STATISTICA, anno LXVIII, n. 2, 2008

DUMMY COVARIATES IN CUB MODELS

Maria Iannario

1. INTRODUCTION

In several disciplines, the analysis of qualitative data and ordinal variables is a relevant issue for the study of phenomena whose information are obtained by surveying a sample of population[1]. Thus, in Marketing and Medicine, in Linguistics and Economics, in Psychology and Cognitive sciences it is quite common to ask people to express ordered evaluation on several items (*rating analysis*) or to grade preference/liking/concern towards a set of comparable items/objects/services from the best to the worst (*ranking analysis*).

In fact, in rating analysis distinct but generally dependent qualitative evaluations are expressed by subjects on several related items using a convenient scale; instead, in ranking analysis people are asked to order a set of objects/items/services from the best to the worst degree of feeling/affection/concern, and so on. Although there is a substantial difference between these two methods of collecting information on ordinal data, in both cases the survey ends up with a sequence of integers expressing preferences or feeling with respect to a set of items.

Thus, it is worth to analyse such topics within a probabilistic framework where the response is a discrete random variable whose distribution and moments should be consistent with the observed behaviour of respondents. In fact, the study of ordinal data can not be simply assimilated to the discrete ones since a naive approach lowers both efficiency and interpretation.

One of the most accredited theory concerns Generalized Linear Models (GLM), promoted by Nelder and Wedderburn (1972), McCullagh and Nelder (1989) and for ordinal data by McCullagh (1980). Related studies of ordinal data derive from the development of multinomial logit models for *discrete-choice modelling* and *ordered response models*, as in Agresti (2002). A wide literature uses the latent variable approach as a convenient way to assess the distribution of multinomial responses, as mainly discussed by Moustaki (2000, 2003), Moustaki and Knott (2000), Cagnone *et al.* (2004), Bock and Moustaki (2007).

In this area, a random variable based on a mixture distribution has been for-

[1] We refer to Agresti (2002), Dobson (1990), Johnson and Albert (1999), Lloyd (1999), Marden (1995), Power and Xie (2000), Simonoff (2003) for a discussion about these topics.

mally introduced by D'Elia and Piccolo (2005a) and then applied to several data sets. Explicit reference to subjects' covariates (Piccolo, 2006) and objects' characteristics (Piccolo and D'Elia, 2008) has been also examined by the introduction of the class of *CUB* models. Specifically, in this paper, we deepen the role of covariates in *CUB* models when they are related to time lags, clusters, sub-populations, space, groups, that is under dichotomous circumstances; thus, they may be specified as dummy variables.

The paper is organized as follows: after an introduction to this class of models, in section 3 we specify the statistical implications of dummy covariates by emphasizing the interpretation of the estimated parameters. In section 4, a simulation study is pursued to show how the use of dummy covariates in *CUB* models allows a sharp discrimination among sub-populations. Then, in section 5 we support the previous analysis by some empirical evidence on real data sets both when dummy refers to the same survey and when different samples are to be compared. Some concluding remarks end the paper.

2. ORDINAL DATA AND *CUB* MODELS

If we observe how people select a single choice out of a set of m items (ranking) or assign a value (rating) within a range of ordered responses from 1 to m, we register two main factors that specify the outcome: a strictly *personal judgement* towards the objects caused by several latent factors and an *inherent indecision* in the choice mechanism that reflects the rater's uncertainty. Thus, the discrete response is a mixture of two elements (both continuous and latent) that should be modelled by discrete random variables.

The first component, that we call *feeling*, is the result of many unobservable subjective variables and thus it may be interpreted as a discretization of a Gaussian random variable. In this respect, by appropriate selection of the thresholds, a shifted Binomial random variable has been proved to be an effective choice to take into account several possibilities that arise when we transform a unimodal continuous random variable into a discrete one whose support is the set of the first m integers (D'Elia, 2000).

The second component, that we call *uncertainty*, is the result of the unavoidable indecision of any person taking a definite choice; then, it may be so extreme to give a constant probability to each element of the support (and thus the respondent acts by means of a purely random mechanism) or to be not present at all (and thus the respondent gives an answer only on the basis of the feeling component). In real cases the behaviour of respondents is intermediate between these two extreme situations. Thus, it seems reliable to model this component through the propensity of acting according to the Uniform discrete random variable defined on the support $\{1,2,\ldots,m\}$.

If we have a set of $m > 3$ discrete choices[2], with a given and known m, this

[2] The constraint $m > 3$ rules out the possibility of a degenerate random variable $(m = 1)$, and of an indeterminate $(m = 2)$ or saturated model $(m = 3)$.

choice mechanism implies that the observed response r_i of the *i*-th subject, in a sample survey of size *n*, is the realization of a discrete random variable *R* whose probability distribution is a mixture of a discrete Uniform and a shifted Binomial random variables (D'Elia and Piccolo, 2005a; 2005b). Then, it is defined by:

$$Pr(\mathrm{R}=r)=\pi\binom{m-1}{r-1}(1-\xi)^{r-1}\xi^{m-r}+(1-\pi)\frac{1}{m}, \qquad r=1,2,\ldots,m. \tag{1}$$

Since $\pi\in(0,1]$ and $\xi\in[0,1]$, the parametric space $\Omega(\pi,\xi)$ is the unit square:

$$\Omega(\pi,\xi)=\{(\pi,\xi):0<\pi\leq 1;\ 0\leq\xi\leq 1\}.$$

Recently, Iannario (2009) proved that this model is identifiable for $m>3$.

A strong point in favor of this distribution is the circumstance that its shape is extremely flexible as it accounts for right to left skewness, high peaked and platykurtic, symmetric and completely flat distributions (Piccolo, 2003).

For the interpretation of the parameters we observe that $(1-\pi)/m$ is the constant proportion of probability uniformly spread over the support, and we define this quantity as *uncertainty share*. Thus, the parameter π is inversely related to the uncertainty.

Instead, the feeling is related to ξ in the sense that if we rate a list of objects/items/services in such a way that the best is set to 1 and the worst to *m*, then a large ξ is a direct measure of our positive taste, preference, liking, etc. On the contrary, if we give a score to items (as a vote, increasing from 1 to m as the liking increases) then the interpretation of ξ is reversed, and $(1-\xi)$ must be considered as the actual measure of preference.

Although the value of the response is not metric (as it stems from a qualitative judgement), it may be useful for comparative purposes to compute the expected value of *R*:

$$E(\mathrm{R})=\pi(m-1)\left(\frac{1}{2}-\xi\right)+\frac{(m+1)}{2}. \tag{2}$$

In fact, this quantity is related to the mean value of the latent variable that expresses people feeling, and thus it is useful for comparative purposes.

