Gazi University Journal of Science
GU J Sci
27(3):953-963 (2014)

Catanova Method for Determining of Zero Partial Association Structures in Multidimensional Contingency Tables

Hülya OLMUŞ[1,♠], Semra ERBAŞ[1]

Gazi University, Faculty of Science, Department of Statistics, 06500,Teknikokullar-Ankara, TURKEY

Received: 12/06/2013 Accepted: 28/12/2013

ABSTRACT

Zero partial association models correspond to the conditional independence relation of a variable pair, given the rest of variables in the model. For determining the model that best fits the data from the contingency tables measured at a nominal level, this study has used the Catanova test (with response as one of its variables, and factor as another one of its variables) instead of the chi-square test, which is used as the test statistic in Wermuth's backward elimination method with zero partial associations. Numerical analyses were performed on two samples, and the associations between these statistics were evaluated. Interpretations were provided for the obtained results.

Keywords: multidimensional contingency tables; zero partial association; CATANOVA; model search

1. INTRODUCTION

Modelling is a useful process both for prediction of future observables and for describing the relationships between factors. The saturated model, in other words the model that includes the effects of variables and interactions, always provides a perfect fit of the data. However, smaller models have more powerful interpretations and are often better predictive than large models. The goal is to find the smallest model that fits the data. To this end, many methods for model selection are available for multi-dimensional contingency tables. For this reason, this study has made use of the backward elimination model (Wermuth's method) among the different selection methods (Christensen, 1990). Wermuth (1976a) has proposed a backward elimination technique that is restricted to the decomposable models or multiplicative models. She focuses on identifying pairs of factors that can be viewed as conditionally independent or zero partial association (z.p.a.), that is, by partially independent pairs of variables. Either the likelihood ratio test statistic or the standard χ^2 statistic was used to test whether pairs with this z.p.a. were conditionally independent.

Many problems in analysis of nominal categorical data can be formulated in the general factor response framework. In such a set-up the appropriate measure of variation in response variable is the sum of squared deviations of individual observations from their mean. This definition, however, does not work for nominal data where the mean is an undefined concept.

This variation has examined by analysis of variance for categorical data (CATANOVA). Light and Margolin (1971) have used the Gini's measure of variation and proposed an ANOVA analogue method, known as

[♠] Corresponding author, e-mail: hulya@gazi.edu.tr

CATANOVA method, to analyze the categorical data. Margolin and Light (1974) have investigated exact small sample behaviour of CATANOVA and its two competitors namely, chi-square statistic and the likelihood ratio-statistic in two way classification table. They found that in small samples the null distribution of CATANOVA statistic is better approximated by a chi-square distribution than is the null distribution of chi-square statistic. The Chi-squared test requires the expected frequency be not very small, preferably at least five, while no such restriction is necessary for CATANOVA test. Hence, the CATANOVA method should be preferred for practical applications (Singh, 1993, 1996). Anderson (1980) has extended the CATANOVA methodology of Light and Margolin (1971) to multidimensional contingency tables obtained from the cross classification involving several factors and a response variable measured on a nominal scale.

Using an appropriate measure of total variation for multinomial data, partial and multiple association measures have developed as R^2 quantities which parallel the analogous statistics in multiple linear regression for quantitative data. He showed that the information obtained through association measures is useful as a basis for selecting a subset of factors which are most important for explaining the variability of a response variable.

In this study, the CATANOVA test was first provided for the Wermuth's backward elimination method; it was then provided for the two-way and multi-way contingency tables. In light of this information, and based on the assumption that one of the variables was a response variable, the two different methods were considered, and it was demonstrated that the CATANOVA test was applicable. The obtained results were also evaluated.

2. WERMUTH'S BACKWARD ELIMINATION METHOD FOR MULTIPLICATIVE MODELS IN MULTIDIMENSIONAL CONTINGENCY TABLES

The interrelations among several variables can more easily be understood if they can be characterized by a pattern of association. Multiplicative models form one class of such patterns of association. Wermuth's backward elimination technique is restricted to multiplicative models, therefore all patterns under consideration are interpretable in terms of zero partial association of variable pairs. This class of pattern is frequently being studied whenever independence hypotheses are tested in a contingency table.

