

Research of optimal hierarchy of ordinal explanatory variables to explain an ordinal response variable¹

Christian Derquenne²

Electricité de France Research & Development, e-mail: christian.derquenne@edf.fr

Abstract: In many applications, in marketing for instance, the modelling of a response variable has not only the goal to know the main explanatory variables, but also the most contributive and the top that the importance order. We propose a method to put in order explanatory of ordinal variables to modelize a response ordinal variable. The main problem is the number of explanatory variables. A first answer is provided by stepwise approach to select variables, but it does not offer a solution concerning order of explanatory variables. Indeed, it must analyse all permutations of these last ones. The proposed method uses order statistic, ordinal logit model and simulated annealing algorithm.

Keywords: semi-parametric models, ordinal logit model, simulated annealing, marketing

1. Framework, motivation and objective

In framework of consumer's behaviour, and more precisely in survey satisfaction, several studies are made. Particular, it can be interesting for a marketing expert to know the main factors and their associated contributions which influence the market. Indeed, a strategy can be built linked with a hierarchy of these factors. A reasonable solution is to search the most contributive explanatory variables which play a role on overall factor (response variable) and to build associated standardized weights. For instance, if the response variable is a global satisfaction, the explanatory variables will be partial factors. In this paper, we focus on ordinal variables. The main problem to build a hierarchy of factors is the number of partial variables. Indeed, if the stepwise regression offers a first answer to select the most contributive variables, it does not give an optimal permutation of these variables. We propose, in this paper, a method which provides an optimal hierarchy of ordinal variables to fit a response variables. This method is based on the order statistic, the ordinal logit model (or gaussian linear model) and the simulated annealing algorithm.

2. Proposed method

Let Y be the global factor (response variable) distributed on ordinal scale with R categories and let $X_1, \dots, X_p$, p partial factors (explanatory variables) on ordinal scale with $m_1, \dots, m_p$ categories.

One solution (bivariate) is firstly to search a correlation between each X_j and Y , with the Kendall's τ_b or the Stuart's τ_c (1990). To evaluate the correlation level, it is possible to use the ratio between correlation and its asymptotic standard error. The more absolute

²Address of correspondence: 1, avenue du Général de Gaulle - 92141 Clamart Cedex - France

value of ratio is big, the more contribution of X_j to Y is high. In Some cases, hierarchy can appear obvious, but it is not all the time true. In addition, individual contribution of each explanatory variable is not taken into account. This problem increases, if the number of variables is high. To get round this problem, it is possible to use stepwise regression. A restricted number of explanatory variables can be obtained with different criteria, as the marginal likelihood ratio-test in ordinal logit model, for instance. Then a hierarchy can be obtained with the retained variables, putting in increase order associated p -values. However, this solution is not fully satisfying if the expert wishes to keep the all variables to measure its mutual contribution. Then, it can be interested to build statistical way to "estimate" the importance order of all the variables. A solution is to calculate a statistic to each permutation of the p variables allowing to evaluate order contribution of these variables with respect to Y . Then a new hierarchy of these variables can be provided.

Proposed approach: For one permutation of the p variables, that is, for each variable, the ordinal categories are put in increasing ou decreasing order. Rank 1 is given to the first chain, then if current chain is not equal to previous, then current rank is equal to 2, and so on. For the n observations, we obtain one variable of ranks associated to one permutation of p explanatory variables. This approach is based on the four following properties.

Prop 1: Minimum number of different ranks is equal to the number of categories of variable X_j which has the greatest number of categories, whereas the maximum number of ranks is equal to the product of $m_1, \dots, m_p$ categories, independently of Y .

Prop 2: Number of different ranks is the same whatever permutation independently of Y .

Remark 1: These two properties allow to work on the same reference support in terms of distribution of the rank variable, whatever the permutation. This offers a means to compare the different statistic values obtained on each permutation.

Prop 3: For each category of Y , the distribution of ranks is globally different for each permutation.

Remark 2: This property allows to propose the following rule: for each ordered category of Y , the less rank distributions are scattered and the more centers are different, then the more hierarchy of X 's is good in term of correlation with Y . The Gaussian Linear Model of Analysis of Variance with one effect given by Y , whereas rank variable is the response variable, seems to be the more suited to answer to previous rule. The Fisher statistics is used to calculate the link between Y and rank variable.

Prop 4: For each rank, the distribution of Y is globally different for each permutation.

Remark 3: This property allows to propose the following rule: the less distribution of Y by ordered rank is scattered and the more cumulated distribution functions are different, then the more hierarchy of X 's is good in term of correlation with Y . The Ordinal Logit Model is preferred to answer to previous rule. In this model, the response variable corresponds to Y and the rank variable is related to explanatory variable. The estimated likelihood ratio is used to measure the link between Y and rank variable.

These two statistics seem to be suited to do appear the quality of hierarchy. But, the curse of dimension linked to the number of hierarchies is very present. Indeed, this number is not only composed by the $p!$ permutations of p variables, but also increasing or decreasing order of each ordinal variable X_j with 2^p possibilities. However, this order is symmetric. Indeed, if the p variables are put in increasing order, then results will be the same to a decreasing order. Finally, the total number of hierarchies is $2^{p-1}p!$.

There are a lot of solutions in scientific literature. The first applications are related to optimal way of commercial traveller. Indeed, if he must visit each town, what is the shortest way ? There are $n!$ ways and criteria to optimize (minimization, in this case) is the covered distance. In our case, the Fisher's F or likelihood ratio are the criteria to optimize. These two statistics are to maximize. Then the main solution chosen in literature is stochastic way, that is to calculate a probability to continue or to replace proposed solution. We have chosen a classical method : the Simulated Annealing introduced by Kirkpatrick (1983).