Notice that both parameters contribute to assess the expected response. This is a relevant issue, since it confirms that - given the correctness of the model - the numerical value of the expectation is unable to convey all the information of the stochastic choice mechanism[3].

As a consequence, the introduction of subjects' covariates should not be related to the expectation (as it is common in GLM proposals); instead, it is con-

[3] Specifically, random variables specified by (1) with substantially different parameters values produce the same expectation (Piccolo, 2006, pp. 43-44).

venient to introduce a more general approach in order to take account of the circumstance that both parameters express different aspects of the choice mechanism.

When we relate the parameters of the mixture random variable to the subjects' covariates[4] we are defining *CUB* models introduced by Piccolo (2006). Although their logic is related to the GLM framework, the approach we adhere to is similar to King *et al.* (2000) who proposed only two components for a general statistical model:

– a *stochastic component* which defines the response variables R_i by a discrete probability distribution.

$$f(r;\boldsymbol{\theta}_i;\boldsymbol{\alpha}), \quad i=1,2,\dots,n$$

where the $\boldsymbol{\alpha}$ parameters may be also constants among subjects;

– a *systematic component* which explains the $\boldsymbol{\theta}_i$ parameters by means of explanatory variables $\mathbf{x}_i$ and parameter vector $\boldsymbol{\beta}$ that is:

$$\boldsymbol{\theta}_i = g(\mathbf{x}_i\boldsymbol{\beta}),$$

where $g(.)$ is the link function.

This simplified paradigm is more general than *GLM* structures since the probability distribution is not compelled to belong to exponential family and the parameters are not necessarily related to explanatory variables via the expectation. In fact, the link function is a complete general mapping between a real and a parametric space. However, from an operational point of view, the use of a logistic function is sufficient and adequate for the fitting and explanation of several data sets. Instead, it should be modified if some asymmetries are suspected in the tails of the distributions.

Suppose we have a sample of ordinal data $\mathbf{r}=(r_1,r_2,\dots,r_n)'$ and we collect several measurements on the subjects summarized in the $\mathbf{Y}$ and $\mathbf{W}$ matrices, whose *i*-th rows, for $i=1,2,\dots,n$, are defined by:

$$\mathbf{y}_i=(y_{i0},y_{i1},y_{i2},\dots,y_{ip}); \qquad \mathbf{w}_i=(w_{i0},w_{i1},w_{i2},\dots,w_{iq}),$$

respectively[5]. For establishing a consistent terminology, we use the acronyms *CUB(0,0)*, *CUB(p,0)*, *CUB(0,q)*, *CUB(p,q)* in order to refer to models without covariates, with covariates for π, with covariates for ξ, and with covariates for (π,ξ), respectively. Thus, the parameters to be estimated are denoted by $\boldsymbol{\theta}$ and are specified by: $(\pi,\xi)'$, $(\boldsymbol{\beta}',\xi)'$, $(\pi,\boldsymbol{\gamma}')'$ and $(\boldsymbol{\beta}',\boldsymbol{\gamma}')'$, respectively.

[4] In fact, it is possible to introduce also objects' covariates in C*UB* models, as in Piccolo and D'Elia (2008).

[5] For making the notation more compact, we introduce the variables Y_0 and W_0 that assume the constant value 1 for all the sample units.

Then, if we let:

$$\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'; \qquad \boldsymbol{\gamma} = (\gamma_0, \gamma_1, \ldots, \gamma_q)',$$

a general *CUB* model with a logistic link is defined, for any $i = 1, 2, \ldots, n$, by the following probability distribution[6]:

$$Pr(R = r \mid y_i, w_i, \beta, \gamma) = \frac{1}{1 + e^{-y_i\beta}} \left[\binom{m-1}{r-1} \frac{(e^{-w_i\gamma})^{r-1}}{(1 + e^{-w_i\gamma})^{m-1}} - \frac{1}{m} \right] + \frac{1}{m}. \tag{3}$$

Specifically, the log-likelihood function of a *CUB* model without covariates, given the observed frequencies n_r, that is the absolute frequencies of ($R = r,\ r = 1, 2, \ldots, m$), is defined by:

$$\log L(\pi, \xi) = \sum_{r=1}^{m} n_r \log \left[\pi \binom{m-1}{r-1} (1-\xi)^{r-1} \xi^{m-r} + (1-\pi)\frac{1}{m} \right]. \tag{4}$$

It may be easily generalized when covariates exist. Then, based on this extended information set, the joint efficient estimation of the parameters may be pursued by maximum likelihood (ML) method. Piccolo (2006) adapted the EM algorithm to such models and derived the observed information matrix of these estimators. Then, the significance of the parameters estimates, the relevance of the covariates and several asymptotic tests may be obtained.

Finally, the comparison among log-likelihoods and difference of deviances for the estimated *CUB* models (without and with covariates) are the common tools for checking and validate different models.

3. DUMMY COVARIATES AND *CUB* MODELS

In this section, we specify *CUB* models for checking significant differences in the behaviour of respondents when circumstances are dichotomous. This frequently happens when we compare agreement or satisfaction with respect to time, space, environment, gender, classes of age, groups of consumers, and so on.

In this regard, according to sampling procedures, we should distinguish two main cases:

– there are *two different samples* and we wish to compare the two observed distributions of the responses via the estimated models, under the constraint that all other circumstances are equal. For instance, we compare models estimated on two samples generated in different times, spaces, rules, and so on.

[6] Notice that we use different notations for the variables explaining the uncertainty and feeling of respondents, respectively. However, the previous definition is completely general since the Y 's variables (or a subset of them) may also coincide with the W 's variables (or a subset of them), as it will happen in some applications we will discuss in section 5.

– there is *a unique sample* and we wish to test if there is a significant effect on the behaviour of groups characterized by dichotomous situations. For instance, we compare the effect of gender, age classes, job, education in the same population.

The logical consequences and the statistical interpretation of the previous cases are different but, from a formal point of view, we may address them in a similar way. For easiness of notation, we will discuss the case of a *CUB* model where only a dummy variable is relevant.

