^2_{(I-1)(K-1)}$$
$$C_{012} = (n-1)(I-1)\frac{IntSS}{TSS} \cong \chi^2_{(I-1)(J-1)(K-1)} \quad (16)$$

3. Analysis of significant components

CATANOVA enables us to know if there is significant dependence between independent and dependent variables and which exploratory variables are significant to explain the response, but which exploratory categories are influential for the response variable?

In order to describe the dependence relationship between independent and dependent variables we propose to carry out Non-Symmetrical Correspondence Analysis (NSCA).

NSCA looks for the orthonormal basis which accounts for the largest part of inertia to visualize the dependence structure between the variables in a lower dimensional space. This leads us to the extraction of the eigenvalues λ_α and eigenvectors $\mathbf{u}_\alpha$ associated to the eigen-system

$$\left(\frac{1}{n}\right)\mathbf{C}^T\left(\mathbf{P}_{AB}-\mathbf{P}_M\right)\mathbf{u}_\alpha=\lambda_\alpha\mathbf{u}_\alpha \tag{17}$$

NSCA enables us to measure and visualize the strength of the asymmetrical relationship between the dependent and independent categories, to analyze which categories are significant for the response, to carry out confidence circles for identifying those categories that are not statistically influential in helping to explain the response.

The focus of this paper is to explore the significant components by NSCA.

Let us consider the matrix of the A variable effect after eliminating the B effect

$$\mathbf{S}_{A/B}=\underbrace{\mathbf{C}^T(\mathbf{I}_N-\mathbf{P}_B)\mathbf{A}}_{\mathbf{Q}^T}\underbrace{(\mathbf{A}^T(\mathbf{I}_N-\mathbf{P}_B)\mathbf{A})^{-1}}_{\mathbf{D}^-}\underbrace{\mathbf{A}^T(\mathbf{I}_N-\mathbf{P}_B)\mathbf{C}}_{\mathbf{Q}} \tag{18}$$

where $\mathbf{D}^-$ is a generalized inverse. NSCA leads us to the extraction of the eigenvalues λ_α and eigenvectors u_α associated to the eigen-system

$$\mathbf{Q}^T\mathbf{D}^-\mathbf{Q}\mathbf{u}_\alpha=\lambda_\alpha\mathbf{u}_\alpha \tag{19}$$

The response and A variable coordinates on α axis are given by

$$\boldsymbol{\psi}_\alpha=\sqrt{\lambda_\alpha}\mathbf{u}_\alpha \qquad \varphi_\alpha=\mathbf{D}^{-1/2}\mathbf{Q}\mathbf{u}_\alpha \tag{20}$$

respectively. It is easy to verify that

$$\varphi_\alpha^T\varphi_\alpha=\mathbf{u}_\alpha^T\mathbf{Q}^T\mathbf{D}^-\mathbf{Q}\mathbf{u}_\alpha=\lambda_\alpha \qquad \sum_\alpha\varphi_\alpha^T\varphi_\alpha=\sum_\alpha\lambda_\alpha \tag{21}$$

These coordinates are especially useful for describing the dependence relationship between the dependent and independent variables. In particular, A variable coordinates close to the origin will infer that their categories do not help predict the response categories. Response coordinates close to the origin indicate that very few explanatory categories are influential in determining the outcome of those response categories. Similarly coordinates far from the origin will highlight that, if they are associated with the explanatory variable, those categories are influential factors for the response variable. If a response category lies far from the origin then there will be explanatory factors that influence its position.

To complement the correspondence plots, more formal tests of the influence of particular categories may be made by considering the confidence circles for NSCA proposed by Beh and D'Ambra (2005). If one considers only the dependence between the row (predictor) and column (explanatory) variable, the C-statistic can be expressed in terms of the predictor coordinates such that

$$C_0 = \frac{(n-1)(I-1)}{\left(1-\sum_{i=1}^{I} p_{i..}^2\right)} \sum_{j=1}^{J}\sum_{\alpha=1}^{M} p_{.j.}\varphi_{j\alpha}^2 \cong \chi^2_{(I-1)(J-1)} \qquad (22)$$

where $p_{i..}$ and $p_{.j.}$ are the *i*-th and *j*-th marginal proportions so that $\sum_{i=1}^{I} p_{i..} = \sum_{j=1}^{J} p_{.j.} = 1$. Beh and D'Ambra showed that 95% confidence circles for the *j* explanatory column category represented in a two dimensional non-symmetrical correspondence plot has radii length

$$r_j^J = \sqrt{\frac{5.99\left(1-\sum_{i=1}^{I} p_{i..}^2\right)}{p_{.j.}(n-1)(I-1)}} \qquad (23)$$

Note that (23) depends on the *j*-th marginal proportion. Thus, for a very small classification in the *j*-th(explanatory) category, the radii length will be relatively large. Similarly, for a relatively large classification, the radii length will be relatively small.