A pair of variables X$_i$, X$_j$ having z.p.a. is conditionally independent on values of the other variables, and a constellation of z.p.a.s defines a model with a more or less complex pattern of conditional independencies between variable pairs, assuming underlying Poisson or multinomial distribution theory. A subclass of z.p.a. models, the 'multiplicative' or 'decomposable' models, characterized by the fact that the likelihood function can be factorized, has been extensively studied recently (Darroch, et. al., 1980; Edwars and Kreiner, 1983; Wermuth, 1976a, 1976b). In this method, test statistics can be calculated by using marginal associations, without the need for maximum likelihood estimations. To counter the risk of misinterpreting the marginal associations, it is necessary to test the zero partial associations of all possible variable pairs.

2.1. Definitions and notation

A multiplicative model or decomposable model states how a joint distribution may be factored into marginal distributions. Suppose that the joint distribution of four variables can be factored as

$$f(x_1x_2x_3x_4x_5) = \frac{f(x_1x_2x_3)\,f(x_1x_2x_4)}{f(x_1x_2)} \qquad (1)$$

where f denotes probability function. Using of these four variables, the joint distribution can have different multiplicative models. Generally, 1,2,3… numerical expressions are used instead of X_1, X_2,… variables, respectively. Then the notation for multiplicative models is 123 / 124 in equation (1). In this model, 3 and 4 variables are conditionally independent given (1,2) variable pairs, that is, this pair has z.p.a. Likelihood ratio tests may be used to evaluate whether a hypothesized model or pattern fits the data. The test statistics under multinomial distribution with observed cell counts n_{ijk} will be

$$\chi^2 = -2ln\prod\left[\frac{\frac{n_{i.K}n_{.jK}}{n_{..K}}}{n_{ijK}}\right]n_{ijK}, \quad i = 1,2,\dots,I; \quad j = 1,2,\dots,J \qquad (2)$$

where $n_{i.K} = \sum_j n_{ijK}$ and $n_{..K} = \sum_{ij} n_{ijK}$. These tests have $(I_i - 1)\left(I_j - 1\right)\prod_{r\epsilon K} I_r$ degrees of freedom (d.f) with I_l as the number of categories for the lth variable (Wermuth, 1976b). While the variable pairs (i,j) have z.p.a., K denotes all indices in the combination. The variable pair (3,4) was selected to have z.p.a for pattern 123 / 124. The corresponding chi-square statistics was computed as

$$2\left[\left(\sum n_{ijkl} lnn_{ijkl}\right) - \left(\left(\sum n_{ijk.} lnn_{ijk.}\right) + \left(\sum n_{ij.l} lnn_{ij.l}\right) - \left(\sum n_{ij..} lnn_{ij..}\right)\right)\right] \qquad (3)$$

The d.f. for z.p.a of (3,4) given variables 1 and 2 are $I_1 I_2 (I_3 - 1)(I_4 - 1)$.

3. ON THE ANALYSIS OF VARIANCE METHOD FOR NOMINAL DATA (CATANOVA)

Analysis of variance (ANOVA) is a method to decompose the total variation of the observations into sum of variations due to different factors and residual component. When the data are categorical, the usual approach of considering the total variation in response variable as a measure of dispersion about the mean is not well defined (D'ambra, et. al., 2005)

For a two-way or multi-way contingency table, the Pearson chi-squared statistic is commonly used when it can be assumed that the categorical variables are symmetrically related. However, for a two-way contingency table, it may be that one variable can be treated as factor and the second variable can be considered a response variable. For such a variable structure, the Pearson chi-squared statistic is not an appropriate measure of association. There are many situations in which the association between two categorical variables is not symmetric. The CATANOVA method enables us to know if there is significant dependence between in dependent and dependent variables and which factors are significant to explain the response (Camminatiello and D'ambra, 2010; Lombardo and Camminatiello, 2010).

3.1. CATANOVA for two-way contingency tables

Light and Margolin (1971) proposed an analysis of variance called CATANOVA for one response variable and one factor, both of them measured on a nominal scale.