Thus, we begin with a dichotomous situation where the sample is characterized by two groups G$_0$ and G$_1$, respectively (for instance, males and females, young and elderly, etc.). Denote by D_i a variable assuming values 0 and 1 when the i-th subject S_i, for $i=1,2,\dots,n$ belongs to one of the groups G$_0$ and G$_1$, respectively. Formally,

$$D_i=\begin{cases}0, & \text{if } S_i\in \mathrm{G}_0;\\ 1, & \text{if } S_i\in \mathrm{G}_1;\end{cases} \quad i=1,2,\dots,n.$$

If we suppose that this membership is relevant for explaining a different effect of the uncertainty and/or the feeling components, we specify a *CUB* model where the corresponding parameters are function of the dummy covariate, that is:

$$\pi_i=\frac{1}{1+e^{-(\beta_0+\phi D_i)}};\quad \xi_i=\frac{1}{1+e^{-(\gamma_0+\psi D_i)}},\ i=1,2,\dots,n.$$

This implies that uncertainty and feeling parameters in the two groups are:

$$(\pi_i\mid S_i\in \mathsf{G}_0)=\pi_0=\frac{1}{1+e^{-\beta_0}};\qquad (\pi_i\mid S_i\in \mathsf{G}_1)=\pi_1=\frac{1}{1+e^{-(\beta_0+\phi)}};$$

$$(\xi_i\mid S_i\in \mathsf{G}_0)=\xi_0=\frac{1}{1+e^{-\gamma_0}};\qquad (\xi_i\mid S_i\in \mathsf{G}_1)=\xi_1=\frac{1}{1+e^{-(\gamma_0+\psi)}};$$

A simple algebra shows that:

$$\pi_1>\pi_0 \Leftrightarrow \phi>0;\qquad \xi_1>\xi_0 \Leftrightarrow \psi>0.$$

Now, the uncertainty is inversely related to the parameter π; instead, the ξ parameter supports different interpretations according to the nature of the ordinal variable. More precisely, if we are working on *ranking data* (where $r=1$ denotes the maximum of preference/liking/concern towards the objects/services/items while $r=m$ denotes the minimum), then the parameter ξ is a direct measure of feeling. On the contrary, if we are working on *rating data* (and we score the items by giving $r=1$ and $r=m$ to the minimum and maximum satisfaction, respectively), then the parameter that measures the feeling is $(1-\xi)$.

The following scheme summarizes this discussion and offers an immediate interpretation of the relationships among the parameters of the dummy covariate and the components of the *CUB* model[7].

TABLE 1

Interpretation of model components

Dummy coefficients	Interpretation in terms of model components	CUB parameters
$\phi > 0$	Uncertainty (G_0) > Uncertainty (G_1)	$\pi_1 > \pi_0$
$\psi > 0$	Ranking: Feeling (G_0)< Feeling (G_1) Rating : Feeling (G_0)> Feeling (G_1)	$\xi_1 > \xi_0$

It is also interesting to relate the parameters β_0 or γ_0 of the *CUB* models to the uncertainty or feeling components of the groups. In fact, by a similar algebra, we obtain:

$$\beta_0 = \log\left(\frac{\pi_0}{1-\pi_0}\right); \quad \gamma_0 = \log\left(\frac{\xi_0}{1-\xi_0}\right).$$

Is is evident that the constants are only related to uncertainty and feeling of the G_0 group, which acts as a sort of reference group.

Similar expressions may be obtained for the parameters ϕ and ψ , respectively:

$$\phi = \log\left(\frac{\pi_1}{1-\pi_1}\right) - \log\left(\frac{\pi_0}{1-\pi_0}\right);$$

$$\psi = \log\left(\frac{\xi_1}{1-\xi_1}\right) - \log\left(\frac{\xi_0}{1-\xi_0}\right).$$

Then, we discuss the relationships among the conditional expectations of the response and the parameters, when one of them is fixed: this situation happens when only π or ξ is significantly explained by the dummy covariate.

Suppose that we are comparing the expected response for a given π . From the formula:

$$\mathrm{E}(R \mid D = j) = \frac{m+1}{2} + \pi(m-1)\left(\frac{1}{2} - \xi_j\right), \quad j = 1, 2, \tag{5}$$

we obtain:

$$\frac{\mathrm{E}(R \mid D = 1) - \mathrm{E}(R \mid D = 0)}{(m-1)\pi} = \xi_0 - \xi_1. \tag{6}$$

[7] Of course, for real data sets, we need to substitute the dummy coefficients with the corresponding estimated parameters.

Thus, an increase in the expectation of R (that is, a reduction of the agreement expressed by ranking the items) is strictly proportional to the corresponding decrease in the feeling parameters. The result confirms that - for a fixed π - feeling and expectation are inversely related.

Then, suppose that we are comparing the expected response for a given ξ, and only the uncertainty parameter is conditioned by the dummy covariate. In this case, we get:

$$\frac{\mathrm{E}(R \mid D=1)-\mathrm{E}(R \mid D=0)}{(m-1)\left(\dfrac{1}{2}-\xi\right)}=\pi_1-\pi_0 . \tag{7}$$

Thus, if we remember that the *uncertainty share* is $(1-\pi)/m$, the result shows that, for a given ξ, an increase in the expectation of R corresponds to a proportional reduction in the uncertainty component of the mixture when $\xi<0.5$, and vice versa when $\xi>0.5$.

Both results might be expressed as displacements of points on the parametric space $\Omega(\pi,\xi)$. In the first situation, when $\xi_1<\xi_0$, if we move vertically down from the point (π,ξ_0) to the point (π,ξ_1) we get a proportional increase from $\mathrm{E}(R \mid D=1)$ to $\mathrm{E}(R \mid D=0)$. Instead, in the second situation, when we move horizontally up from the point (π_0,ξ) to the point (π_1,ξ), the expectation $\mathrm{E}(R \mid D=1)$ increases with respect to $\mathrm{E}(R \mid D=0)$ if $\xi<0.5$, and decreases if $\xi>0.5$. Thus, the variation in the corresponding expectations depends also on the sign of $\left(\dfrac{1}{2}-\xi\right)$.

The previous discussion may be also pursued with reference to the shape of the distribution since, as noticed by Piccolo (2003), the expectation of R increases (decreases) as long as the asymmetry moves towards negative (positive) values. In this regard, we observe that this random variable is perfectly symmetric if and only if $\xi=1/2$, and the sign of the asymmetry is the same of $(\xi-1/2)$. Thus, the expected preference of the raters towards a fixed object increases (decreases) with respect to the mid-range together with the negative (positive) value of the asymmetry measure. Then, a positive (negative) asymmetry is associated with a preference (adversity) towards the object.

As a consequence, it is immediate to derive that γ_0 is a measure of symmetry of the probability distribution of the choices for the reference group G_0.

More precisely,

$\gamma_0<0 \Leftrightarrow$ Left skewness;

$\gamma_0=0 \Leftrightarrow$ Null skewness;

$\gamma_0>0 \Leftrightarrow$ Right skewness.

On the contrary, the G1 group distribution is symmetric if and only if $\gamma_0 + \psi = 0$, and thus:

$\gamma_0 + \psi < 0 \Leftrightarrow$ Right skewness;
$\gamma_0 + \psi = 0 \Leftrightarrow$ Null skewness;
$\gamma_0 + \psi > 0 \Leftrightarrow$ Left skewness.

