BANCA D'ITALIA

Area Ricerca Economica e Relazioni Internazionali

Una classe di modelli per valutazioni e preferenze: analisi statistiche ed esperienze reali

Domenico Piccolo

Dipartimento di Scienze Statistiche

Università degli Studi di Napoli Federico II

20 Aprile 2009

domenico.piccolo@unina.it

Outline

- **I. Introduction**
  - Ordinale data modelling
  - Perception and evaluation

Outline

- ***I. Introduction***
  - Ordinale data modelling
  - Perception and evaluation
- ***II. CUB models: foundations, interpretation and inference***
  - A mixture random variable for ordinal data models
  - Main characteristics of **CUB** models (without and with covariates)
  - Sample data, log-likelihood function and E-M algorithm
  - Testing and fitting measures for **CUB** models
  - Fields of applications and real experiences

Outline

- **_I. Introduction_**
  - Ordinale data modelling
  - Perception and evaluation
- **_II. CUB models: foundations, interpretation and inference_**
  - A mixture random variable for ordinal data models
  - Main characteristics of **CUB** models (without and with covariates)
  - Sample data, log-likelihood function and E-M algorithm
  - Testing and fitting measures for **CUB** models
  - Fields of applications and real experiences
- **_III. Experiences on real data with CUB models_**
  - Preference for cities where to live
  - Emergencies in a metropolitan area
  - Subiective survival probability to age 75 and 90 years

Outline

- **I. Introduction**
  - Ordinale data modelling
  - Perception and evaluation
- **II. CUB models: foundations, interpretation and inference**
  - A mixture random variable for ordinal data models
  - Main characteristics of **CUB** models (without and with covariates)
  - Sample data, log-likelihood function and E-M algorithm
  - Testing and fitting measures for **CUB** models
  - Fields of applications and real experiences
- **III. Experiences on real data with CUB models**
  - Preference for cities where to live
  - Emergencies in a metropolitan area
  - Subiective survival probability to age 75 and 90 years
- **IV. Further developments**
  - *Shelter choices* and extended **CUB** models
  - Generalized **CUB** models and *IRT* paradigm
  - Concluding remarks

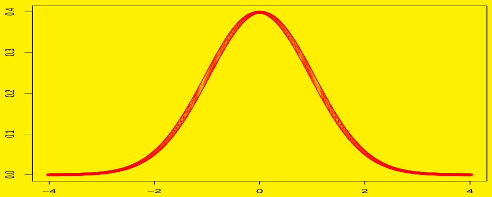
Part I

Introduction

From unobserved components to discrete choices

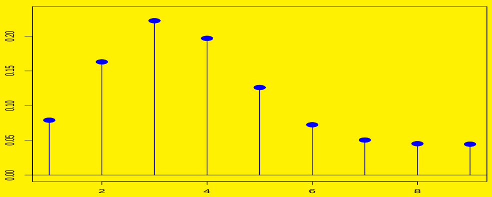
Perception

continuous latent variable



Evaluation

mixture of random variables



Discrete choices

finite discrete support

1	2	...	r	...	m

Perception

- *Perception* is a basic component in the formation of a concept, it is the ability to see, hear or understanding things and usually it expresses the awareness of something via the senses.

- Formally, perception is a cognitive act by which person interprets and organizes several sensations in order to identify a specific object/situation. Thus, perception of an object/service/item is a psychological process by which a subject synthesizes sensory data in forms that are meaningful for his/her conscience.

- Indeed, when we ask a person to answer a specific topic on a questionnaire, we are looking for his/her perception of the problem; specifically, we are asking to summarize his/her perception into a well defined category (qualitative, quantitative, mixed, verbal, and so on).

- Since this perception is a complex function of several causes (personal, family, environmental, social, and so on), the expressed survival probability is affected both by the real consideration of problems and by inherent uncertainty that accompanies any human decision.