Assume that there are g experimental groups and r unordered categories. Each response is in one and only one r categories. One common model for categorical data assumes that responses in different groups are stochastically independent and the responses in jth group follows a multinomial law:

$$P(n_{1j}, n_{2j}, \dots, n_{rj}) = (n_{1j}, n_{2j}, \dots, n_{rj}) \prod_{i=1}^{r} (p_{ij})^{n_{ij}} \qquad (4)$$

The null hypothesis in model (4) is that all the g groups have same multinominal probability structure.

$$H_0 : p_{ij} = p_i \ , \ j = 1,2, \dots g \qquad (5)$$

To study relationship between a response variable and a factor, let us calculate the following association measure:

$$R^2 = \frac{BSS}{TSS} \qquad (6)$$

Here, BSS, the between-group sum of squares, and TSS, the total sum of squares, are defined below:

$$TSS = (n/2) - (1/2n) \sum_{i=1}^{r} n_{i.}^2 \qquad (7)$$

$$BSS = \frac{1}{2} \sum_{j=1}^{g} \frac{1}{n_{.j}} \left(\sum_{i=1}^{r} n_{ij}^2 \right) - (1/2n) \sum_{i=1}^{r} n_{i.}^2 \quad (8)$$

where

$n_{i.} = \sum_{j=1}^{g} n_{ij}$ and $n_{.j} = \sum_{i=1}^{r} n_{ij}$.

To test the significance of the association measure, Light and Margolin (1971) developed the following C statistic

$$C = (n-1)(r-1)R^2 \cong \chi^2_{(r-1)(g-1)}$$
(9)

They showed that the C-statistic is asymptotically chi-squared distributed with (r-1)(g-1) degrees of freedom.

3.2. CATANOVA for multi-way contingency tables

Anderson (1980) extended the CATANOVA methodology of Light and Margolin (1971) to multidimensional contingency tables obtained from the cross-classification of the response variable Y with several factors. For this purpose, let$j$=1,2,…,J index the categories of Y, i_1=1,2,…I_1 index the categories of X_1, and i_2=1,2,…I_2 index the categories of X_2. Under the assumption that Y follows the product multinomial distribution with parameters $\prod_{i_1 i_2 j}$ and $n_{i_1 i_2 .}$ corresponding to the i_2 th level of X_2 within the i_1 th level of X_1, the null hypothesis of conditional independence of Y and X_2 given X_1 can be stated as

$$H_0 : \prod_{i_1 i_2 j} = \prod_{i_1 . j}, \quad i_2 = 1,2, \dots, I_2 \qquad (10)$$

or

H_0: The pair of variables Y and X_2 are conditionally independence of given X_1.

He proposed test statistic for the null hypothesis of conditional independence of Y and X_2 given X_1 in terms of the $R^2_{02\backslash 1}$ partial association criterion. $R^2_{02\backslash 1}$ has "proportion of explained variation" interpretations. This information is useful as a basis for selecting a subset of factors which are most important for explaining the variability of a response variable. The subscript 0 denotes the dependent variable Y and the subscripts 1 and 2 correspond to the factors X_1 and X_2, respectively. To explain the variability of a response variable, let us calculate the following association measure:

$$R^2_{02/1} = \frac{\sum_{i_1 i_2 j} n^2_{i_1 i_2 j}/n_{i_1 i_2 .} - \sum_{i_1 j} n^2_{i_1 . j}/\, n_{i_1 ..}}{n - \sum_{i_1 j} n^2_{i_1 . j}/\, n_{i_1 ..}} \qquad (11)$$

where $n_{i_1 i_2 j}$ denotes the number of subjects in the sample which are jointly classified as belonging to the i_1 th level of X_1, the i_2 th level of X_2, and the jth level of Y; $n_{i_1 i_2 .}$ denotes the marginal total number of subjects classified as belonging to the i_1 th level of X_1, the i_2 th level of X_2; $n_{i_1 . j}$ denotes the marginal total number of subjects classified as belonging to the i_1 th level of X_1, the j th level of Y; $n_{i_1 ..}$denotes the marginal total number of subjects classified as belonging to the i_1 th level of X_1.

He derived a test statistic for assessing the statistical significance of the measure of partial association proposed in (11). This test statistic is as follows:

$$C_{02\backslash 1} = (n - I_1)(J - 1)R^2_{02\backslash 1} \qquad (12)$$

Under H_0, $C_{02\backslash 1}$ asymptotically follows the chi-square distribution with $I_1(I_2 - 1)(J - 1)$ degrees of freedom (Anderson, 1980).