These results may be summarized as follows:

$$Skewness = 0 \Rightarrow \xi = \frac{1}{2} \Rightarrow \begin{cases} \gamma_0 = 0, & \text{if } D_i = 0; \\ \gamma_0 = -\psi, & \text{if } D_i = 1. \end{cases}$$

It is worth to notice that all the results about the interpretation of the dummy variable parameters are still valid when several covariates are present in *CUB* models. In these cases, one should add the standard convention (as in multiple regression modelling) that the effects of the covariates are consistently explained *ceteris paribus*, that is all other variables being constant.

4. DISCRIMINATING POWER OF DUMMY VARIABLES IN *CUB* MODELS

In this section a simulation study is pursued to assess the discriminating power of *CUB* models with regard to the presence of two sub-populations. Although the experiment cannot be considered exhaustive, it strongly supports the usefulness of dummy covariates for discriminating purposes.

To set up the simulation design, we let $m = 9$ and suppose that two independent random samples of size n_0 and n_1, respectively, are generated by CUB model where the parameters (π, ξ) are specified as in Table 2. In this way, we are comparing two groups with a constant *uncertainty share* (measured by $(1-\pi)/m = 0.028$) and a shifting in feeling, measured by τ.

TABLE 2

Design of the simulation experiment

Groups	Parameters		Samples sizes				
G0	$\pi = 0.75$	$\xi_0 = 0.10$	100	150	200	150	250
G1	$\pi = 0.75$	$\xi_1 = 0.10 + \tau$	100	150	200	250	150

Then, on the basis of the observed samples, for increasing $\tau > 0$, we test:

$$H_0 : (\pi = 0.75,\ \xi = 0.10) \quad \textit{versus} \quad H_1 : (\pi = 0.75,\ \xi = 0.10 + \tau).$$

We reject *H0* when the log-likelihood ℓ_{01} of the estimated *CUB*(0,1) model with a dummy covariate for explaining a difference in feeling (ξ parameter) in the second sample is significantly greater than the corresponding log-likelihood ℓ_{00} of a *CUB*(0,0) model fitted to a joint sample. Thus, our asymptotic critical region of nominal size $\alpha = 0.05$ is defined by:

$$2(\ell_{01} - \ell_{00}) > \chi^2_{(0.05,1)} = 3.841.$$

We simulate 1000 times a couple of samples generated under *H₀* and *H₁*, respectively, and we estimate the probability of rejecting *H₀*. If $\tau = 0$, this probability is an estimate of the nominal size α, while for varying τ it is an estimate of the power function of the test as long as *H₁* differs from *H₀*. Formally, we define:

$$\gamma(\tau) = P_r(\textit{reject}\; H_0 \mid \tau).$$

For our experiment, we have chosen three balanced and two unbalanced sample sizes for *H₀* and *H₁*.

Table 3 presents the estimated power function for the different sample size combinations, and we list the main points derived from these results:

TABLE 3

Simulated probability of rejecting H₀, *given* $\tau \geq 0$

Cases	A	B	C	D	E
τ	$n_0 = 100$	$n_0 = 150$	$n_0 = 200$	$n_0 = 150$	$n_0 = 250$
	$n_1 = 100$	$n_1 = 150$	$n_1 = 200$	$n_1 = 250$	$n_1 = 150$
0.00	0.047	0.038	0.060	0.048	0.055
0.01	0.076	0.082	0.120	0.100	0.111
0.02	0.155	0.206	0.266	0.260	0.268
0.03	0.278	0.410	0.520	0.489	0.468
0.04	0.460	0.590	0.732	0.690	0.687
0.05	0.599	0.785	0.902	0.850	0.857
0.06	0.755	0.903	0.956	0.943	0.948
0.07	0.869	0.960	0.987	0.979	0.987
0.08	0.934	0.986	0.999	0.998	0.998
0.09	0.969	0.996	1.000	0.999	0.998
0.10	0.985	0.998	1.000	1.000	0.999
0.11	0.988	1.000	1.000	1.000	1.000
0.12	0.999	1.000	1.000	1.000	1.000
0.13	1.000	1.000	1.000	1.000	1.000

These issues are easily conformed by the patterns of power functions shown in Figure 1.

1. As expected, power function increases uniformly with sample size;

2. a small difference in feeling (as measured by $\tau = 0.08$, say) is already detected for moderate sample sizes with a probability always greater than 0.9;

3. the steep slope of the power function causes some problems with respect to the nominal size (for moderate sample size, it results lower than expected);

4. no serious problems arise when the sample sizes are not balanced, in both directions.

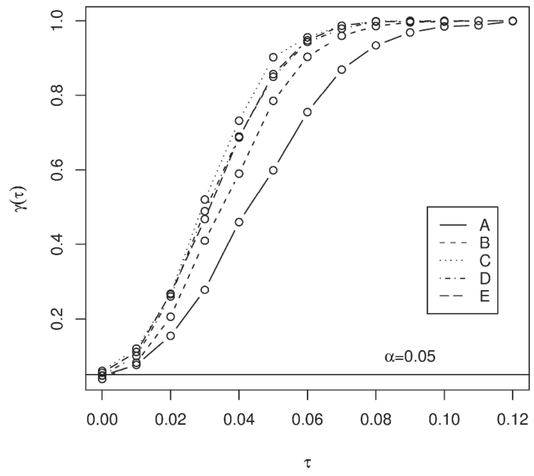

Figure 1 – Power functions for varying sample size.

5. SOME EMPIRICAL EVIDENCE

We will discuss the previous results with reference to some real data sets[8] where *CUB* models proved to be effective for explaining the behaviour of respondents' preferences.

In the first case, we measure the effect of gender as a significant covariate for explaining the agreement toward a color (thus, data refers to *rankings* of the items). In the second case, we introduce a dummy covariate to explain the characteristic of two groups attending University Orientation services (thus, data refers to *ratings* of a service). In the third case, we compare the difference in concern expressed by respondents with reference to serious problems in a large city during 2004 and 2006 surveys, respectively (again, data refers to *ranking* but we are joining two samples).

Thus, the first two case studies use dummy covariates within a unique sample while the last one refers to disjoint samples, collected in different times[9].

[8] To be precise, we cannot assure that these data sets are realizations of random samples as they are collected from people attending University in different circumstances. However, the distribution of the relevant covariates is not substantially different with respect to a general population of respondents and we found no *a priori* reason to invoke a selection bias for our samples; thus, statistical inference may be confidently pursued.

[9] For an effective statistical analysis of *CUB* models one should consider samples of moderate/large sizes. In our data sets, the minimum sample size is $n = 169$ (in the first instance). This is caused by the circumstance that if we perform some inference with a smaller data set it is high the probability to observe no ordinal values for some $r = 1, 2, \ldots, m$, and this implies inefficient evaluation of the probability distribution.