Evaluation

- Evaluation can be described as the psychological process which a subject has to perform when he/she is requested to give a determination of merit regarding an item (the attributes of a service, a product or in general, any tangible or intangible object) using a certain ordinal scale.

- The mechanism governing individual choices between a set of possible alternative options has been widely studied by the latent variables theory.

- Sample surveys gather measures of satisfaction which are a manifest expression of respondents' psychological constructs. Simply, we need to transform a mental process into a discrete state in order to assign an evaluation referred to the graded scale proposed by the interviewer.

- Then, the psychological mechanism, by which the choice is made, is the result of a personal *feeling* for the object under judgement and an inherent *uncertainty* associated with the selection of the ordinal value of the response.

Latent components for expressing subjective probability

- Such psychological processes manifest themselves as the result of two main factors:
  - a *primary* component, generated by the sound impression of the respondent, related to *awareness and full understanding* of the problem, personal or previous experience, group partnership, and so on;
  - a *secondary* component, generated by the *intrinsic uncertainty* of the final choice. This may be due to the amount of time devoted to the answer, the use of limited set of information, nature of the chosen scale, partial understanding of the item, lack of self-confidence, laziness, apathy, and so on.

- Then, the responses are realizations of a stochastic phenomenon and it should be analyzed with statistical methods that focus on the generating data process.

- *Feeling (agreement)* is usually related to subjects' motivations whereas *uncertainty (fuzziness)* mostly depends on circumstances that surround the process of judging.

Part II

CUB models: foundations, interpretation and inference

Ordinal variables and discrete choices

- ***Ordinal variables*** associate *integers* to discrete choices in several circumstances.

Ranking

Numbers convey the `location/preference` of the “object” in a given ordered list

Rating

Numbers convey the `level` of a “stimulus” as perceived by the respondent

Qualitative assessment

Numbers convey a `qualitative judgment` about a situation as perceived by the respondent

Ranking and *rating*: similarities and differences

- ***SIMILARITIES***
  - Response is the expression of a continuous latent variable, which is compelled to be expressed by means of discrete values.
  - Response is the result of sequential or paired comparisons.
  - $\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$
- ***DIFFERENCES***
  - Ranking analysis of m “objects”produces a permutation of the first m integers (a vector), that is the realization of a *multivariate random variable* of dimensions $(m-1)$.
  - Rating analysis of an “object” on a scale $[1, m]$ produces a number (a unique integer), that is the realization of a discrete *univariate random variable* defined on the support $\{1, 2, \ldots, m\}$
  - $\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$
- ***RELATIONSHIP AMONG THEM***
  - ***Ranking*** will be analysed as a marginal distribution of the given “object”.
  - Any ***ranking*** always includes a ***rating***, not vice versa.

