Multivariate Logistic Regression for the Estimate of Response Functions in the Conjoint Analysis

Amedeo De Luca[1],
Università Cattolica del "Sacro Cuore" di Milano, Dipartimento di Scienze Statistiche
amedeo.deluca@unicatt.it

Sara Ciapparelli,
Università Cattolica del "Sacro Cuore" di Milano, Dipartimento di Scienze Statistiche

Abstract: *In the Conjoint Analysis (COA) model proposed here - an extension of traditional COA - the polytomous response variable (i.e. evaluation of the overall desirability of alternative product profiles) is described by a sequence of binary variables. To link the categories of overall evaluation to the factors levels, we adopt a multivariate logistic regression model at the aggregate level. The model provides several overall desirability functions (aggregated part-worths sets), as many as the overall ordered categories are, unlike the traditional metric e non metric COA, which gives only one response function. We provide an application of the model.*

Keywords: **Aggregate Level Analysis**, **Conjoint analysis**, **Multivariate Logistic Regression**.

1. Introduction

In the Conjoint Analysis (COA) approach proposed here it is assumed that the respondent's evaluative judgement on the *overall* desirability, that is to be expressed on each profile of the new product, consists in a choice of one of the K desirability categories. To link the *overall* desirability (ordinal dependent variable Y, with modalities Y_k, $k = 1, 2, \ldots, K$) with the levels of experimental factors (independent variables X), the summarizing vector of the choice probability of one of the K said ordered categories has been interpreted via a multivariate multiple logistic regression model.

2. Estimation of Response Functions in the Conjoint Analysis

The model proposed concerns the full-profile COA and it is based on overall desirability categories chosen by a respondent's sample, for each of S hypothetical product profiles. The total number of profiles S, resulting from the total number of possible combinations of levels of the M attributes (X), constitute a *full-facto*rial experimental design. The focus of this study is to estimate the relationship between dependent and independent variables via a multivariate logistic regression model. In the approach, for a given respondent j, we let y_{ksj} denote the desirability category k of the sth concept for the respondent j. In terms of probabilities, the effects of the factors express the variations of the

[1] A. De Luca developed the model and its probability interpretation, and also wrote this note. S. Ciapparelli was responsible for the computer processing of the application.

probabilities p_{ks} - if k is the overall category - associated with the vector $\mathbf{z}_s$ corresponding to the combination s (s = 1, 2, …, S) of levels of the M factor, as follows:

$$p(Y_k = 1 \mid \mathbf{z}_s) = \pi_k(\mathbf{z}_s) = \exp(\delta_{k0} + \delta_{k1}^{'}\mathbf{z}_s) / [1 + \exp(\delta_{k0} + \delta_{k1}^{'}\mathbf{z}_s)] \tag{1}$$

where:

$\boldsymbol{\delta}^{'} = (\delta_0, \boldsymbol{\delta}_1^{'})$ is the unknown vector of regression coefficients of the predictor variables;

$\mathbf{z}_s$ is the vector of the dummy explanatory variables relative to the combination or concept s.

To estimate said probabilities $\pi_k(\mathbf{z}_s)$, we use an *aggregate* level model across the J homogeneous research respondents (Moore, 1980, p. 517), whose evaluations, on each product profile, are considered J repeated observations. To estimate the relationship between Y_k (k = 1, 2, …, K) dependent variable and m = 1, 2, ..., M, qualitative independent variables (factors X), with levels l = 1, 2, …, l_m, the K overall categories (Y_k,) are codified as K dummy variables; also the independent variables are codified as dummy variables (Z). The judgment evaluations are pooled across respondents (*pooled model*) and the novelty value in our approach is that one set of aggregated part-worths is estimated in connection with *each overall* category Y_k (see Table 1, Figure 1). In the configured multivariate model, owing to the interrelationship between the K dependent variables, the Kth equation can be drawn from the remaining q = K–1 equations. The q univariate logistic regression equations, *without* intercept, after transforming the dependent into a logit variable, are expressed as follows:

$$\begin{aligned} g_k(\mathbf{z}) &= \ln[\pi_k(\mathbf{z})/(1 - \pi_k(\mathbf{z})] \\ &= \mathbf{Z}\,\boldsymbol{\delta}_k, \qquad k = 1, 2, \ldots, q; \quad s = 1, 2, \ldots, S \end{aligned} \tag{2}$$

where: $g_k(\mathbf{z})$ is the *logit transformation*, $\mathbf{Z}$ is a fixed design matrix, $\boldsymbol{\delta}_k$ is a column vector of the unknown regression coefficients for the response function k.

To resolve the linear dependency between the independent variables the model is reparametrized using $z_1^{(m)}$ as reference category, and the model *with intercept* is, in alternative algebraic form:

$$g_k(\tilde{\mathbf{z}}_s) = \tilde{\delta}_{k0} + \sum_{m=1}^{M}\sum_{l=2}^{l_m} \tilde{\delta}_{kl}^{(m)}\, \tilde{z}_{lsj}^{(m)} + e_{ksj}, \tag{3}$$

where: $g_k(\tilde{\mathbf{z}}_s)$ is the logit of the sth profile with regard to the kth dependent variable; $\tilde{\delta}_{k0}$ is the constant term; $\tilde{\delta}_{kl}^{(m)}$ is the unknown regression coefficient for the lth level of the m factor; $\tilde{z}_{lsj}^{(m)}$ is the dummy variable for the lth level of the m factor in the combination s; e_{ksj} is the error term pertinent to the stimulus s and subject j (j = 1, 2, …, J). The $\tilde{\mathbf{Z}}$ denotes the design matrix below equation (3). The q equations $g_k(\tilde{\mathbf{z}}_s)$ can be expressed compactly as follows:

$$\mathbf{g}^* = \tilde{\mathbf{Z}}^*\, \tilde{\boldsymbol{\delta}}^*, \tag{4}$$

where: $\mathbf{g}^*$ is a compound vector (*vec*) of q column vectors $g_k(\tilde{\mathbf{z}})$; $\tilde{\mathbf{Z}}^*$ is a *square* compound diagonal matrix, containing q×q submatrices $\tilde{\mathbf{Z}}$; $\tilde{\boldsymbol{\delta}}^*$ is a compound vector of the q column vectors $\tilde{\boldsymbol{\delta}}_k$.

To estimate the (4) multivariate model parameters we need to consider the following variance-covariance matrix Φ, between the Y_k, with elements Var(Y_{ksj}) = $p_{ksj}(1-p_{ksj})$, where p_{ksj} is the probability for a j respondent to choose the k category for the s combination, and Cov(Y_{ksi}, Y_{qsj}) = -$p_{ksj}\,p_{qsj}$:

$$\Phi = \begin{bmatrix}
p_{111}(1-p_{111}) & \cdots & 0 & -p_{111}p_{211} & \cdots & 0 & -p_{111}p_{q11} & \cdots & 0 \\
\vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
0 & \cdots & p_{1SJ}(1-p_{1SJ}) & 0 & \cdots & -p_{1SJ}\,p_{2SJ} & 0 & \cdots & -p_{1SJ}\,p_{qSJ} \\
-p_{211}p_{111} & \cdots & 0 & p_{211}(1-p_{211}) & \cdots & 0 & -p_{211}p_{q11} & \cdots & 0 \\
\vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
0 & \cdots & -p_{2SJ}\,p_{1SJ} & 0 & \cdots & p_{2SJ}(1-p_{2SJ})\ \cdots & 0 & \cdots & -p_{2SJ}\,p_{qSJ} \\
 & & & & & \vdots & & & \\
-p_{q111}p_{111} & \cdots & 0 & -p_{q11}p_{211} & \cdots & 0 & p_{q11}(1-p_{q11}) & \cdots & 0 \\
\vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
0 & \cdots & -p_{qSJ}\,p_{1SJ} & 0 & \cdots & -p_{qSJ}\,p_{2SJ} & 0 & \cdots & p_{qSJ}(1-p_{qSJ})
\end{bmatrix}$$

