AN EXPLORATORY METHOD FOR THE SELECTION OF CATEGORICAL VARIABLES

J. Norberto W. Dachs
and
Mauro S. F. Marques

RELATÓRIO INTERNO Nº 182

ABSTRACT: A method is presented for the selection of categorical variables, using the measure of association λ of Guttman (1941). The method is exploratory and does not assume any previous choice of model or class of models.

Universidade Estadual de Campinas
Instituto de Matemática, Estatística e Ciência da Computação
IMECC – UNICAMP
Caixa Postal 1170
13.100 - Campinas - SP
BRASIL

Outubro - 1980

1. INTRODUCTION

The problem of building a forecasting model which relates p "independent variables" to a "dependent" one, and the associate problem of selecting those p variables from a set of q possible candidates (q >> p, in may instances) has been fully discussed for the case when the q+1 variables are quantitative ones. Several selection procedures are available: Anscombe (1967), Tukey (1967), Allen (1971), Mallows (1973), Furnival and Wilson (1974), among many others, including the well known stepwise techniques.

When the variables are categorical the proposals are few: Grizzle, Starmer and Koch (1969), Goodman (1971 and 1973), Wermuth (1976). They suppose a model or a class of models and then discuss the problem of selection of variables in this framework. In many applications this approach may be unfeasible or, at best, risky. In pilot studies, many times, the researcher does not have enough knowledge of the problem to start from a model, even a saturated one, and then select variables.

The technique proposed here is an exploratory one, to select among a large number of variables a much smaller set, without any previous postulation or choice of model, or class of models. It is based on the use of the measure of association λ, proposed by Guttman (1941) and whose properties and advantages are presented by Goodman and Kruskal (1954, 1959, 1963 and 1972).

2. THE MEASURE OF ASSOCIATION λ

Consider two polytomies L and C with classes, respectively, $(L_1, L_2, \ldots, L_r)$ and $(C_1, C_2, \ldots, C_s)$. Let ρ_{ij} be the probability of an individual belonging to the cell (L_i, C_j), so that $\rho_{i.} = \sum_j \rho_{ij}$ is the probability of belonging to L_i and $\rho_{.j} = \sum_i \rho_{ij}$ of belonging to C_j. Let $\rho_{(m).} = \max_i \rho_{i.}$ and $\rho_{(m)j} = \max_i \rho_{ij}$ and suppose that $\rho_{(m).}$ and the $\rho_{(m)j}$'s are unique. Without any condition of underlying

continuity and order for the polytomies suppose that C precedes L in some sense, causaly or not. Consider then the following probabilistic model: one individual is selected at random from the population and his L class has to be predicted when: 1 - no information is given on the C - class to which the subject belongs; 2- given the C-class. In the first case the best prediction is the class L_i such that $\rho_{i.} = \rho_{(m).}$ and in the second case the best prediction, if the individual belongs to class C_j is to chose the L-class such that $\rho_{ij} = \rho_{(m)j}$. Guttman (1941) proposes the following measure of association

$$\lambda(L/C) = \frac{(\text{Probability of error in the first case}) - (\text{Probability of error in the second case})}{(\text{Probability of error in the first case})} =$$

$$= \frac{\sum_j \rho_{(m)j} - \rho_{(m).}}{1 - \rho_{(m).}} \qquad (2.1)$$

It is easy to see that the assumptions of uniqueness of the maxima are not necessary in practice since the existence of more than one maximum in either case does not affect the probability of error for the choice of L-class that has been made. This measure of association has the following properties:

P1 - $\lambda(L/C)$ becomes undetermined if and only if the population in concentrated (with probability one) in one line (one L-class).

P2 - If $\lambda(L/C)$ is not undetermined then $0 \leq \lambda(L/C) \leq 1$.

P3 - $\lambda(L/C) = 0$ if and only if knowledge of C does not help in the prediction of L, i.e.
$\lambda(L/C) = 0 \iff \exists k;\ \rho_{kj} = \rho_{(m)j},\ \forall j.$

P4 - $\lambda(L/C) = 1$ if and only if knowledge of C completely specifies the L-class, i.e.
$\lambda(L/C) = 1 \iff \exists k = k(j);\ \rho_{ij} = 0,\ i \neq k.$

P5 - If L and C are statistically independent then $\lambda(L/C) = 0$.