4. NUMERICAL EXAMPLE

In this section, two different contingency tables were considered, and the best model was sought by using the backward elimination technique proposed by Wermuth (1976b). The results obtained by using the chi-square (χ^2) statistic and the C statistic were considered when determining the best model.

4.1. Coppen data

Table 1 shows data for a set of four binary variables concerning symptoms of 362 psychiatric patients (Wermuth, 1976b). The symptoms are X_1 : stability (0 =extroverted, 1 = introverted); X_2 : validity (0 = psychasthenic, 1 = energetic); X_3 : acute depression (0 = yes, 1 = no); X_4 : solidity (0 = hysteric, 1 = rigid).

Table 1.Data on symptoms of psychiatric patients.

		X_4			
		0		1	
		X_3			
X_1	X_2	0	1	0	1
0	0	15	23	25	14
	1	9	14	46	47
1	0	30	22	22	8
	1	32	16	27	12

Table 1 shows the observed cell counts n_{ijkl} for each of the symptom combinations. Wermuth (1976b) analysed these data using backward elimination techique, and obtained the best model. This pattern was the result of the backward selection procedure displayed in Table 2. When four different variable pairs were considered in this method, it was possible for 6 variable pairs to assume zero partial association.

When Table 2 was reviewed, the χ^2 statistic was initially determined by using the method proposed by Wermuth (1976a, 1976b) for all variable pairs. The minimum insignificance test statistic was then determined by using the C statistic proposed by Anderson (1980) and Light and Margolin (1971). This provided the variable with zero partial association. For example, at the first step of the selection, variable pair (2,3) was selected to have z.p.a., the corresponding chi-square statistic was computed as

$$2\left[\left(\sum n_{ijkl} \ln n_{ijkl}\right) - \left(\left(\sum n_{ij.l} \ln n_{ij.l}\right) + \left(\sum n_{i.kl} \ln n_{i.kl}\right) - \left(\sum n_{i..l} \ln n_{i..l}\right)\right)\right] = 3.93$$

Denote the number of categories for each variable as I_1=2, I_2=2, I_3=2 and I_4=2, then the d.f. for z.p.a. of (2,3) given variables 1 and 4 are (I_2-1)(I_3-1) I_1 I_4=4.

Also, variable pair (2,3) was selected to have z.p.a.; based on the assumption that the X_2 variable is a response variable (Y), the corresponding R^2 and C statistic were computed as:

$$R^2_{YX_3.X_1X_4} = \frac{\sum_{i_1 j i_3 i_4} n^2_{i_1 j i_3 i_4}/n_{i_1 . i_3 i_4} - \sum_{i_1 j i_4} n^2_{i_1 j . i_4}/n_{i_1 .. i_4}}{\sum_{i_1 j i_4} n^2_{i_1 j . i_4}/n_{i_1 .. i_4}} = 0.009$$

and

C=(362-4)(0.009)=3.222

Denote the number of categories for each variable as I_1=2, I_2=2, I_3=2 and I_4=2, then the d.f. for z.p.a. of (2,3) given variables 1 and 4 are $I_1 I_4$(J-1)(I_3-1) = 4.

Also, variable pair (2,3) was selected to have z.p.a.; based on the assumption that the X_3 variable is a response variable (Y), the corresponding R^2 and C statistic were computed as:

$$R^2_{YX_2.X_1X_4} = \frac{\sum_{i_1 i_2 j i_4} n^2_{i_1 i_2 j i_4} / n_{i_1 i_2 . i_4} - \sum_{i_1 j i_4} n^2_{i_1 . j i_4} / n_{i_1 .. i_4}}{\sum_{i_1 j i_4} n^2_{i_1 . j i_4} / n_{i_1 .. i_4}} = 0.010$$

and

C=(362-4)(0.010)=3.580

The d.f. for z.p.a. of (2,3) given variables 1 and 4 are $I_1 I_4 (J\text{-}1)(I_2\text{-}1) = 4$.

C statistics (which are calculated in four steps) were obtained by using the method first proposed by Anderson, which makes use of conditional independence associations. Here, the aim depends on the selection of one of the variable pairs as the dependent variable. For example, the (1,2) variable having z.p.a. describes the conditional independence that is valid for the variables 1 and 2 when variables 3 and 4 are provided. In other words, when the coefficient of determination is calculated, the calculations are initially carried out by assuming that the variable pair 1 and the variable pair 2 were dependent variables. However, in the sub-table 13 obtained during the third step, the C statistic was calculated with the method proposed by Light and Margolin by separately considering each case in which 1 variable was dependent and that 3 variables were independent, or in which 1 variable was independent and 3 variables were dependent.

