

Multivariate Multiple Regression on Dummy Variables for the Estimate of the Response Function in the Conjoint Analysis

*Regressione Multipla Multivariata su Variabili Dummy per la Stima della
Funzione di Risposta nella Conjoint Analysis*

Amedeo De Luca, Sara Ciapparelli¹

Università Cattolica di Milano, Istituto di Statistica, amedeo.deluca@unicatt.it

Riassunto: Per la stima delle funzioni di utilità parziali dei fattori sperimentali – nella *Conjoint Analysis* (COA) condotta sulla base di punteggi di gradimento globale, espressi su versioni alternative di un nuovo prodotto – la variabile di risposta (*overall*) è descritta come funzione di variabili indicatrici di natura dicotomica. Per collegare l'*overall* alle modalità degli attributi del prodotto si propone un modello di regressione lineare multivariata generalizzata vincolata, la quale evita il ricorso alla tradizionale trasformazione di scala (operata, tipicamente, con la regressione monotona di Kruskal) che renda “metrica” la scala di gradimento. I parametri del modello COA, di livello *aggregato*, sono stimati con la programmazione quadratica. Del modello proposto si dà un’applicazione.

Keywords: Aggregate level analyses, Conjoint analysis, Ordinal response.

1. Introduction

In a market study performed using the Conjoint Analysis (COA) methodology it is assumed that the person’s judgement on the *overall* desirability that is to be expressed on each profile of the new product consists in a choice of the K categories.

To link the *overall* desirability (Y_k , $k = 1, 2, \dots, K$), ordinal dependent variable, with the levels of experimental factors (independent variables X) relative to the product the summarizing vector of the choice probability of one of the K said ordered categories or modalities has been interpreted via a multivariate multiple regression.

2. Estimation of the Response Function in Conjoint Analysis with a Restricted Multivariate Multiple Regression on Indicator Variables

This approach considers COA based on *satisfaction* or *utility* scores assigned by a subject’s sample for each of S hypothetical profiles of a product. The total number of profiles or cards resulting from all possible combinations of the M factors (X) levels constitute a *full factorial* design. The focus of this study is to estimate the relationship be-

¹ A. De Luca developed the concept, methodological system and probability interpretation of the model, and also wrote this note. S. Ciapparelli collaborated in the formal description of the model and was responsible for the computer processing of the application.

tween independent and dependent variables via a multivariate model (De Luca, Zanella, Cantaluppi, 2004). With the parametrization adopted in the proposed model the effects of the factors express the variations of the probabilities p_{ks} resulting from the different combinations s ($s = 1, 2, \dots, S$) of the M factor levels, with respect to a reference group (*baseline*). To estimate said probabilities p_{ks} we use an aggregate level model and the person's judgements on each experimental stimulus (with the full concept method) are considered J repeated observations (Moore, 1980, p. 517). To estimate the relationship between Y_k ($k = 1, 2, \dots, K$) and M qualitative independent variables (factors X_l , with levels $l = 1, 2, \dots, l_m$; $m = 1, 2, \dots, M$) the K overall categories (Y_k) are expressed in function of K binary variables, that assume value 1 if the respondent j ($j = 1, 2, \dots, J$) chooses k of the ordered categories – if category k th is observed – 0 otherwise; the factors levels are codified by the Z_l binary variable, so that $Z_l = 1$ if level l th is observed, in all other cases, $Z_l = 0$. The preference judgments are pooled across all respondents (*pooled model*) and one set of overall utility function is estimated for each category Y_k . In the configured multivariate regression model the K linear regression equations on indicator variables are interconnected, consequently the K th equation can be drawn from the remaining $q = K-1$ equations. The q univariate regression equations, *without intercept*, are expressed as follows:

$$\mathbf{Y}_k = \mathbf{Z} \boldsymbol{\delta}_k + \mathbf{e}_k, \quad k = 1, 2, \dots, q, \quad (1)$$

where: $\mathbf{y}_k$ is the $(SJ) \times 1$ column vector, with generic observation y_{ksj} on the k th dependent variable associated with stimulus s and to respondent j ; $\mathbf{Z}$ is a $(SJ) \times \sum_{m=1}^M l_m$ fixed matrix of the experimental design; $\boldsymbol{\delta}_k$ is a column vector, with dimensions $\sum_{m=1}^M l_m \times 1$, of the unknown regression parameters; $\mathbf{e}_k$ is a $(SJ) \times 1$ column vector of *heteroscedastic* errors. The algebraic form of the model *with intercept*, reparametrized following the drop – for the M factors – of the column related to the constant term (*baseline*) correspondent to the first category ($\tilde{z}_1^{(m)}$) for obtaining univocal estimates of the parameters, is:

$$y_{ks} = \tilde{c}_k + \sum_{m=1}^M \sum_{l=2}^{l_m} \tilde{\delta}_{kl}^{(m)} \tilde{z}_{ls}^{(m)} + e_{ks}, \quad k = 1, \dots, q; s = 1, 2, \dots, S, \quad (2)$$

where: $\tilde{c}_k$ is conditional mean $E(Y_{k|\tilde{z}_1^{(1)}, \tilde{z}_1^{(2)}, \dots, \tilde{z}_1^{(M)}}) = \tilde{\delta}_{k1}$; $\tilde{\delta}_{kl}^{(m)}$ is the unknown parameter relevant to the l modality of the m factor; $\tilde{z}_{ls}^{(m)}$ is the indicator variable correspondent to l level of the m factor in the stimulus s ; e_{ks} is the error term pertinent to the s th stimulus. The $\tilde{\mathbf{Z}}$ experimental matrix below equation (2) has dimension $(JS) \times (1 + \sum_{m=1}^M l_m - M)$. The q equations (2), can be expressed compactly as follows:

$$\mathbf{y}^* = \tilde{\mathbf{Z}}^* \tilde{\boldsymbol{\delta}}^* + \mathbf{e}^*, \quad (3)$$

where: $\mathbf{y}^*$ is a compound vector (*vec*) from q column vectors, $(SJ) \times 1$, each of which contains the observations of the dependent variable for each of the J respondents on the

S stimuli; $\tilde{\mathbf{Z}}^*$ is a $[q(JS)] \times [q(1 + \sum_{m=1}^M l_m - M)]$ square compound diagonal matrix, containing $q \times q$ submatrices, of which the q $\tilde{\mathbf{Z}}$ submatrices disposed along the principal diagonal (all equal) give the independent indicator variables relative to the various equations in the column, while the remaining submatrices are compounds of zero elements; δ^* is a compound vector of the q column vectors of the regression coefficients $\tilde{\delta}_k$, each with dimensions $(1 + \sum_{m=1}^M l_m - M) \times 1$; $\mathbf{e}^*$ is a compound vector of q column vectors of the errors $\mathbf{e}_k$. Referring to (3), the *inequality* condition: $0 \leq p_{ks} \leq 1$ is expressed – indicating with $\tilde{\mathbf{z}}'_s$ the indicator variables vector of the stimulus s of the $\tilde{\mathbf{Z}}$ matrix – in the constraint:

$$0 \leq \tilde{\mathbf{z}}'_s \tilde{\delta}_k \leq 1, \quad k = 1, 2, \dots, q; s = 1, 2, \dots, S. \quad (4)$$

For the estimation of the model (2) parameters, the (4) requires the application of the Constrained Least Squares method and the Quadratic Programming (QP) for each univariate equation. To later obtain the estimation of the multivariate model (3) parameters, we need to consider the covariances between the Y_k , $k = 1, 2, \dots, q$. If we consider that for each respondent j the y_{kj} evaluation is described by a multinomial observation with K components and that there is a stochastic independence at the j variation, the variance-covariance matrix of the model can be defined as follows:

$$\Phi = \begin{bmatrix} p_{111}(1-p_{111}) & \cdots & 0 & -p_{111}p_{211} & \cdots & 0 & -p_{111}p_{q11} & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & p_{1sJ}(1-p_{1sJ}) & 0 & \cdots & -p_{1sJ}p_{2sJ} & 0 & \cdots & -p_{1sJ}p_{qsJ} \\ -p_{211}p_{111} & \cdots & 0 & p_{211}(1-p_{211}) & \cdots & 0 & -p_{211}p_{q11} & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & -p_{2sJ}p_{1sJ} & 0 & \cdots & p_{2sJ}(1-p_{2sJ}) & \cdots & 0 & \cdots & -p_{2sJ}p_{qsJ} \\ \vdots & & \vdots & & & \vdots & & & \vdots \\ -p_{q11}p_{111} & \cdots & 0 & -p_{q11}p_{211} & \cdots & 0 & p_{q11}(1-p_{q11}) & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & -p_{qsJ}p_{1sJ} & 0 & \cdots & -p_{qsJ}p_{2sJ} & 0 & \cdots & p_{qsJ}(1-p_{qsJ}) \end{bmatrix}$$