The estimates of the Φ matrix elements are calculated on the basis of estimations $\hat{p}_{ksj}$ obtained by performing logistic regression analysis separately on each dependent variable, using the maximum likelihood method to each equation (3). To estimate the multivariate logistic regression model (4) we minimize the following mathematical expression (where $\hat{\Phi}^{-1}$ is the inverse matrix of the $\hat{\Phi}$):

$$F = (\mathbf{g}^* - \tilde{\mathbf{Z}}^*\,\tilde{\delta}^*)'\,\hat{\Phi}^{-1}\,(\mathbf{g}^* - \tilde{\mathbf{Z}}^*\,\tilde{\delta}^*). \tag{5}$$

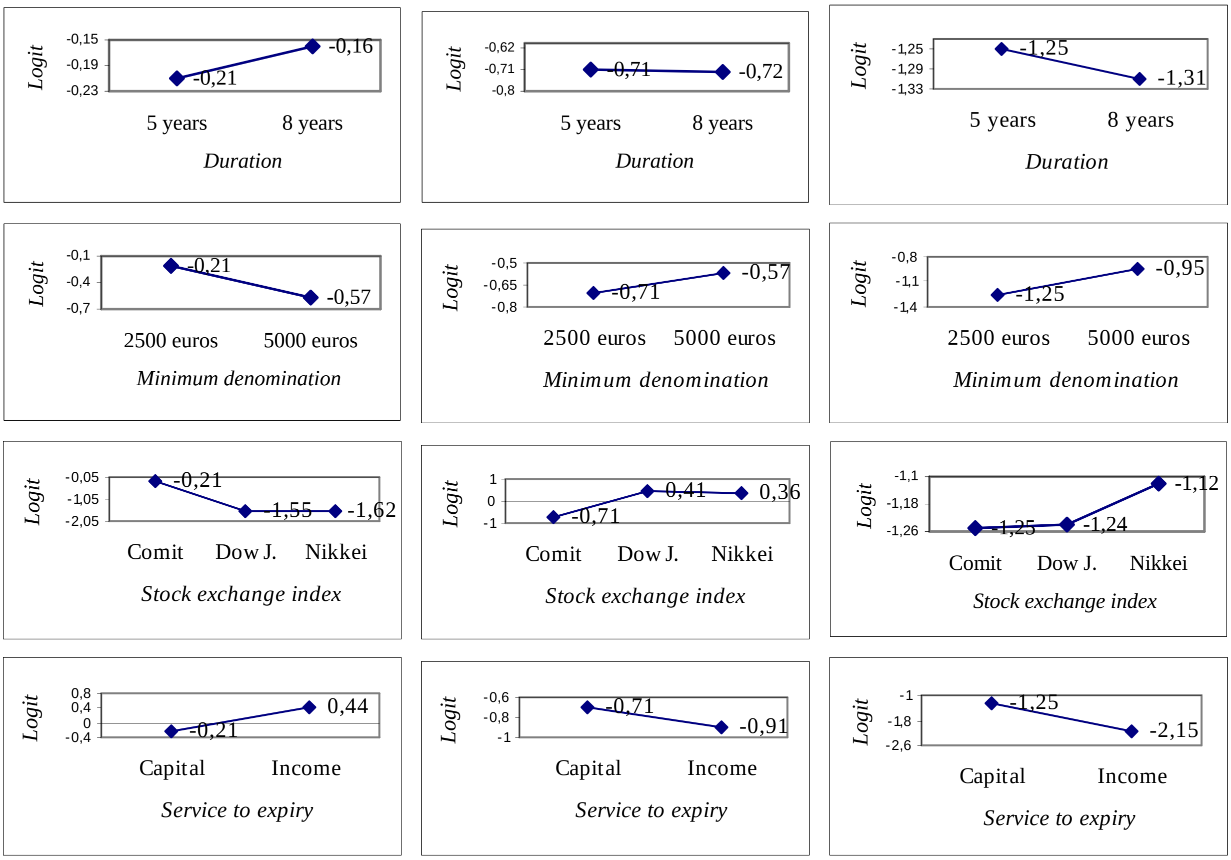
3. The Application of the Proposed Model and Conclusions