4. A numerical example

The data was collected in a enterprize of Local Public transport in Naples. We will treat Satisfaction as the response variable and age and profession as the predictor variables. The response is measured on a scale ranging from 1 (Low) to 4 (High), the age has five categories ($<$18, 19-25, 26-40, 41-65, $>$65), also the profession has five categories (student, employee, housewife, pensioner, other). The passengers are 400.

For our table $R^2 = 0.06$ which has an associated C_0-statistic of 73.61 (p-value =0.001). Therefore we can conclude that the age and profession influence the Passenger Satisfaction. To further investigate the source of this asymmetrical relationship, we carry out the CATANOVA.

Table 1 shows the results of decomposition of $TSS = SS_A + SS_{B/A} + IntSS + WSS$.

Table 1. Decomposition of $TSS = SS_A + SS_{B/A} + IntSS + WSS$

Source	d.f.	SS	Test	p-value
Age (adjusted)	12	4.857	23.161	0.008
Profession	4	3.447	16.439	0.001
Age x Profession	48	7.134	34.016	0.016
Within	335	235.595		
Total	399	251.035		

Table 2 shows the results of decomposition of $TSS = SS_B + SS_{A/B} + IntSS + WSS$.

Table 2. Decomposition of $TSS = SS_B + SS_{A/B} + IntSS + WSS$

Source	d.f.	SS	Test	p-value
Profession (adjusted)	12	3.060	14.595	0.0583
Age	4	5.244	25.005	0
Age x Profession	48	7.134	34.016	0.016
Within	335	235.595		
Total	399	251.035		

CATANOVA shows that the age is a more influential factor in level of satisfaction than the profession. Moreover the age effect after eliminating the profession effect is significant at 1 %, the profession effect after eliminating the age effect is significant at 6 %. So we proceed to apply the NSCA on the matrix related age effect after eliminating the profession effect. Of course we could carry out the NSCA of each component.

The results of NSCA show that the first two factors explain the 97 % of variability, so they allow us to have a quite complete view of the dispersion of phenomenon.

The plan produced by first two factors (Figure 1) shows that passengers who are less than 18 and between 41 to 65 tend to be not too satisfied, those between 18 to 25 tend to be pretty satisfied, those who are more than 65 and between 26 to 40 tend to end up either not satisfied or very satisfied.

For Figure 1 95 % confidence circles have been included. Since the origin, which is associated with zero predictability of the response variable given the explanatory variables (ie. independence), does not lie within any of the circles, all of the categories of the age variable are statistically influential in helping to determine the passenger satisfaction.

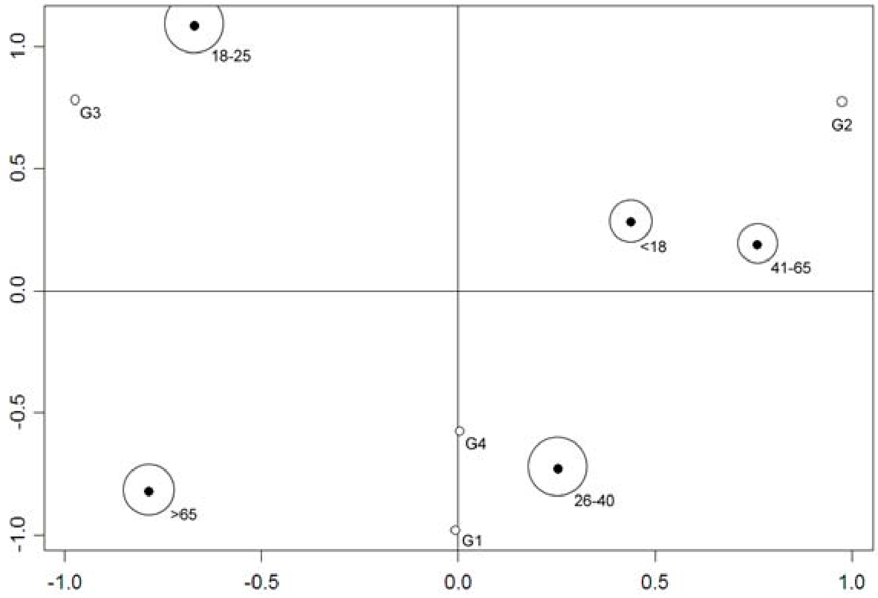

Figure 1. NSCA plot

5. Final remarks

We applied CATANOVA to study if there is significant association between independent and dependent variables and which exploratory variables are significant to explain the response.

In order to measure and visualize the strength of the asymmetrical relationship between the dependent and independent variables and to study which categories are significant to explain the response we proposed to carry out the NSCA of significant factors.