5.1. *Preferences towards colors*

This data set has been exploited in several studies on preference and motivated many advances in the methods and experiences on ordinal data (see: D'Elia *et al.*, 2001; D'Elia, 2003); it is concerned with a sample of $n = 169$ University students during 1998. We limit the following discussion to compare the ranking of Black color with reference to the dummy covariate Smoking (=0 nosmokers, =1 smokers).

Black color received a high score (the average rank is 4.248, among a set of 12 colors) with a sharp mode at $r = 1$ but with a significant *uncertainty share* of 4.8% distributed over all the support; thus, the total amount of uncertainty in the responses is estimated as 58%. Indeed, a $CUB\,(0{,}0)$ model shows a relevant feeling ($\hat{\xi} = 0.948$) but the great uncertainty expressed by respondents shifts the expectation towards values not so extreme. The fitting measure[10] $Diss = 0.104$ of the estimated model is acceptable (even the classical Chi-square test is not significant).

A substantial improvement in the model has been obtained by the inclusion of the smoking habit as explanatory covariate of the uncertainty parameters, as in Table 4. Notice that 34% of respondents declared to smoke.

TABLE 4

Estimation of CUB models for Black color preferences

Models	Uncertainty	Feeling	Log-lik
CUB(0,0)	$\hat{\pi} =\ \ 0.421\,(0.056)$	$\hat{\xi} = 0.948\,(0.014)$	-374.596
CUB(1,0)	$\hat{\beta}_0 = -0.791\,(0.305)$	$\hat{\xi} = 0.948\,(0.014)$	-370.135
	$\hat{\phi} =\ \ 1.313\,(0.450)$		

The asymptotic likelihood ratio test, as implied by: $2\,(\ell_{10} - \ell_{00}) = 8.922$, is highly significant to confirm the improvement acquired by the introduction of the dummy covariate.

If we apply the interpretation discussed in section 3, we observe that $\hat{\phi} > 0$. Thus, when we move from the group G_0 (nosmokers) to G_1 (smokers) we expect an increase in the parameter π, and so a decrease in uncertainty.

In Table 5, we compare the distribution of respondents of two groups, and it seems evident that the responses of smokers are mostly concentrated on the first two ranks. This confirms the more limited uncertainty of smokers respondents.

[10] The dissimilarity index is a normalized sum of the absolute differences among observed relative frequencies and estimated probabilities of a given model. It measures the proportion of respondents to move among categories in order to reach a perfect fit.

TABLE 5

Black color preferences with respect to Smoking

Groups	1	2	3	4	5	6	7	8	9	10	11	12
G_0 (nosmokers)	27	15	11	12	5	4	8	2	7	7	10	3
G_1 (smokers)	23	13	4	5	5	0	1	1	3	2	0	1

The estimated *CUB*(0,1) accounts for this result by producing the same ξ estimate (for the feeling component) and two π estimates (as related to uncertainty) expressed, respectively, by:

$$\hat{\pi}_0 = \frac{1}{1+e^{0.791}} = 0.312; \qquad \hat{\pi}_1 = \frac{1}{1+e^{0.791-1.313}} = 0.628.$$

It turns out that the estimated uncertainty parameter expressed by nosmokers is about twice than smokers.

In Table 6, we compare the estimated distributions of *CUB* (0,0) and *CUB*(1,0) models (given $D_i = 0$ and $D_i = 1$, respectively) with the corresponding relative frequencies f_r of the two sample groups. Thus, the splitting into two subsets produces also a closer fitting within the groups.

TABLE 6

Observed and estimated distributions, conditioned by Smoking

	Sample (n=169)		Nosmokers (D=0)		Smokers (D=1)	
r	fr	Pr(R=r)	fr	Pr(R=r)	fr	Pr(R=r)
1	0.29586	0.28155	0.24324	0.23037	0.39655	0.37916
2	0.16568	0.18990	0.13514	0.16208	0.22414	0.24176
3	0.08876	0.08733	0.09910	0.08616	0.06897	0.08902
4	0.10059	0.05470	0.10811	0.06210	0.08621	0.04061
5	0.05917	0.04893	0.04505	0.05787	0.08621	0.03209
6	0.02367	0.04827	0.03604	0.05738	0.00000	0.03112
7	0.05325	0.04822	0.07207	0.05734	0.01724	0.03104
8	0.01775	0.04822	0.01802	0.05734	0.01724	0.03104
9	0.05917	0.04822	0.06306	0.05734	0.05172	0.03104
10	0.05325	0.04822	0.06306	0.05734	0.03448	0.03104
11	0.05917	0.04822	0.09009	0.05734	0.00000	0.03104
12	0.02367	0.04822	0.02703	0.05734	0.01724	0.03104

5.2. *Rating of Orientation services*

This data set derives from a survey that has been carried out by University of Naples Federico II, at the end of each year, with reference to an extensive Orientation program provided to its students, in order to check the students' satisfaction. The survey is based on a questionnaire where each student was asked to give a score for expressing his/her overall satisfaction towards the Orientation service, only if he/she used it during the year. The answers range from 1 ("completely unsatisfied") to 7 ("completely satisfied") and they concern 5 items: Willingness (W) and Competence (C) of the Orientation staff, Clearness of the information (I), Adequateness of timetable (T), and a Global evaluation (G).

The complete data set consists of $n = 2000$, 2457, 2975 subjects for the years 2002, 2003, 2004, respectively[11]. In the following pages, we limit ourselves to introduce a dummy covariate for explaining the responses to the global evaluation, and specifically[12] we analyse "Frequency of usage" (=0 if occasionally, =1 if very frequent) on the global evaluation of the services during 2003.

As a matter of fact, the global satisfaction towards the service received high scores with a sample average of 5.626 and a mode at $(R = 6)$; moreover, as shown in Figure 2, a *CUB* model is well fitted to the responses.

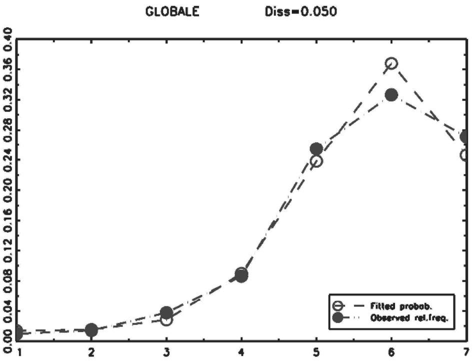

Figure 2 – Observed and estimated distribution for global preference.

Table 7 presents *CUB* models fitted to global preference, for all data and for the two subsets characterized by the "Frequency of usage". For a correct interpretation, one should remember that uncertainty is related to $(1-\pi)$ while the positive feeling towards the service is now related to $(1-\xi)$, since the response is a score whose value increases with a positive evaluation of the service.