3. Application to survey satisfaction

A survey satisfaction about services and products has been conducted on 923 persons. We have selected a part of questionnaire related to one product A. This part contains a question related to global satisfaction and 16 questions about partial factors of satisfaction. Y is the response variable and corresponds to the global satisfaction, and $X_1, \dots, X_{16}$ are related to the explanatory variables. In our example, the number of hierarchies is very high and is equal to $2^{15} 16! = 6.86E + 17$. So, it is not reasonable to calculate all permutations.

We have applied the both proposed methods of permutations with Fisher statistic, named "perm-anova" and with likelihood ratio, named "perm-logit". Then we have compared to the other following methods.

(1) The Ordinal Logit Model is applied on Y as response variable and the p partial factors as explanatory variables.

(2) The Sequential Regression developed by Bachelet (1996). This method consists in realizing a backward stepwise regression on all p variables (the non significant variables are also retained). Then a hierarchy is obtained and standardized weights are calculated for each variable, with respect to the more significant variable.

(3) The Bayesian Network (Pearl, 1965) is an oriented graph without circuit, associated to a probability distribution in which the nodes represent probabilised variables and the links correspond to the relations of probabilistic dependencies. In our case, Y is the target node and the X 's are the child nodes.

(4) The Partial Maximum Likelihood Regression, developed by Derquenne (2004) for categorical variables, is an extension of the Partial Least Squares Regression (Wold et al., 1983). This type of regression allows to avoid multicollinearity between explanatory variables. In our case, this property is very interesting, because the partial satisfactions are generally linked. The results provide a standardized weight (contribution) for each explanatory variable to Y .

(5) Clustering variables and PML Approach. In our application on satisfaction survey, as we have specified the partial factors are generally correlated and even can form a single dimension. However, if there is a dimensional structure, it is very interesting to apply a clustering of variables (Derquenne, 1998) to obtain some new dimensions containing the initial variables. Then, on the structural clusters obtained, we have applied the PML Approach (Derquenne, 2005) dedicated to categorical variables and which is an extension of the PLS Approach (Wold, 1982). In this case, the p partial satisfaction factors and the global satisfaction are the manifest variables and the links with own latent variables (new dimensions coming from clustering variables) correspond to the outer model. The inner model put in linking each latent variable with the global satisfaction.

(6) Clustering variables and RFGPC Approach (Regression on the First Generalized Principal Components). This process is the same as (5), but we use the RFGPC (Derquenne, 2005) which is an extension to the RFPC Approach (Derquenne et al., 2002). This method provides unbiased tests for the structural model estimation.

Main results: The Perm-anova, Perm-logit and PML Regression provide the same hierarchy on the three first variables : X_5 , X_{14} , X_{12} . The logit ordinal model gives nearly same order, putting X_{12} in fourth position. The hierarchy of Perm-logit is relatively close to the hierarchy provided by logit ordinal model. One reason is that it is the same criteria (likelihood ratio) to optimize. Even if the rank variable is a single explanatory variable in Perm-logit and the X 's are explanatory variables in ordinal logit model.

4. Contributions - limits - applications - extensions

The proposed methods provide a hierarchy of ordinal explanatory variables to fit an ordinal response variable. These ones are interesting because they are based on a semi-parametric approach. Indeed, using of ranks gives non-parametric side, whereas Ordinal Logit Model provides parametric side. In addition, multivariate way in logistic regression, sequential regression, PML regression and, PML and RFGPC Approaches is also present in our both methods, because the rank variable takes into account all explanatory variables. On the other hand, high number of permutations can be a disadvantage to obtain "optimal" hierarchy, but we have provided a reasonable solution with the Simulated Annealing algorithm. But our proposed method does not allow to provide individual weight for each explanatory variable. However, further researches are in the process of solution.

References

- Bachelet D., (1996), La mesure de la satisfaction du consommateur ou la chaine, l'arbre et la cascade, *ESOMAR, Customer Satisfaction*, 199-228, Madrid, Espagne.
- Derquenne Ch., (1998), Clustering of Mixture of Variables to Research Significant Ideas in Electric Heating Survey, IFCS, Roma, Italy.
- Derquenne Ch. & al., (2002), Une méthode alternative à l'approche PLS : Comparaison et application aux modèles conceptuels marketing, 35^{èmes} *Journées de Statistique*, Belgique.
- Derquenne Ch., (2003), A Multivariate Modeling Method based on the Partial Maximum Likelihood, *Third International Symp. on PLS and Related Methods*, Lisbon, Portugal.
- Derquenne Ch., (2005), Generalized Path Modelling based on the Partial Maximum Likelihood Approach, *PLS'05, Fourth Int. Symp. on PLS and Rel. Methods*, Barcelona, Spain.
- Kendall M. & al., (1990), *Rank Correlation Methods*, 5th Edition, Edwards Arnold.
- Kirkepatrick S., Gelatt C., Vecchi M., (1983), Optimization by Simulated Annealing, *Science*, 20, pp. 671-680.
- Pearl J., (1980), Bayesian Networks: A Model of Self-activated Memory for Evidential Reasoning, *In proceedings of the 7th Conference of the Cognitive Science Society*, University of California, Irvine, CA, pp. 329-334.
- Wold H., (1982), Soft modeling: the basic design and some extensions, In: K.G. Jreskog & H. Wold (eds.), *Systems under Indirect Observations: Causality, Structure, Prediction*, Vol 2, 1-54, Amsterdam, North-Holland.
- Wold S. & al., (1983), Patter Recognition: Finding and Using Regularities in Multivariate Data in *Proc. IUFOST Conf. "Food Research and Data Analysis"*, Martens J. (Ed.), Applied Sciences publications.