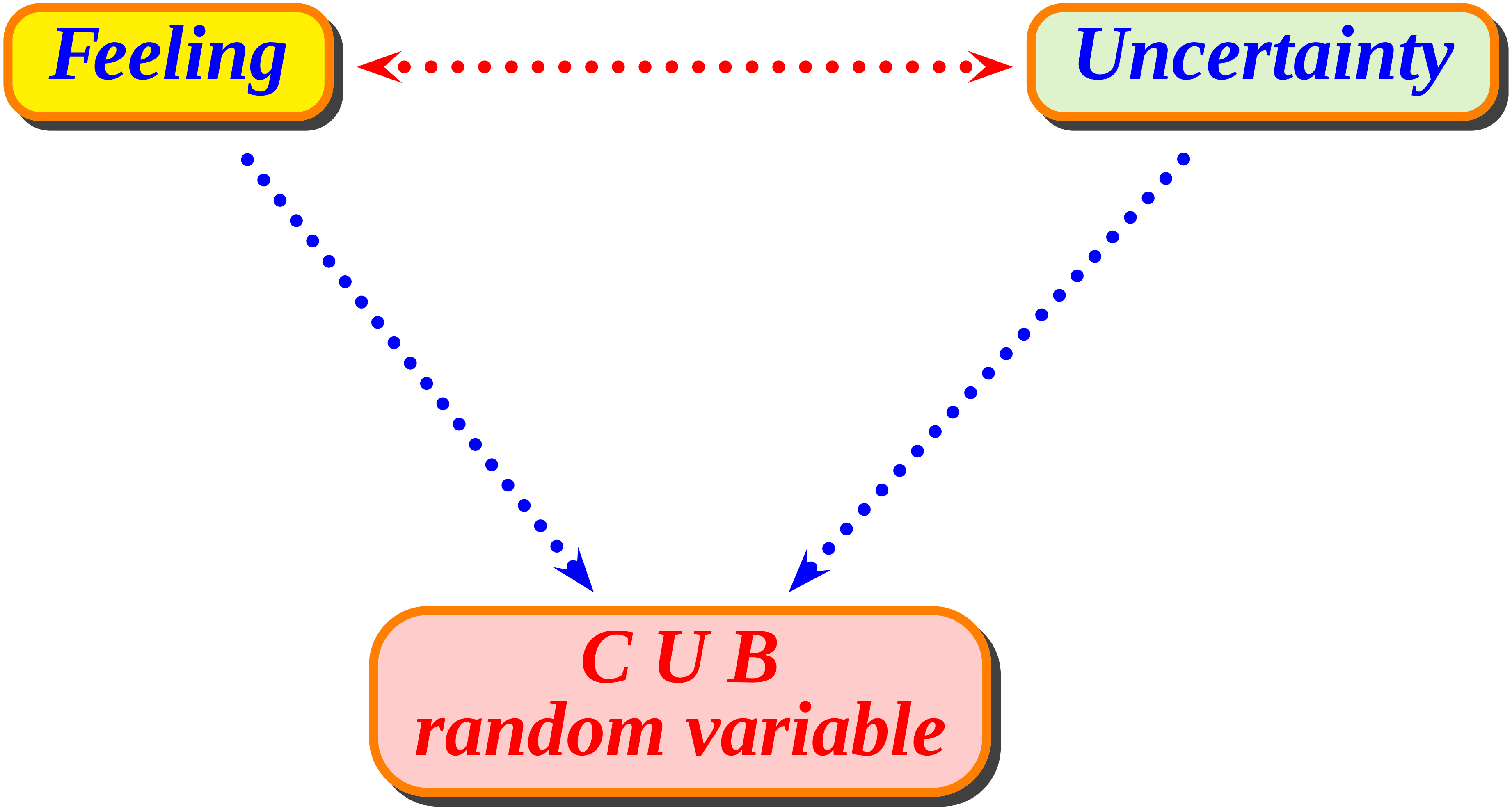
Why a *new* model for ordinal data?

All models are wrong ...... but some are useful *(George P. Box)*

- ***Any statistical model should combine logical facts (about phenomena to interpret) with empirical evidences (generated by available data), in the framework of a parsimonious and consistent specification.***

- Main motivations for a new model:
  1. Logical foundations derived by psychological behaviour of respondents.
  2. Estimate, interpret and forecast explicitly $Pr\,(R = r)$ for an ordinal response without preliminary transformation.
  3. Parameters immediately related to unobservable components.
  4. Flexible with data and consistent with interpretation, at least as previous models.

Unobserved components of a discrete choices



Feeling

Uncertainty

C U B
random variable

Characteristics of the feeling

- ***Feeling*** is the result of a continuous random variable that becomes a discrete one, since we compel the subject to express the preferences into m prefixed bins.

- The judgement process is intrinsically continuous and, since the basic feeling depends on several causes, it could be thought to follow a Gaussian distribution.