TABLE 7

CUB models of Global satisfaction for Orientation services

Parameters	Occasionally (=0)	Very frequent (=1)	Whole sample
$\hat{\pi}$	0.912 (0.014)	0.923 (0.016)	0.904 (0.012)
$\hat{\xi}$	0.242 (0.005)	0.129 (0.006)	0.202 (0.004)
n	1657	800	2457
Diss	0.037	0.047	0.050
AIC/n	3.097	2.637	3.030

[11] Complete documentation reports and several information about this project are available on *http://www.dipstat.unina.it/ricerca/progettoVER*. Some CUB models for these data sets were also discussed by Iannario and Piccolo (2009) in a work related to Customer Satisfaction analysis.

[12] Here, we will not discuss the open problem of the selection of covariates in CUB models. In this case, the choice might be suggested by a substantial difference of the expressed average scores between the groups, that is 5.413 and 6.065 when the "Frequency of usage" is 0 and 1, respectively. The large sample sizes and the limited standard deviation of the scores $(\cong 1.1/1.2)$ support the significance of such difference.

We observe that the groups present the same uncertainty in the response but feeling parameter is about doubled when we move from occasional to frequent users. This aspect is more evident if we compare the two distributions (Figure 3) where different modes and shapes are shown.

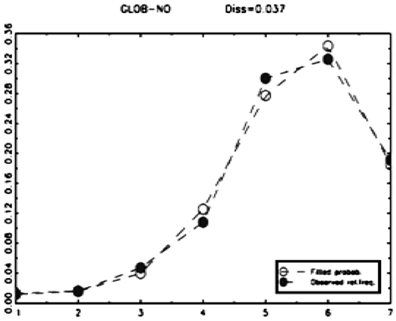

Figure 3 – Global preference: occasional users.

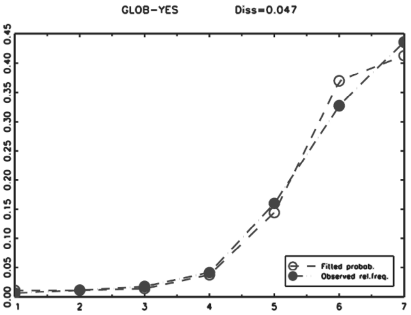

Figure 4 – Global preference: frequent users.

The results suggest to include a dummy covariate to model the effect of "Frequency of usage" of the services, and we obtained the *CUB*(0,1) model estimates as in Table 8.

TABLE 8

Estimated CUB models for Orientation services

Models	Estimated parameters	Log-lik	AIC
CUB(0,0)	$\hat{\pi} =$ 0.904 (0.012)	-3720.285	7444.570
	$\hat{\xi} =$ 0.202 (0.004)		
CUB(0,1)	$\hat{\pi} =$ 0.917 (0.011)	-3612.769	7231.538
Frequency of usage	$\hat{\gamma}_0 = -1.138$ (0.027)		
	$\hat{\psi} = -0.775$ (0.056)		

We observe a high significance of parameters and a sensible reduction of log-likelihood functions caused by the dummy covariate. Moreover, as expected from the schemes of section 3, the parameter $\psi < 0$ suggests that moving from the group G_0 (=occasional users) to G_1 (=very frequent users), we get an increase in feeling. In fact, frequent users present a distribution with a sharp mode at (R=7), that is the maximum score.

5.3. *Time differences about concern on urban problems*

In this data set a sample of respondents living in Naples ranked the main perceived problems of the city among a list of 9 preselected items[13]. We classified

[13] After a preliminary investigation, the following items were selected as relevant: 1. *Political patronage and corruption (CORRUP)*. 2. *Organized crime (ORCRIM)*. 3. *Unemployment (UNEMPL)*. 4. *En-*

them as "emergencies" and asked to people to rank them with respect to the degree of worry/anxiety/concern they caused. Thus, in this case study, *feeling* is indeed the degree of *concern* about an urban problem.

The survey was submitted for the first time in December 2004 and repeated in December 2006 for a similar sample, in a rigorous and homogeneous way[14], and we get complete answers from n_0=354 and n_1=419 respondents, respectively. In order to have a general idea of the difference of the rankings expressed by respondents in different years, we present in Figure 5 the parametric space where the estimated parameters of the *CUB* (0,0) models are shown.

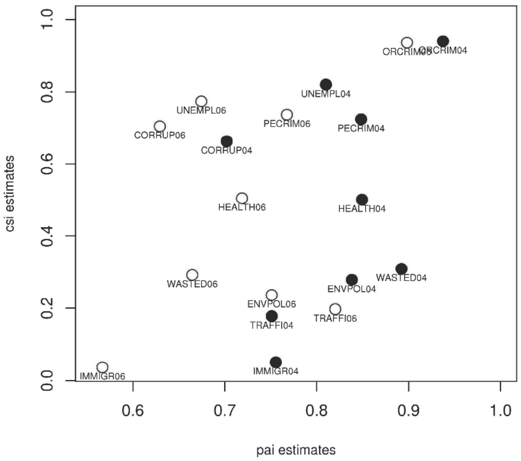

Figure 5 – Estimated CUB models for 2004 (•) and 2006 (◦) surveys.

We register a substantial stability about the level of the perception of these problems, and the representation[15] in the parametric space is a useful tool for visualizing the joint dynamics of both concern and uncertainty of respondents.

First, we register a quite stable ordering of the items between the years[16], as only few of them register a small variation in concern, as for instance *Unemploy-*

vironmental pollution (ENVPOL). 5. *Public health shortcomings (HEALTH)*. 6. *Petty crimes (PECRIM)*. 7. *Immigration (IMMIGR)*. 8. *Streets cleanness and waste disposal (WASTED)*. 9. *Traffic and local transport (TRAFFI)*. The list were submitted as indicated, according to Italian alphabetic order.

[14] Gender and age distributions, professions, residences, etc. are quite similar among the samples. Further analyses about these data sets are discussed by D'Elia and Piccolo (2005b) and Iannario (2007) for the 2004 and 2006 surveys, respectively.

[15] For a correct interpretation of Figure 5, one should be aware that π scale have been doubled to allow a sharper reading of the graph. Thus, vertical displacements are relatively more important than horizontal ones.

[16] It seems surprising that different samples of hundredth of people give exactly the same global ordering to 9 items after two years; however, this confirms both the robustness of the procedure for detecting the real perception of residents and also the general context of the city that does not seem to evolve through the years.

ment, *Political patronage and corruption* and *Environmental pollution*. Thus, it interesting to test the ability of dummy covariates approach to detect such small variations.