P6 - $\lambda(L/C)$ in invariant under permutations in the lines or columns (L and C classes, respectively).

The choice of this measure of association to develop the selection technique was based on the facts that it can be used without any strong assumptions on the polytomies, it has a strong appeal on prediction abilities and it has a very simple interpretation.

3. THE ASYMPTOTIC DISTRIBUTION OF λ

If P_{ij} denotes the relative frequency of individuals in cell (L_i, C_j) then, with notation analogous to (2.1),

$$\hat{\lambda}(L/C) = \frac{\sum_j P_{(m)j} - P_{(m).}}{1 - P_{(m).}} \qquad (3.1)$$

is an estimator of $\lambda(L/C)$. The theory involved in studying the exact distribution of $\hat{\lambda}$ is extremely complicated as can be seen in Greenwood and Glasgow (1950) and Koselka (1952 and 1956) for the determination of the exact distribution of the maximum frequency observed in multinomial sampling. On the other hand the asymptotic distribution as an approximation is fairly good as seen in Goodman as Kruskal (1963). This justifies the use of the asymptotic results for testing and constructing confidence intervals for $\hat{\lambda}$ needed ahead.

The fundamental theorem states that under multinomial sampling for the whole population, if

S1 - For each j, the index i such that $\rho_{(m)j} = \rho_{ij}$ is unique;

S2 - The index i such that $\rho_{(m).} = \rho_{i.}$ is unique;

S3 - $\rho_{(m).} \neq 1$;

then, for $0 < \lambda(L/C) < 1$:

$$\sqrt{n}\,(\hat{\lambda}(L/C) - \lambda(L/C))\,\sqrt{V} \tag{3.2}$$

where

$$V = \frac{(1 - P_{(m).})^3}{(1 - \sum_j P_{(m)j})(\sum_j P_{(m)j} + P_{(m).} - 2\,\Sigma^* P_{(m)j})} \tag{3.3}$$

is asymptotically normal with mean zero and variance one. If i_j is the index corresponding to each j in condition S1 above and i* the one in S3, the symbol Σ^* is the summation $\sum_j P_{i_j j}$ over all j such that $i_j = i^*$.

If $\lambda(L/C) = 0$ the probability that $\hat{\lambda}$ is zero converges to one. If $\lambda(L/C) = 1$ then $\hat{\lambda}$ is one without any asymptotic approximations. Hypothesis of the type $\lambda = 0$ or $\lambda = 1$, at any level of significance are rejected if $\hat{\lambda} \neq 0$ or $\hat{\lambda} \neq 1$, respectively. Confidence intervals must exclude 0 and 1 unless $\hat{\lambda} = 0$ or 1, respectively, in which case the "intervals" consist of the corresponding points 0 or 1 only.

In applications a serious problem may arise in the computation of the asymptotic variance when there are ties in the $P_{i.}$'s. As an example consider the situation in table 1. As pointed out before $\hat{\lambda}(L/C)$ is well determined and equal to (33-16)/(50-16) = 0.5.

	C_1	C_2	C_3	
L_1	9	0	2	11
L_2	2	10*	4	16
L_3	1	1	14*	16
L_4	1	0	6	7
	13	11	26	50

Table 1: Example with ties in the $P_{i.}$'s.

But, since $P_{(m).} = P_{2.}$ or $P_{3.}$ there are two ways of computing the asymptotic variance V of (3.3). This is a consequence of the fact that there are two different values for $\Sigma^* P_{(m)j}$, namely 10/50 and 14/50. Consequently there are two possible approximate confidence intervals for $\lambda(L/C)$. In this example the two approximate intervals of 95% confidence coefficient are, resctively, (.2805, .7195) and (.3132, .6868). This problem does not have a theoretical solution, but, for practical purposes the mean of the possible values of the variance can be used, or possibly better , the maximum, since this procedure would assure a lower limit for the confidence coefficient.

Another problem would be $P_{(m).} = 1$. This means that it is not possible to estimate $\lambda(L/C)$. The problem is circumvented by the fact that $P_{(m).} = 1$ is a strong indication of a bad choice for the L-classes, and a redefinition of these is necessary since all individuals have fallen in the same class.