- Indeed, a *latent variable approach* for the analysis of ordinal data assumes that the observations are generated by an unobserved continuous variable (say R^*) normally distributed, and defining a correspondence with a discrete ordinal random variable R by means of ordered threshold parameters to be estimated.

$$\begin{array}{lcl}
-\infty < R^* \leq \tau_1 & & R = 1 \\
\tau_1 < R^* \leq \tau_2 & & R = 2 \\
\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \Longleftrightarrow & \ldots\ldots\ldots\ldots \\
\tau_{m-2} < R^* \leq \tau_{m-1} & & R = m-1 \\
\tau_{m-1} < R^* < +\infty & & R = m
\end{array}$$

Rationale for modelling the first component

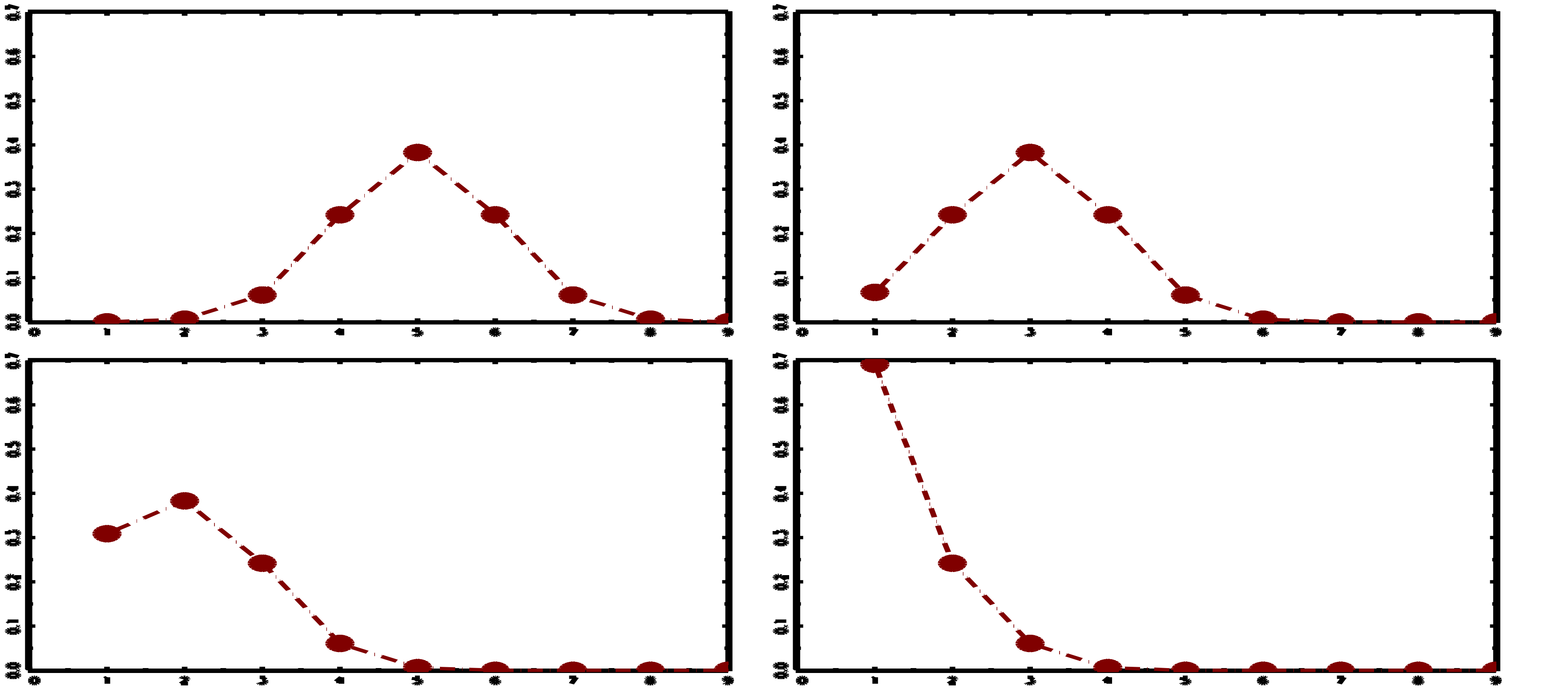
➤ A *shifted Binomial* random variable is introduced for the perception component and legitimated by two arguments:

- From a ***statistical*** point of view, a standard Binomial distribution is generated by adding several independent and identically distributed Bernoulli choices. Then, we may think that when a subject chooses a rating (among m possible categories) he/she excludes the others by a pairwise comparison.
- From a ***heuristic*** point of view, the shifted Binomial distribution is able to map a continuous latent variable (characterized by a single mode distribution: Normal, Student-t, logistic, etc.) into a discrete set of values $\{1, 2, \ldots, m\}$; this happens with just one parameter. The shape of the resulting distribution depends on the way the cut-points are originally chosen. This fact adds further flexibility in modelling the observations since it allows for very different mode location, flatness and skewness.

From a Normal feeling to a shifted Binomial rating

➤ Following this idea, a suitable model for achieving the mapping of the unobserved continuous variable R^* into a discrete random variable defined on the support $r = 1, 2, \ldots, m$, may be the ***shifted Binomial*** distribution.

➤ Indeed, the next figure shows how, by *varying the ordered thresholds*, a standard Normal random variable can be made discrete, according to features (mode, skewness, etc.) that are well fitted by a shifted Binomial random variable.



Characteristics of the uncertainty

➤ ***Uncertainty*** is a vague component that needs some clarification.

➤ Uncertainty ***is not*** the stochastic component related to the sampling experiment (so that different people generates different rankings), but is the personal component intrinsically related to the choice mechanism.

➤ Uncertainty ***is*** the result of convergent and related factors:

- *Knowledge/Ignorance* of the problems and/or the characteristic of the objects.
- *Personal interest/Engagement* in similar activities, objects, opinions, etc.
- *Time spent* for assuming the decision
- *Laziness/Apathy* in the selection mechanism
- ....

Uncertainty and randomness

➤ An important point concerns the very nature of the second component and it should be stressed in order to reduce some ambiguity:

- We speak of ***uncertainty*** with reference to the *subjective* respondents' indecision. This aspect is related to the very nature of human choices. In our modelling approach we will explicitly consider this structural component.

- We speak of ***randomness*** with reference to modality of collecting data from a subset of a given population. This aspect is related to sampling selection, measurement errors and limited knowledge. In any modelling approach this issue is considered by using a random variable paradigm.

➤ We face with the first issue by the structure of the model we are going to introduce whereas the second issue is intrinsic if we model sample data by means of a probability model.

Extreme uncertainty

- In case the subject shows indifference (=*equipreference*) towards a given item, it seems appropriate to model choice by means of a discrete Uniform random variable U with a probability mass function defined by:

$$Pr\,(U = r) = \frac{1}{m}, \quad r = 1, 2, \ldots, m\,.$$

- In this way, the choice is the result of a complete randomized mechanism since any item has the same probability of receiving any rank $r \in [1, m]$.

- From a probability point of view, the discrete Uniform random variable maximizes the entropy, among all the discrete distributions with finite support $\{1, 2, \ldots, m\}$, for a fixed m.

- However, we are not stating that people answer questions in a purely random manner. Instead, we are saying that the uncertainty affecting any choice can, at worst, be constituted by a situation where no category prevails over the others, and that is the case of a Uniform distribution.

- In fact, we are choosing the ***discrete Uniform*** distribution as a building block for modeling the uncertainty in the ordinal modeling.

Specification of a CUB model

Formally, **CUB** models are specified by considering the ordinal response r as a realization of a discrete random variable R defined on the support $\{r = 1, 2, \dots, m\}$. For a given $m > 3$, the random variable R is a mixture of Uniform and shifted Binomial random variables and its probability mass function is given by:

$$\boxed{Pr(R = r) = \pi \underbrace{\left[\binom{m-1}{r-1}(1-\xi)^{r-1}\xi^{m-r}\right]}_{\textit{feeling}} + (1-\pi) \underbrace{\left[\frac{1}{m}\right]}_{\textit{uncertainty}}, \quad r = 1, 2, \dots, m;}$$

where $\pi \in (0, 1]$ and $\xi \in [0, 1]$. Thus, the parametric space is the (left open) unit square:

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

➤ From a theoretical point of view, Iannario (2009) proved that **CUB** models are fully identifiable for any $m > 3$.