In the second step of the selection, the smallest insignificance test statistic was χ^2= 4.99 and C=4.680. This value indicated that the (3,4) variable pair had zero partial association. In the third step of the selection, the smallest insignificance test statistic was χ^2= 5.49 and C=5.400. This value indicated that the (1,2) variable pair had zero partial association. The method was discontinued during the fourth step, as all values for χ^2 were significant at a significance level of α=0.05.

When all results are considered, the model that best describes the obtained data is the 13 / 14 / 24 model. According to this model, it is clear that the (1,2), (2,3) and (3,4) variable pairs had zero partial associations. In addition, the marginal associations of the (1,3), (1,4) and (2,4) symptoms were sufficient for describing the associations between all four symptoms.

Table 2. Model search for the four symptoms using χ^2 and C statistics

	Step 1					Step 2					
Variable pair	χ^2 statistics for z.p.a.	d.f	Determination of coefficients (R^2)	C statistics	d.f	Sub-table	χ^2 statistics for z.p.a.	d.f	Determination of coefficients (R^2)	C statistics	d.f
(1,2)	4.78	4	$R^2_{YX_2.X_3X_4}$=0.012 $R^2_{YX_1.X_3X_4}$=0.013	4.296 4.654	4 4	124	5.49	2	$R^2_{YX_2.X_4}$=0.015 $R^2_{YX_1.X_4}$=0.015	5.400 5.400	2 2
(1,3)	12.87	4	$R^2_{YX_3.X_2X_4}$=0.035 $R^2_{YX_1.X_2X_4}$=0.036	12.530 12.888	4 4	134	13.58	2	$R^2_{YX_3.X_4}$=0.037 $R^2_{YX_1.X_4}$=0.037	13.320 13.320	2 2
(1,4)	33.00	4	$R^2_{YX_4.X_2X_3}$=0.079 $R^2_{YX_1.X_2X_3}$=0.089	28.282 31.862	4 4	-*	×	×	×	×	×
(2,3)	3.93	4	$R^2_{YX_3.X_1X_4}$=0.009 $R^2_{YX_2.X_1X_4}$=0.010	3.222 3.580	4 4	X*	×	×	×	×	×
(2,4)	22.38	4	$R^2_{YX_4.X_1X_3}$=0.048 $R^2_{YX_2.X_1X_3}$=0.061	17.184 21.838	4 4	124	19.73	2	$R^2_{YX_2.X_1}$=0.052 $R^2_{YX_4.X_1}$=0.054	18.720 19.440	2 2
(3,4)	7.64	4	$R^2_{YX_4.X_1X_2}$=0.021 $R^2_{YX_3.X_1X_2}$=0.023	7.518 8.234	4 4	134	4.99	2	$R^2_{YX_3.X_1}$=0.013 $R^2_{YX_4.X_1}$=0.014	4.680 5.040	2 2
Selected model			124 / 134						124 / 13		

Table 2. Continued

	Step 3					Step 4					
Sub-table	χ^2 statistics for z.p.a.	d.f	Determination of coefficients (R^2)	C statistics	d.f	Sub-table	χ^2 statistics for z.p.a.	d.f	Determination of coefficients (R^2)	C statistics	d.f
124	5.49	2	$R^2_{YX_2.X_4}$=0.015 $R^2_{YX_1.X_4}$=0.015	5.400 5.400	2 2	×	×	×	×	×	×
13	10.02	1	$R^2_{YX_1}$=0.028 $R^2_{YX_3}$=0.028	10.108 10.108	1 1	13	10.02	1	$R^2_{YX_1}$=0.028 $R^2_{YX_3}$=0.028	10.108 10.108	1 1
124	30.80	2	$R^2_{YX_1.X_2}$=0.082 $R^2_{YX_4.X_2}$=0.084	29.520 30.240	2 2	14	28.03	1	$R^2_{YX_1}$=0.078 $R^2_{YX_4}$=0.089	28.158 32.129	1 1
×	×	×	×	×		×	×	×	×	×	
124	19.73	2	$R^2_{YX_2.X_1}$=0.052 $R^2_{YX_4.X_1}$=0.054	18.720 19.440	2 2	24	16.97	1	$R^2_{YX_2}$=0.077 $R^2_{YX_4}$=0.071	27.797 25.631	1 1
×	×	×	×	×	×	×	×	×	×	×	×
Selected model	13 / 14 / 24					3 / 14 / 24					

*X means that the corresponding variable pair was in a previous step selected to have z.p.a.