Instead, we observe a systematic displacement of almost all the uncertainty parameter estimates (π) towards the left side[17], and often for a substantial amount; then, all the 2006 respondents include more uncertainty in their ranking with respect to 2004. Given the constancy of the conditions, we may deduce that people feel, as a whole, more uncertain about evaluation of the problems. Specifically, they do not change the order of *Public health shortcomings*, *Streets cleanness and waste disposal* and *Immigration*[18]; however, uncertainty towards these items increased by a very large amount. As a consequence, we infer that a dummy covariate related to *Time* should be significant in these cases.

Specifically, we join the two surveys to get a unique sample of n=773 observations, that are quite homogeneous but *Time*, and we will study the *Environmental pollution* item that presents a low concern (the average rank is 6.635) and whose *CUB*(0,0) model shows an acceptable fit (*Diss*=0.102), as confirmed by Figure 6.

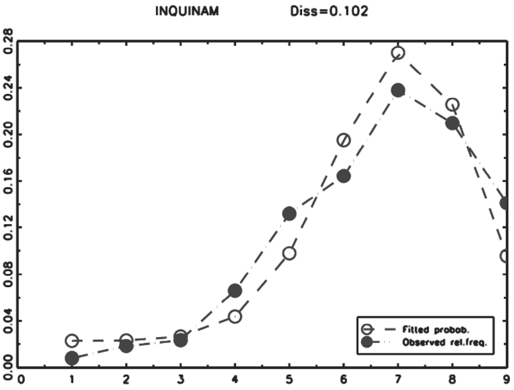

Figure 6 – Frequency and probability distributions of ranks.

Then, in Table 9, we show estimated *CUB* models for the *Environmental pollution* item: we get a significant improvement in likelihood measures as long as we include the covariates gender, ln(age)[19] and a dummy covariate for *Time* (=0 for 2004, =1 for 2006).

[17] The only exception to this pattern is given by the *TRAFFI* variable whose uncertainty parameter moves on the right side of the parameter space.

[18] Notice that the 2006 survey was collected some months before the significant upsurge of the "waste disposal" problem in Naples, at the beginning of 2007.

[19] We transform age by logarithm to reduce the variability of this covariate (in our data set, the age ranges from 18.0 to 57.8 years). As a consequence, we get a variance-covariance matrix of the estimators about 500 times more stable (as measured by the condition number); this transformation does not modify the other parameters values but improved their significance. A similar effect may be obtained if we consider the deviations of the age from its average. Instead, we have not found relevant covariates for uncertainty: in fact, it spreads over the support by a very limited amount (2%).

TABLE 9

Estimated CUB models for Environmental pollution

Models	$\hat{\pi}$	$\hat{\xi}$	Log-lik	AIC
CUB(0,0)	$\hat{\pi} = 0.796$ (0.030)	$\hat{\xi} = 0.258$ (0.008)	-1483.315	2970.630
CUB(0,1)	$\hat{\pi} = 0.804$ (0.029)	$\hat{\gamma}_0 = -0.857$ (0.049)	-1465.270	2936.541
Gender		$\hat{\gamma}_1 = -0.487$ (0.083)		
CUB(0,2)	$\hat{\pi} = 0.831$ (0.027)	$\hat{\gamma}_0 = -3.163$ (0.375)	-1447.139	2902.277
Gender		$\hat{\gamma}_1 = -0.345$ (0.082)		
ln(Age)		$\hat{\gamma}_2 = 0.710$ (0.113)		
CUB(0,3)	$\hat{\pi} = 0.830$ (0.027)	$\hat{\gamma}_0 = -3.044$ (0.381)	-1445.075	2900.151
Gender		$\hat{\gamma}_1 = -0.347$ (0.082)		
ln(Age)		$\hat{\gamma}_2 = 0.697$ (0.114)		
Time		$\hat{\psi} = -0.154$ (0.076)		

We observe that men are more concerned than women: indeed, moving from males (Gender=0) to females (Gender=1) the coefficient of this variable shows a decrease in concern; instead, with increasing age, people are more worried about this item.

Finally, the dummy covariate *Time* seems to show a limited but significant effect by reducing concern of respondent towards this item. This interpretation is consistent with the general analysis of section 3 since the parameter $\psi < 0$ suggests that, *ceteris paribus*, moving from 2004 survey (Time=0) to the 2006 survey (Time=1) we get a reduction in concern. In this regard, it is worth to observe that a standard analysis of the expressed average ranks of 6.531 and 6.732 of concerns in 2004 and 2006, respectively, cannot reject the hypothesis of the same expected ranks between the years[20].

It is interesting to compare the *CUB* models (with the same covariates) obtained in separate and aggregate surveys, as reported in Table 10 (where parameters estimates of γ_1, γ_2, ψ refers to covariates Gender, ln(Age) and *Time*, respectively). It turns out that parameters estimates are quite stable within the years and thus the dummy variable related to *Time* adds new information to the statistical interpretation.

[20] This conclusion is based on a *t*-test of $t_c = -0.054$, which is supported by the large sizes of the samples we are comparing.

TABLE 10

Comparison of CUB models for separate and aggregate surveys

Parameters	2004	2006	2004 & 2006
$\hat{\pi}$	0.868 (0.035)	0.792 (0.040)	0.830 (0.027)
$\hat{\gamma}_0$	-2.889 (0.542)	-3.357 (0.538)	-3.044 (0.381)
$\hat{\gamma}_1$	-0.325 (0.111)	-0.378 (0.124)	-0.347 (0.082)
$\hat{\gamma}_2$	0.649 (0.163)	0.745 (0.163)	0.697 (0.114)
$\hat{\psi}$			-0.154 (0.076)
Sample size	354	419	773
AIC/n	3.6861	3.8164	3.7518

If we desire to apply the discussion of section 3, we need to isolate the effect of the Gender and Time covariates. However, given the nature of the covariates (two of them are dichotomous while Age is continuous), it is possible to plot their combined effects both for the ξ parameter (a direct measure of concern for the item) and for the expected rank derived by the estimated models (an inverse measure related to a continuous *proxy* of the peoples' concern), as shown in Figures 7 and 8, respectively.

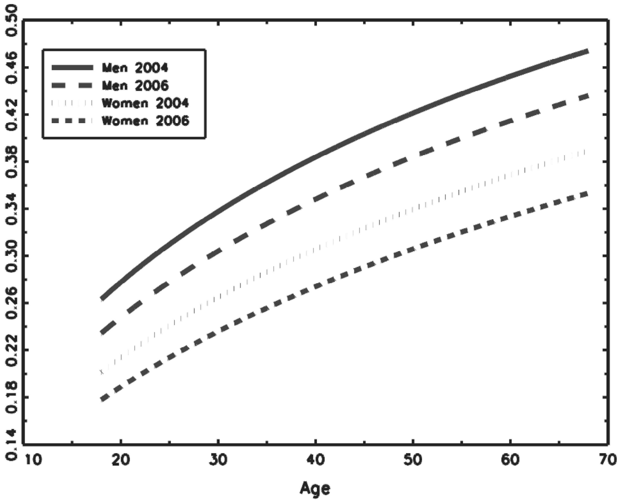

Figure 7 – Feeling parameter, given Gender and Time, for varying Age.