Meaning of a mixture distribution

- It is important to make clear that we are not conjecturing that the population is composed of two subgroups of respondents, each behaving according to one of the two above-mentioned probability distributions.

- Indeed, each respondent acts with a ***propensity*** to adhere to a thoughtful and to a completely uncertain choice, measured by (π) and $(1-\pi)$, respectively.

- As a consequence, $(1-\pi)$ is a *measure of uncertainty* whereas $(1-\xi)$ may be interpreted as a *measure of adhesion* to the proposed choice.

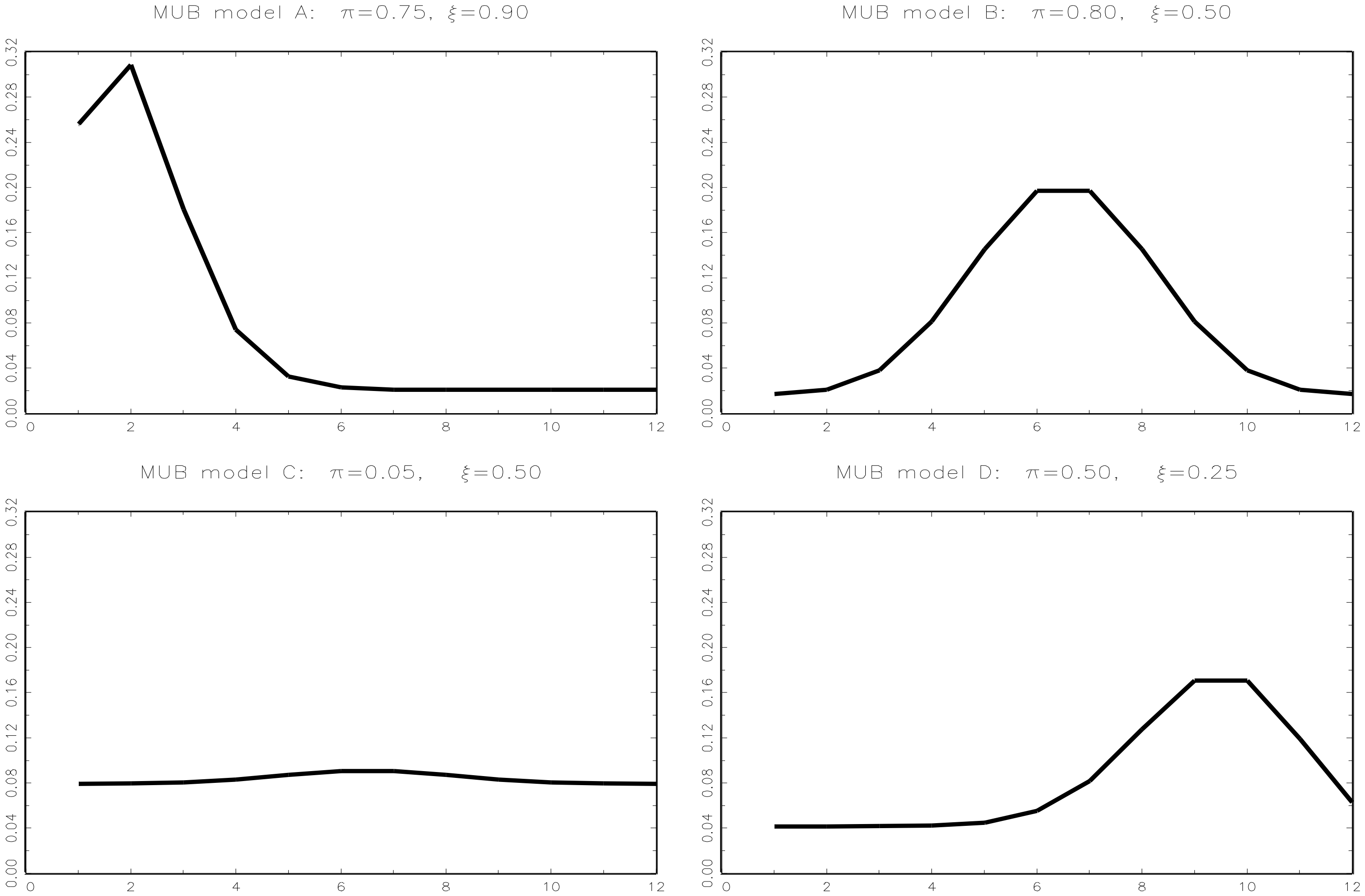
Interpreting parameters of a CUB model

- The ξ parameter is related to the degree of ***feeling*** of respondent in the following way:
  - In *rating* analysis –where evaluation is higher for preferred items– agreement towards the "object" increases with $1 - \xi$.
  - In *ranking* analysis –where preferred object is ranked first– agreement towards the "object" increases with ξ.

- The π parameter is inversely related to the degree of ***uncertainty*** expressed by respondents, thus uncertainty increases with $1 - \pi$. More specifically:
  - If $\pi \longrightarrow 0$, respondent manifests a great *propensity* towards an extreme indecision in the choice.
  - If $\pi \longrightarrow 1$, respondent manifests a minimum *propensity* towards an extreme indecision and its choice is more resolute and determined mostly by a feeling attitude.

CUB model distributions are very flexible

- The proposed mixture distribution is extremely flexible and, depending on the parameters, it is able to assume different shapes: symmetric or extremely skewed, rather flat or with definite mode.