4. THE METHOD OF SELECTION

Consider a sample of the polytomies $L, C^{(1)}, C^{(2)}, \ldots, C^{(k)}$. The first step consists of selecting a C-polytomy with the greatest possible association with L. For this purpose all the k cross classification $L \times C^{(\ell)}, \ell = 1,2,\ldots,k$ are considered and the $\lambda(L/C)^{(\ell)}$) are computed. It is then choosen the $C^{(\ell_1)}$ with greatest $\hat{\lambda}$. Having selected $C^{(\ell_1)}$, k-1 new polytomies are constructed, denoted by $C^{(\ell_1\ell)}$, $\ell = 1,2,\ldots,k$, $\ell \neq \ell_1$. Each of these is obtained by considering possible combinations of $C^{(\ell_1)}$-classes with all the classes of each $C^{(\ell)}$, $\ell \neq \ell_1$.

The selection then proceeds in choosing the $C^{(\ell_1\ell)}$ such that $\lambda(L/C^{(\ell_1\ell)})$ is maximized. If $C^{(\ell_1\ell_2)}$ is this polytomy the method consists of then forming new polytomies $C^{(\ell_1\ell_2\ell)}$, $\ell = 1,2,\ldots,k$, $\ell \neq \ell_1$ and $\ell \neq \ell_2$ by considering all possible combinations of $C^{(\ell_1\ell_2)}$-classes with all the classes of each $C^{(\ell)}$, $\ell \neq \ell_1, \ell_1 \neq \ell_2$, and so successively. The procedure is best ilustrated by the

following example, with polytomies L, $C^{(1)}$ and $C^{(2)}$. In table 2a there is a sample of 20 individuals, classified according to these three polytomies. In table 2b and 2c are the cross-tabulations of L X $C^{(1)}$ and L X $C^{(2)}$, respectively. In this case

$$\hat{\lambda}(L/C^{(1)}) = \frac{(3+3+1+3) - 9}{20 - 9} = 1/11$$

and

$$\hat{\lambda}(L/C^{(2)}) = \frac{(1+3+3+2+2) - 9}{20 - 9} = 2/11$$

The first selection is $C^{(2)}$. A new polytomy is then formed combining each class of $C^{(2)}$ with every $C^{(1)}$-class. The classes of the new polytomy $C^{(21)}$ are denoted by 11, 21, 31,...,55, as in table 2d, in which is presented the crossing of L X $C^{(21)}$. Then

$$\hat{\lambda}(L/C^{(21)}) = \frac{(1+1+1+1+2+2+1+1+1+1+2) - 9}{20 - 9} = 5/11$$

It is very important in simplifying the method computationaly that the classes of any aggregation can be renumbered and the number of classes can be reduced by ignoring all those that have no elements, since they do not provide any information on the joint behavior of L and any further aggregation since they are already empty. In the example of table 2 the number of classes of $C^{(21)}$ can be reduced from 20 to 11, ignoring the empty ones, marked with * in table 2d.

In using the method the fundamental result is that λ is non-decreasing in each step. To show this consider polytomies A, B, C and (BC), this last one resulting from the aggregation of B and C.

C	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
L	2	2	1	2	3	1	3	2	2	3	2	2	1	3	3	1	2	1	2	3
$C^{(1)}$	2	3	1	1	1	1	4	2	4	1	2	4	2	4	4	1	1	3	1	3
$C^{(2)}$	2	4	3	5	2	4	2	4	2	1	2	4	2	4	4	1	1	3	5	3

(a)

$L/C^{(1)}$	1	2	3	4	
1	3	1	1	0	5
2	3	3	1	2	9
3	2	0	1	3	6
	8	4	3	5	20

(b)

$L/C^{(2)}$	1	2	3	4	5	
1	1	1	2	1	0	5
2	1	3	0	3	2	9
3	1	2	1	2	0	6
	3	6	3	6	2	20

(c)