The combined approach has been applied on the data collected in a enterprize of Local Public transport in Naples obtaining interesting results

The paper focused on nominal data. In the next we will focus on ordered categorical variables.

References

1. Anderson, R . G. and Landis, R. **Catanova for multimendional contingency tables: nominal-scale response**, Communications in Statistics, 9, 1980, pp. 1191-1206
2. D'ambra, A., Ciavolino, E. and De Franco, D. **Un approccio statistico multivariato per la valutazione dell'utente nell'ambito dei trasporti: il caso AMTS**, "Proceedings of MTISD'06 Conference", Napoli, 2006
3. D'ambra, L., Beh, E. J. and Amenta, P. **Catanova for two-way contingency tables with ordinal variables using orthogonal polynomials**, Communications in Statistics, Theory and Methods, 34, 2005, pp. 1755-1769
4. D'ambra, L. and Lauro, N. C. **Non symmetrical analysis of three-way contingency tables**, In Coppi, R. and Bolasco, S. (Eds.), Multiway Data Analysis, North Holland, 1989 , pp. 301-314
5. Light, R. and Margolin, B. **An analysis of variance for categorical data**, Journal of the American Statistical Association, 66, 1971, pp. 534-544
6. Onukogu, I. B. **An analysis of variance of nominal data**, Biometrics Journal, 27, 1985, pp. 375-384
7. Singh, B. **On CATANOVA method for analysis of two-way classified nominal data**, Sankhya: The Indian Journal of Statistics, 58, 1996, pp. 379-388
8. Takeuchi, K., Yanai, H. and Mukeherjee, B. N**. The foundations of multivariate analysis**, Wiley, New York, 1981

[1] Ida Camminatiello is currently contract researcher (Formez). She holds a PhD in statistics from the University of Naples Federico II. She worked as a Lecturer at the Faculty of Economics, Second University of Naples, where she taught statistics, time series analysis and informatics. She has participated to numerous national and international conference. The main research topics are related to robust regression, multinomial logit model, categorical analysis of variance and non-simmetrical correspondence analysis with applications in environmental field, transport, customer satisfaction and sensory analysis. The most important and recent publications are:

1. Camminatiello, I., D'Ambra, L., Meccariello, G. and Della Ragione, L. **A study of instantaneous emissions through the decomposition of directional measures for three-way contingency tables with ordered categories,** Journal of Applied Sciences (to appear).
2. Camminatiello, I. and Lucadamo, A. **Estimating multinomial logit model with multicollinear data,** Asian Journal of Mathematics and Statistics (to appear).
3. Lombardo, R. and Camminatiello I. **CATANOVA for two-way cross classified categorical data,** Statistics: A Journal of Theoretical and Applied Statistics, 2009
4. Camminatiello, I. **A robust approach for partial least squares regression,** in "Metodi, modelli e tecnologie dell'informazioni a supporto delle decisioni, II. Applicazioni" Franco Angeli, Milano, 2008, pp. 31-38,

5. Camminatiello, I. and D'Ambra A. **Evaluation of Passenger Satisfaction using three-way contingence table with ordinal variables,** Rivista di Economia e Statistica del Territorio, 1, 2008, pp. 25-38

[2] Luigi D'Ambra is a professor of statistics, University of Naples Federico II. In December 2005 he has been invited by a School of Quantitative Methods and Mathematical Sciences (University of Western Sydney to participate to several conferences and to carry out research activity (Project title: The non-symmetrical correspondence analysis of ordinal categorical data). He has participated to numerous national and international conference as invited lecturer and discussant. He has been the chairman of steering committee of the schools of Italian Statistics Society on the " Statistical Methods for Customer Satisfaction Evaluation " and " Statistical Methods for of the Healthcare Services Evaluation " . He has been School President of the European Courses in Advanced Statistics - ECAS 2003 too. The main research topics are related to the non-symmetric techniques, to the analysis of multi-way data tables, with respect to qualitative and quantitative variables, and to the interpretative aspects of automatic classification. The most important and recent publications are:

1. Beh, E. and D'ambra, L. **Some interpretative tools for non-symmetrical correspondence analysis,** Journal of Classification, vol. 26, 2009, pp. 55-76
2. Beh, E., Simonetti, B. and D'ambra, L. **Partitioning a non-symmetric measure of association for three-way contingency tables,** Journal of Multivariate Analysis, vol. 98, 2007, pp. 1391-1411
3. Lombardo, R., Beh, E. and D'ambra, L. **Non-symmetric correspondence analysis with ordinal variables using orthogonal polynomials,** Computational Statistics & Data Analysis, vol. 52, 2007, pp. 566-577
4. Beh, E., Simonetti, B. and D'ambra, L. **Three-way ordinal non symmetrical correspondence analysis for the evaluation of the patient satisfaction,** Statistica & Applicazioni, 2006