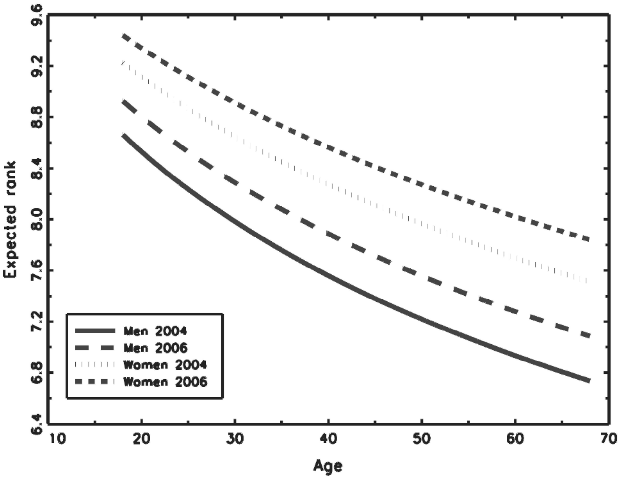

Figure 8 – Expected rank, given Gender and Time, for varying Age.

6. CONCLUDING REMARKS

In this paper, we have discussed interpretations and statistical implications of the use of dummy covariates in *CUB* models by investigating the relationship among the parameters, the feeling and uncertainty components and the expected response.

Direct relationships among dummy parameters and location and skewness measures of the implied distributions of ordinal variables have been derived. These results help the interpretation of estimated *CUB* models, fitted by one or two samples. An extensive simulation to measure the power discrimination among sub-populations by using dummy covariates has confirmed a good performance of the approach when the selection of different groups is a relevant issue.

Finally, for some real data sets, we have shown how a careful discussion about the placement of the estimates on the parametric space may be fruitfully exploited for the subsequent estimation of a *CUB* model with a dummy covariate.

Department of Statistical Sciences MARIA IANNARIO
University of Naples Federico II

REFERENCES

A. AGRESTI (2002), *Categorical data analysis*, 2nd edition, J. Wiley & Sons, New York.

R.D. BOCK, I. MOUSTAKI (2007), *Item Response Theory in a General Framework*, "Handbook of Statistics", 26, North-Holland, Amsterdam, 469-513.

S. CAGNONE, A. GARDINI, S. MIGNANI (2004), *New developments of latent variable models with ordinal data*, Atti della XLII Riunione Scientifica SIS, Bari, vol. I, 1-12.

A. D'ELIA (2000a), *Il meccanismo dei confronti appaiati nella modellistica per graduatorie: sviluppi statistici ed aspetti critici*, "Quaderni di Statistica", 2, 173-203.

A. D'ELIA (2003), *A mixture models with covariates for ranks data: some inferential developments*, "Quaderni di Statistica", 5, 1-25.

A. D'ELIA, E. MAURIELLO, N. SITZIA (2001), *Uno studio sulla scelta dei colori: esperienze e metodi statistici*, Atti del Convegno nazionale su Matematica, Formazione Scientifica e Nuove Tecnologie, Montevarchi, 85-96.

A. D'ELIA, D. PICCOLO (2005a), *A mixture model for preference data analysis*, "Computational Statistics & Data Analysis", 49, 917-934.

A. D'ELIA, D. PICCOLO (2005b), *Uno studio sulla percezione delle emergenze metropolitane: un approccio modellistico*, "Quaderni di Statistica", 7, 121-161.

A. J. DOBSON (1990), *An introduction to generalized linear models*, Chapman & Hall, London.

M. IANNARIO (2009), *A note on the identifiability of a mixture model for ordinal data*, submitted.

M. IANNARIO (2007), *A statistical approach for modelling Urban Audit Perception Surveys*, "Quaderni di Statistica", 9, 149-172.

M. IANNARIO, D. PICCOLO (2009), *A new statistical model for the analysis of Customer Satisfaction*, "Quality Technology and Quantitative Management", forthcoming.

V. E. JOHNSON, J. H. ALBERT (1999), *Ordinal data modeling*, Springer, New York.

G. KING, M. TOMZ, J. WITTENBERG, (2000), *Making the most of statistical analyses: improving interpretation and presentation*, "American Journal of Political Science", 44, 341-355.

C. J. LLOYD (1999), *Statistical Analysis of categorical data*, J. Wiley & Sons, New York.

J.I. MARDEN (1995), *Analyzing and modelling rank data*, Chapman & Hall, London.

P. MCCULLAGH (1980), *Regression models for ordinal data (with discussion)*, "Journal of the Royal Statistical Society", Series B, 42, 109-142.

P. MCCULLAGH, J. A. NELDER (1998), *Generalized linear models*, 2nd edition, Chapman & Hall, London.

I. MOUSTAKI (2000), *A latent variable model for ordinal data*, "Applied Psychological Measurement", 24, 211-223.

I. MOUSTAKI (2003), *A general class of latent variable model for ordinal manifest variables with covariate effects on the manifest and latent variables*, "British Journal of Mathematical and Statistical Psychology", 56, 337-357.

I. MOUSTAKI, M. KNOTT (2000), *Generalized latent trait models*, "Psychometrika", 65, 391-411.

J. A. NELDER, R. W. M. WEDDERBURN (1972), *Generalized linear models*, "Journal of the Royal Statistical Society", Series A, 135, 370-384.

D. PICCOLO (2003), *On the moments of a mixture of uniform and shifted binomial random variables*, "Quaderni di Statistica", 5, 85-104.

D. PICCOLO (2006), *Observed information matrix for MUB models*, "Quaderni di Statistica", 8, 33-78.

D. PICCOLO, A. D'ELIA (2008), *A new approach for modelling consumers' preferences*, "Food and Quality Preference", 19, 247-259.

D. PICCOLO, M. IANNARIO (2008), *A package in R for CUB models inference*, Version 1.1, available at http://www.dipstat.unina.it

D. A.POWER, Y. XIE (2000), *Statistical methods for categorical data analysis*, Academic Press, London.

J. S. SIMONOFF (2003), *Analyzing categorical data*, Springer, New York.

SUMMARY

Dummy covariates in CUB models

In this paper we discuss the use of dummy variables as sensible covariates in a class of statistical models which aim at explaining the subjects' preferences with respect to several items. After a brief introduction to CUB models, the work considers statistical interpretations of dummy covariates. Then, a simulation study is performed to evaluate the power discrimination of an asymptotic test among sub-populations. Some empirical evidences and concluding remarks end the paper.