- means that the corresponding pattern requires iterative fitting.

4.2. Data regarding individuals who underwent open heart surgery

Table 3 shows data from Erbaş and Bayrak (1999) for a set of four binary variables regarding 286 patients who underwent open heart surgery. The symptoms are X_1 :sex (1 =female, 2 = men); X_2 :duration of stay in intensive care (1 = ≤2 day, 2 = >2 day); X_3 :total duration of hospitalization (1 = ≤12 day, 2 = >12 day); X_4 :other disease (1 = absent, 2 = present).

Table 3.Data regarding 286 individuals who underwent open heart surgery

Sex															
female								male							
duration of stay in intensive care								duration of stay in intensive care							
≤2				>2				≤2				>2			
total duration of hospitalization				total duration of hospitalization				total duration of hospitalization				total duration of hospitalization			
≤12		>12		≤12		>12		≤12		>12		≤12		>12	
other disease		other disease		other disease		other disease		other disease		other disease		other disease		other disease	
no	yes	no	yes	no	yes	no	yes	No	Yes	no	yes	no	yes	no	yes
33	11	33	10	8	3	4	6	55	47	31	14	8	7	11	9

We analysed the these data using zero partial associations models, and obtained the best model. This pattern was the result of the backward selection procedure displayed in Table 4. For example, at the first step of the selection, variable pair (1,2) was selected to have z.p.a., the corresponding chi-square statistic was computed as

$$2\left[\left(\sum n_{ijkl} ln n_{ijkl}\right) - \left(\left(\sum n_{.jkl} ln n_{.jkl}\right) + \left(\sum n_{i.kl} ln n_{i.kl}\right) - \left(\sum n_{..kl} ln n_{..kl}\right)\right)\right]=3.85$$

Denote the number of categories for each variable as I_1=2, I_2=2, I_3=2 and I_4=2, then the d.f. for z.p.a. of (1,2) given variables 3 and 4 are (I_3-1)(I_4-1) I_1 I_2= 4.

Also, variable pair (1,2) was selected to have z.p.a.; based on the assumption that the X_1 variable is a response variable (Y), the corresponding R^2 and C statistic were computed as:

$$R^2_{YX_1.X_3X_4} = \frac{\sum_{i_1ji_3i_4} n^2_{i_1ji_3i_4} / n_{i_1.i_3i_4} - \sum_{i_1ji_4} n^2_{i_1j.i_4} / n_{i_1..i_4}}{\sum_{i_1ji_4} n^2_{i_1j.i_4} / n_{i_1..i_4}} = 0.013$$

and C=(286-4)(0.013)=3.666

Denote the number of categories for each variable as I_1=2, I_2=2, I_3=2 and I_4=2, then the d.f. for z.p.a. of (1,2) given variables 3 and 4 are $I_3I_4(I_1$-1) (J-1)=4.

Also, variable pair (1,2) was selected to have z.p.a.; based on the assumption that the X_2 variable is a response variable(Y), the corresponding R^2 and C statistic were computed as:

$$R^2_{YX_2.X_3X_4} = \frac{\sum_{ji_2i_3i_4} n^2_{ji_2i_3i_4} / n_{.i_2i_3i_4} - \sum_{ji_3i_4} n^2_{j.i_3i_4} / n_{..i_3i_4}}{\sum_{ji_3i_4} n^2_{j.i_3i_4} / n_{..i_3i_4}} = 0.014$$

and C=(286-4)(0.014)=3.948

The d.f. for z.p.a. of (1,2) given variables 3 and 4 are $I_3I_4(J$-1)(I_2-1)=4.