The estimations of the Φ matrix elements are calculated on the basis of estimations $\hat{p}_{ksj}$ ($k = 1, 2, \dots, q$), obtained by applying the *Generalized Least Squares* (GLS) method to each equation. To estimate the multivariate regression model using the GLS method, we minimize, under the constraints (4), and under the condition: $\sum_{k=1}^K \tilde{\mathbf{z}}'_j \tilde{\delta}_k = 1$, the following mathematical expression (where $\hat{\Phi}^{-1}$ is the inverse matrix of the $\hat{\Phi}$):

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

3. The Application of the Proposed Model and Conclusions

The model was applied to the overall evaluations (Y_k) expressed (on a graduate scale of

0 -10 – where the higher the score, the greater the preference – and accordingly reclassified into the $K = 3$ categories: “undesirable”, “desirable”, “more desirable”, in order to simplify the model) by a judgmental sample of $J = 100$ insurance officers on $S = 24$ profiles of the *insurance policy*. The $M = 4$ experimental factors were: X_1 = “policy duration” (with levels: 5, 8 years); X_2 = “minimum denomination” (levels: 2,500, 5,000 euros); X_3 = “stock exchange index” (levels: Comit, Dow Jones, Nikkei); X_4 = “service to expiry” (levels: paid-up capital, income for life). The estimators (part-worths), calculated with the QP CNLR program of the SPSS software, are given in Table 1; the standard errors are indicated in parentheses. The percent of explained variance of the model (De Luca *et al.*, 2004) is 45,89%.

Table 1: Estimates of parameters and relative standard errors of the COA model

Overall category	Estimated coefficient of the first equation	Overall category	Estimated coefficient of the second equation	Overall category	Estimated coefficient of the third equation
“undesirable”	$\tilde{c}_1$ 0,4297 (0,0019)	“desirable”	$\tilde{c}_2$ 0,3664 (0,0021)	“more desirable”	$\tilde{c}_3$ 0,2039 (0,0346)
	$\tilde{\delta}_{12}^{(1)}$ 0,0243 (0,0010)		$\tilde{\delta}_{22}^{(1)}$ -0,0093 (0,0012)		$\tilde{\delta}_{32}^{(1)}$ -0,0150 (0,0245)
	$\tilde{\delta}_{12}^{(2)}$ -0,0063 (0,0010)		$\tilde{\delta}_{22}^{(2)}$ 0,0036 (0,0012)		$\tilde{\delta}_{32}^{(2)}$ 0,0027 (0,0245)
	$\tilde{\delta}_{12}^{(3)}$ -0,3321 (0,0018)		$\tilde{\delta}_{22}^{(3)}$ 0,2788 (0,0022)		$\tilde{\delta}_{32}^{(3)}$ 0,0533 (0,0316)
	$\tilde{\delta}_{13}^{(3)}$ -0,3396 (0,0018)		$\tilde{\delta}_{23}^{(3)}$ 0,2563 (0,0022)		$\tilde{\delta}_{33}^{(3)}$ 0,0833 (0,0917)
	$\tilde{\delta}_{12}^{(4)}$ 0,1956 (0,0011)		$\tilde{\delta}_{22}^{(4)}$ -0,0959 (0,0014)		$\tilde{\delta}_{32}^{(4)}$ -0,0997 (0,0265)

With regard to the estimator properties, it must be noted that the sampling properties of the inequality-restricted GLS estimators ($\tilde{\delta}$) – when Φ is estimated – is still an open question in literature. On the other hand, since in our application the estimators associated with the restricted minimum relative to (5) coincide with those of the absolute minimum, we can state (De Luca *et al.*, 2004, pp. 586-587) that the estimators $\tilde{\delta}_{kl}^{(m)}$ are tendentially – as Φ is estimated – unbiased and efficient. In addition, these estimators have the following properties: $\sum_{k=1}^K \tilde{c}_k = 1$; $\sum_{k=1}^K \tilde{\delta}_{kl}^{(m)} = 0$, $l = 2, 3, \dots, l_m$; $m = 1, 2, \dots, M$.

A reading of the coefficients (effects) in Table 1, we can see the modalities of the factors that contribute to the increase/decrease of the p_k ($k = 1, 2, 3$) values, and consequently the relative importance of each attribute as well as which levels of each attribute are most preferred. The analysis model here proposed provides the advantage of the use of the $\hat{p}_{ks}$ satisfaction probability as an “average response”, which does not required scale adjustments to render the preference scale “metric”.

References

- De Luca A., Zanella A., Cantaluppi G. (2004) Un modello di valutazione della customer satisfaction con variabile risposta politomica, *Statistica Applicata*, vol. 16, n. 4, 563-598.
- Moore W.L. (1980) Levels of aggregation in conjoint analysis: an empirical comparison, *Journal of Marketing Research*, 17, 516-523.