The model was applied to the overall desirability evaluations expressed on the K = 3 categories: “undesirable”, “desirable”, “more desirable”, by a sample of J = 100 insurance officers (homogeneous respondents) on S = 24 profiles of the insurance policy. The M = 4 attributes were: X_1 = “policy duration” (with levels: 5, 8 years); X_2 = “minimum denomination” (2,500 €, 5,000 €); X_3 = “stock exchange index” (Comit, Dow Jones, Nikkei); X_4 = “service to expiry” (paid-up capital, income for life). To estimate the parameters of the response functions of the (5) was used the Constrained Non Linear Regression (CNLR) program of the SPSS software. These estimates are given in Table 1.

Overall category	Estimated coefficient of the 1st equation		Overall category	Estimated coefficient of the 2nd equation		Overall category	Estimated coefficient of the 3rd equation	
“undesirable”	$\tilde{\delta}_1$	-0,21	“desirable”	$\tilde{\delta}_2$	-0,71	“more desirable”	$\tilde{\delta}_3$	-1,25
	$\tilde{\delta}_{12}^{(1)}$	0,05		$\tilde{\delta}_{22}^{(1)}$	-0,01		$\tilde{\delta}_{32}^{(1)}$	-0,06
	$\tilde{\delta}_{12}^{(2)}$	-0,36		$\tilde{\delta}_{22}^{(2)}$	0,14		$\tilde{\delta}_{32}^{(2)}$	0,30
	$\tilde{\delta}_{12}^{(3)}$	-1,34		$\tilde{\delta}_{22}^{(3)}$	1,12		$\tilde{\delta}_{32}^{(3)}$	0,01
	$\tilde{\delta}_{13}^{(3)}$	-1,41		$\tilde{\delta}_{23}^{(3)}$	1,07		$\tilde{\delta}_{33}^{(3)}$	0,13
	$\tilde{\delta}_{12}^{(4)}$	0,65		$\tilde{\delta}_{22}^{(4)}$	-0,20		$\tilde{\delta}_{32}^{(4)}$	-0,90

Table 1: *Estimates of three set of the aggregated part-worths utilities of the COA model*

A reading of the coefficients in Table 1 allows us to see the levels of the factors that contribute to the increase/decrease of the $\hat{p}_{ksj}$ (k = 1, 2, 3) values and the relative importance of each attribute as well as which attribute levels are most preferred. The analysis model here proposed provides two main advantages: the use of the probability $\hat{p}_{ks}$ as an average response, which does not require scale adjustments to render the preference scale "metric", and a cross-check of the effects of the attribute levels on the different k categories. Figure 1 shows graphs of the regression coefficients values, which are equal to the constant term plus the corresponding parameters given in Table 1.



a) The first response function b) The second response function c) The third response function

Figure 1: Sets of the aggregated part-worth scores for each of the three response functions

Bibliography

De Luca A., (2006), A logit model with a variable response and predictors on an ordinal scale to measure customer satisfaction, Quality and Reliability Engineering International, 22: 591-602.

Green P.E., Rao V.R. (1971), Conjoint measurement for quantifying judgmental data, Journal of Marketing Research, 8: 355-363.

Hosmer D.W., Lemeshow S. (1989), Applied Logistic Regression, Wiley, New York.

Moore W.L. (1980) Levels of aggregation in conjoint analysis: an empirical comparison, Journal of Marketing Research, 17: 516-523.