Table 4. Model search for the four symptoms using χ^2 and C statistics

	Step 1					Step 2						Step 3					
Variable pair	χ^2 statistics for z.p.a.	d.f	Determination of coefficients (R^2)	C statistics	d.f	Sub-table	χ^2 statistics for z.p.a.	d.f	Determination of coefficients (R^2)	C statistics	d.f	Sub-table	χ^2 statistics for z.p.a.	d.f	Determination of coefficients (R^2)	C statistics	d.f
(1,2)	3.85	4	$R^2_{YX_2.X_3X_4}$=0.014 $R^2_{YX_1.X_3X_4}$=0.013	3.948 3.666	4 4	X*	×	×	×	×	×	×	×	×	×	×	×
(1,3)	8.14	4	$R^2_{YX_3.X_2X_4}$=0.027 $R^2_{YX_1.X_2X_4}$=0.028	7.614 7.896	4 4	134	5.39	2	$R^2_{YX_1.X_4}$=0.019 $R^2_{YX_3.X_4}$=0.017	5.396 4.828	2 2	134	5.39		$R^2_{YX_1.X_4}$=0.018 $R^2_{YX_3.X_4}$=0.017	5.396 4.828	2 2
(1,4)	7.87	4	$R^2_{YX_4.X_2X_3}$=0.026 $R^2_{YX_1.X_2X_3}$=0.028	7.332 7.896	4 4	134	7.04	2	$R^2_{YX_1.X_3}$=0.024 $R^2_{YX_4.X_3}$=0.023	6.816 6.532	2 2	134	7.04		$R^2_{YX_1.X_3}$=0.024 $R^2_{YX_4.X_3}$=0.023	6.816 6.532	2 2
(2,3)	10.97	4	$R^2_{YX_3.X_1X_4}$=0.034 $R^2_{YX_2.X_1X_4}$=0.038	9.588 10.716	4 4	234	8.22	2	$R^2_{YX_3.X_4}$=0.033 $R^2_{YX_2.X_4}$=0.029	9.372 8.236	2 2	23	5.36		$R^2_{YX_3}$=0.019 $R^2_{YX_2}$=0.019	5.415 5.415	1 1
(2,4)	5.22	4	$R^2_{YX_4.X_1X_3}$=0.022 $R^2_{YX_2.X_1X_3}$=0.019	6.204 5.358	4 4	234	4.39	2	$R^2_{YX_2.X_3}$=0.015 $R^2_{YX_4.X_3}$=0.016	4.260 4.544	2 2	×	×	×	×	×	×
(3,4)	5.29	4	$R^2_{YX_4.X_1X_2}$=0.018 $R^2_{YX_3.X_1X_2}$=0.019	5.076 5.358	4 4	-*	×	×	×	×	×	134	2.66		$R^2_{YX_3.X_1}$=0.009 $R^2_{YX_4.X_1}$=0.010	2.556 2.840	2 2
Selected model	134 / 234					134 / 23						13 / 14 / 23					

Table 4. Continued

Step 4						Step 5					
Sub-table	χ^2 statistics for z.p.a.	d.f	Determination of coefficients (R^2)	C statistics	d.f	Sub-table	χ^2 statistics for z.p.a.	d.f.	Determination of coefficients (R^2)	C statistics	d.f
×	×	×	×	×	×	×	×	×	×	×	×
13	3.57	1	$R^2_{YX_3}$=0.013 $R^2_{YX_1}$=0.012	3.705 3.420	1 1	×	×	×	×	×	×
14	5.22	1	$R^2_{YX_1}$=0.018 $R^2_{YX_4}$=0.018	5.130 5.130	1 1	14	5.22	1	$R^2_{YX_1}$=0.018 $R^2_{YX_4}$=0.018	5.130 5.130	1 1
23	5.36	1	$R^2_{YX_3}$=0.019 $R^2_{YX_2}$=0.019	5.415 5.415	1 1	23	5.36	1	$R^2_{YX_3}$=0.019 $R^2_{YX_2}$=0.019	5.415 5.415	1 1
×	×	×	×	×	×	×	×	×	×	×	×
×	×	×	×	×	×	×	×	×	×	×	×
Selected model	14 / 23					1 / 23 / 4					

*X means that the corresponding variable pair was in aprevious step selected to have z.p.a.