$L/C^{(21)}$	11	21	31	41	51	12	22	32	42	52	13	23	33	43	53	14	24	34	44	54	
1	1	0	1	1	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	5
2	1	0	0	0	2	0	2	0	1	0	0	0	0	1	0	0	1	0	1	0	9
3	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	2	0	6
	3	1	1	1	2	0 *	3	0 *	1	0 *	0 *	0 *	2	1	0 *	0 *	2	0 *	3	0 *	20

(d)

Table 2 - An example of selection

Let

ρ_{ij} = probability of cell (A_i, B_j)

ρ_{ijk} = probability of cell (A_i, B_j, C_k)

$$\lambda(A/B) = \frac{\sum_j \rho_{(m)j} - \rho_{(m).}}{1 - \rho_{(m).}} \tag{4.1}$$

$$\lambda(A/BC) = \frac{\sum_j \sum_k \rho_{(m)jk} - \rho_{(m)..}}{1 - \rho_{(m)..}} \tag{4.2}$$

and then

$$\rho_{(m).} = \rho_{(m)..} \tag{4.3}$$

since $\rho_{ij} = \sum_k \rho_{ijk}$, so that $\rho_{(m).} = \max_i \rho_{i.} = \max_i \sum_j \rho_{ij} =$

$$= \max_i \sum_j \sum_k \rho_{ijk} = \max_i \rho_{i..} = \rho_{(m)..}$$

On the other hand

$$\rho_{ijk} \leq \max_i \rho_{ijk} = \rho_{(m)jk} \qquad \forall i,j,k$$

$$\Longrightarrow \rho_{ij} = \sum_k \rho_{ijk} \leq \sum_k \rho_{(m)jk} \qquad \forall i,j$$

$$\Longrightarrow \sum_j \rho_{(m)j} \leq \sum_j \sum_k \rho_{(m)jk} \tag{4.4}$$

Therefore, using (4.3) and (4.4) in the expressions (4.1) and (4.2)

$$\lambda(A/B) < \lambda(A/BC) \tag{4.5}$$

The same result is also valid for the estimators. In a given step B corresponds to the joint polytomy up to that point and C is the new one being aggregated now.

5. HOW TO STOP

Sttoping the process is accomplished in different ways. Initially a desired level of association, $\lambda_o < 1$ can be established. If this association is statistically attained at level α the process of selection stops. It may also happen that all polytomies are included before this level of association is reached. To check whether the association desired has been obtained at level α, at each step the unilateral confidence interval, with confidence coefficient $(1-\alpha)$ is determined, using the asymptotic variance (3.3) and the appropriate normal cut-off point, this being equivalent to testing the hypothesis $\lambda \geq \lambda_o$ against $\lambda < \lambda_o$ at level α, stopping when the hypothesis is accepted.

Also, if aggregation of a new polytomy does not increase λ the process should stop. The third manner in which the process is terminated is when $\hat{\lambda}(L/C) = 1$ is attained at a certain step.

Although the occurrence of an indeterminacy in the computation of $\hat{\lambda}$, or $\hat{\lambda} = 0$ are not considered as stopping criteria, the process must be interrupted. Undetermination of $\hat{\lambda}$, as discussed before, corresponds to a bad choice of division in classes for the polytomy L. If $\hat{\lambda} = 0$, all the maxima in the columns occur at the same L-class. This corresponds to a possible bad choice of variables, and new ones may have to be included in the study if any predictions are to be made on L based on the C's.

Finally, if in a given step two or more polytomies give the same $\hat{\lambda}$ a personal (arbitrary) choice has to be made on which one to include first.

6. FINAL REMARKS

The method was tested on data of Sinisgalli (1980), for 132

children who started the first grade in 1977 in a small town in the state of São Paulo. The "dependent" variable (L) was the child's classification in a test to measure the ability to learn how to read, with scores: high, medium high, medium low and low. Initially ten other variables were considered:

X_1 Time in pre-school - 6 classes

X_2 Place to study in the house - 2 classes

X_3 Type of sport - 4 classes

X_4 Sleeping accomodations - 3 classes

X_5 Type of food at meals - 5 classes

X_6 Travel - 3 classes

X_7 Overprotection - 3 classes

X_8 Location of the house - 3 classes

X_9 Discussions between parents - 3 classes

X_{10} Parents separated - 2 classes

The limit for the desired association was set at $\lambda_o = 0.9$ and $\alpha = 0.05$. The method selected four variables, in the order given in table 3, where limsup is the limit of the corresponding confidence interval.