- means that the corresponding pattern requires iterative fitting

When this table was reviewed, the minimum insignificance test statistics were calculated by estimating the χ^2 statistic and C statistic for all variable pairs in each step. The method was discontinued during the fourth step, as all values for χ^2 were significant at a significance level of α=0.05. Consequently, the model that best describes the obtained data is the 1 / 23 / 4 model. According to this model, it is clear that the (1,2), (1,3), (1,4), (2,4) and (3,4) variable pairs had zero partial associations. Marginal associations (variable pairs 2 and 3) were only obtained for the duration of stay in intensive care and the total duration of hospitalization. Thus, for individuals undergoing open heart surgery, the gender of the patient did not have an effect on the presence of other diseases. The gender of the patient and the presence of another disease were not affected by the duration of stay in intensive care and the total duration of hospitalization.

5. CONCLUSIONS

The purpose of analyzing data is to find structures which are complex enough to fit the data but simple enough to facilitate interpretation. Structures describing interrelations among several variables may be called pattern of association. Wermuth (1976a) characterized a certain class of patterns by the concept of zero partial association and showed that it is this class of patterns which can be studied by fitting multiplicative models to contingency table. A feature of Wermuth's method is that a pair of factors contained in more than one term in the model cannot be considered for zero partial association or conditional independence. For determining the model that best fits the data by using the backward elimination proposed by Wermuth, multiplicative models are sufficient for describing marginal associations and other association models. They also present the means for easily interpreting these associations.

In this study, the CATANOVA test was used to evaluate the associations in two-way contingency tables, and also to evaluate the measure of partial associations in multi-way contingency tables. The objective was to obtain certain results by using the C statistic instead of the chi-square statistic and by benefiting from zero partial associations. When using the C statistic, it was assumed that one of the variables being considered was a response variable.

When these results were evaluated, it was observed that, in the selection of a model, the values of the C statistic were generally smaller than the values of the chi-square statistic. By using both test statistics, it was observed that similar results were obtained in the interpretation of marginal relations between variable pairs, in the test statistics of each model, and in the interpretation of the models. To this end, it was demonstrated that in case one of the variables for model selection was the response variable, the C statistics could be preferred over the χ^2 statistic, so long as the expected frequencies were not required to be too small and the test frequency was found to be smaller than chi-square statistics. This result is in agreement with previously conducted studies in the literature.

CONFLICT OF INTEREST

No conflict of interest was declared by the authors.

REFERENCES

Anderson, R.J. and Landis, J.R.,"Catanova for multidimensional contingency tables: Nominal-scale response", ***Communications in Statistics-Theory and Methods***, 9(11), 1191-1206(1980),

Camminatiello, I. and D'ambra, L., "Visualization of the significant explicative categories using CATANOVA method and non-symmetrical correspondence analysis for evaluation of passenger satisfaction", ***Journal of Applied Quantitative Methods***, 5(1):64-72, (2010).

Christensen, R., Log-linear models and logistic regression. Second edition, ***Springer-Verlag*** New York (1990).

Darroch, J.N., Lauritzen,S.L. and Speed,T.P., "Markov fields and log-linear interactions models for contingency tables", ***Annals Statistics***, 8**:**522-539, (1980).

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***, 34:1755-1769(2005).

Edwards,D., and Kreiner, S., "The analysis of contingency tables by graphical models", ***Biometrika***, 70: 553-562(1983).

Erbaş,E.O. and Bayrak,H., Graphical Models. Bizim Publications Office, ISBN: 975-97011-0-3, Ankara / Turkey (1999).

Light, R. and Margolin,B., "An analysis of variance for categorical data", ***Journal of the American Statistical Association***, 66 :534-544 (1971).

Margolin, B.H. and Light,R.J., "An analysis of variance for categorical data II. Small samples comparisons with Chi-square and other competitors", ***Journal American Statistics Association***, 69 : 755-761(1974).

Lombardo, R. and Camminatiello, I., "CATANOVA for two-way cross classified categorical data", ***Statistics***. 44(1): 57-71 (2010).

Singh,B.,"On the analysis of variance method for nominal data", Sankhya: ***The Indian Journal of Statistics***, 55 (B):40-47, (1993).

Singh, B., "On CATANOVA method for analysis of two-way classified nominal data", Sankhyā: ***The Indian Journal of Statistics***, 58(3): 379-388, (1996).

Wermuth, N., "Analogies between multiplicative models in contingency tables and covariance selection", ***Biometrics***, 32**:** 95-108, (1976a).

Wermuth, N., "Model search among multiplicative models", ***Biometrics,*** 32**:**256-263,(1976b).