Step	Variable	$\hat{\lambda}$	limsup
1	X_3	0.143	0.270
2	X_4	0.312	0.553
3	X_5	0.545	0.756
4	X_7	0.779	0.934

Table 3 - The four variables selected by the method.

From the value of $\hat{\lambda}$ for X_3 which is the maximum of $\hat{\lambda}(L/X_i)$, $1 \leq i \leq 10$, it can be noticed that each of the variables had a low degree of association with the "dependent" variable. One interesting aspect to be pointed out is that X_3 is possibly, among the ten variables, the one that best measures, in this case, the socio-economic status of the child's family, in a small poor community in an underdeveloped area. The 4 classes for this variable are: swimming, bicycle riding but not swimming; ball-games but not swimming nor bicycle riding; none of the former.

An algorithm implementing the method is presented in Dachs, Marques and Tardelli (1980).

GOODMAN, L.A. and KRUSKAL, W. H. (1972). Measures of Association for Cross Classification IV. Simplification of Asymptotic Variances. J. Amer. Statist. Ass., 67, 415 - 421.

GREENWOOD, R.E. and GLASGOW, M.O. (1950). Distribution of Maximum and Minimum Frequencies in a Sample Drawn from a Multinomial Distribution. Ann. Math. Stat., 21, 416 - 424.

GRIZZLE, J.E.; STARMER, C.F. and KOCH, G. G. (1969). Analysis of Categorial Data by Linear Models. Biometrics, 25, 489 - 504.

GUTTMAN, L. (1941). An Ouline of the Statistics Theory of Prediction. Suplementary Study B - 1. Social Science Research, New York, 48, 253 - 318.

KOZELKA, R.M. (1952). On Some Special Order Statistics from the Multinomial Distribution. Ph.D. Dissertation, Harward Univ.

KOZELKA, R.M. (1956). Approximate Upper Percentage Points for Extreme Values in Multinomial Sampling. Ann. Mat. Stat., 27, 507 - 512.

MALLOWS, C.L. (1973). Some Comments on Cp. Technometris, 15, 661-675.

SINISGALLI, F.J. (1980). Maturidade Infantil para a Aprendizagem de Leitura e da Escrita. Master's Dissertation, Univ. Metodista de Piracicaba, Brasil.

TUKEY, J.W. (1967). Discussion of Anscombe's Paper. J.R. Statist. Soc. B, 29, 47 - 48.

WERMUTH, N. (1976). Model Search Among Multiplicative Models. Biometrics, 32, 253 - 263.

REFERENCES

ALLEN, D. (1971). Mean Square Error of Prediction as a Criterion for Selecting Variables. Technometrics, 13, 469-475.

ANSCOMBE, F.J. (1967). Topics in the Investigation of Least Squares, J.R. Statist. Soc. B, 29, 1-29.

DACHS, J.N.W.; MARQUES, M.S.F. and TARDELLI, A.O. (1980). An Algorithm for Selecting Categorial Variables. To be submitted to Applied Statistics.

FURNIVAL, G.M. and WILSON, R.W. (1974). Regression by Leaps and Bounds. Technometrics, 16, 499-511.

GOODMAN, L.A. (1971). The Analysis of Multidimensional Contingency Tables. Stepwise Procedures and Direct Estimation Methods for Building Models for Multiple Classifications. Technometrics, 13, 33-61.

GOODMAN, L.A. (1973). Guided and Unguided Methods for the Selection of Models for a set of T Multidimensional Contingency Tables. J. Amer. Statist. Ass., 68, 165-175.

GOODMAN, L.A. and KRUSKAL, W.H. (1954). Measures of Association for Cross Classification. J.Amer. Statist. Ass., 49, 732-764.

GOODMAN, L.A. and KRUSKAL, W.H. (1959). Measures of Association for Cross Classification II. Further Discussion and References. J. Amer. Statist. Ass., 54, 123-163.

GOODMAN, L.A. and KRUSKAL, W.H. (1963). Measures of Association for Cross Classification III. Approximate Sampling Theory. J. Amer. Statist. Ass., 58, 310-364.