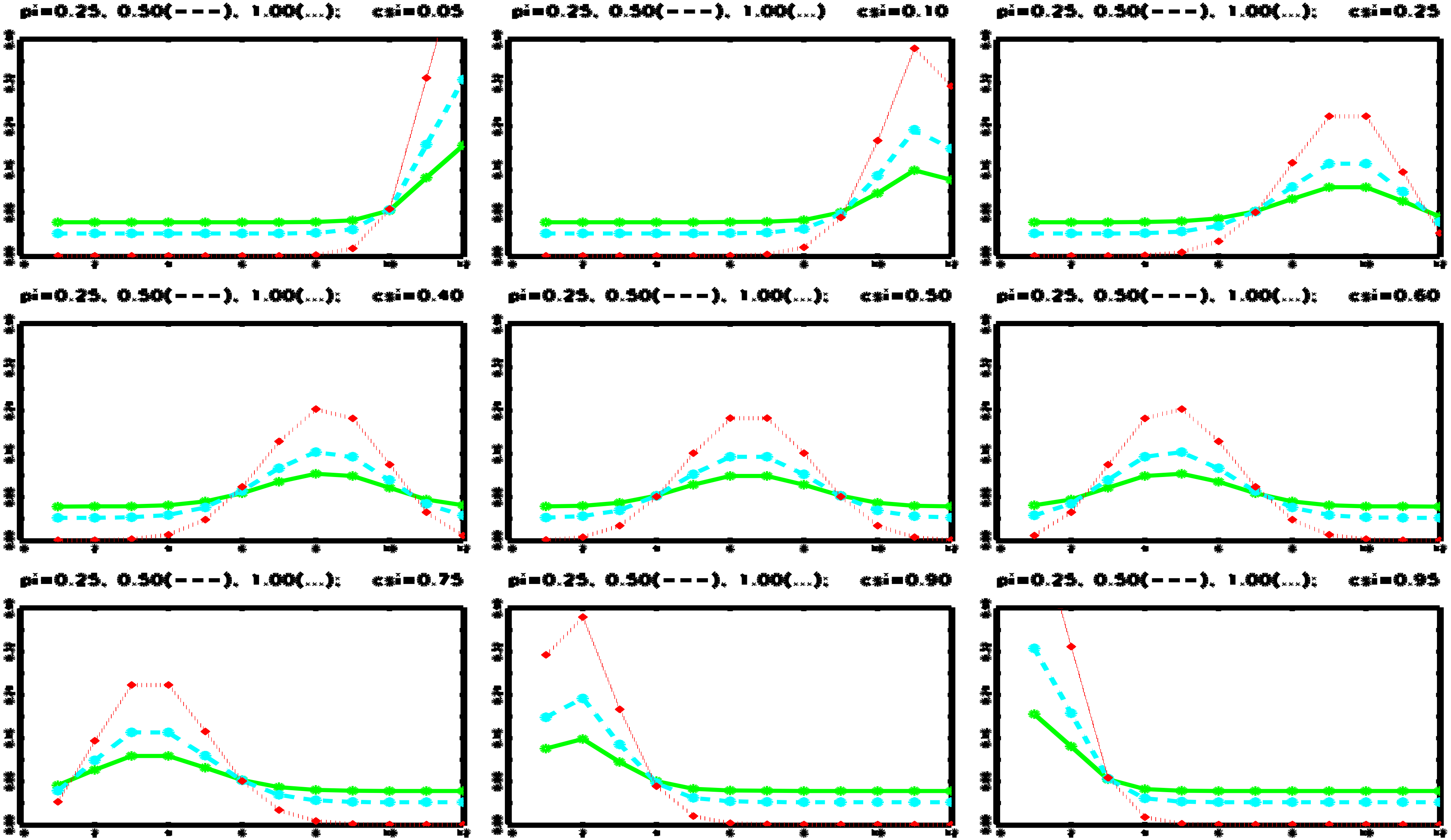

MUB model A: π=0.75, ξ=0.90

MUB model B: π=0.80, ξ=0.50

MUB model C: π=0.05, ξ=0.50

MUB model D: π=0.50, ξ=0.25

CUB distributions for varying π and ξ



pi=0.25, 0.50(---), 1.00(...); csi=0.05

pi=0.25, 0.50(---), 1.00(...) csi=0.10

pi=0.25, 0.50(---), 1.00(...); csi=0.25

pi=0.25, 0.50(---), 1.00(...); csi=0.40

pi=0.25, 0.50(---), 1.00(...); csi=0.50

pi=0.25, 0.50(---), 1.00(...); csi=0.60

pi=0.25, 0.50(---), 1.00(...); csi=0.75

pi=0.25, 0.50(---), 1.00(...); csi=0.90

pi=0.25, 0.50(---), 1.00(...); csi=0.95

Main characteristics of the R distribution

- Some features of **CUB** distributions are more specific for our discussion:
  - It admits a mode at any value of the support $\{1, 2, \ldots, m\}$.
  - The weight of the tails is a function of $(1 - \pi)/m$.
  - It is a symmetric random variable only if $\xi = \frac{1}{2}$.
  - It is a ***reversible*** random variable, since:

$$R \sim \textbf{CUB}\ (\pi, \xi) \Longrightarrow (m + 1 - R) \sim \textbf{CUB}\ (\pi, 1 - \xi)\,.$$

  - It is consistent with the hypothesis that the population is made by two sub-groups of raters (an informed/reflexive set and a more uninformed/instinctive one) and their relative ratio is $\pi/(1 - \pi)$.
  - It emulates many theoretical distributions:
    - ○ A ***Uniform*** distribution, if $\pi = 0$;
    - ○ A ***Shifted Binomial*** distribution, if $\pi = 1$;
    - ○ A ***IHG*** distribution, if $\pi \to 1$ *and* ξ tends to 0 or 1;
    - ○ A ***Normal*** distribution, if $\xi \to \frac{1}{2}$ *and* $m \to \infty$;

Expectation of CUB random variable

- The expected value of the random variable R is given by:

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

- Notice that expectation moves towards the central value of the support depending on the sign of $(\frac{1}{2}-\xi)$. Thus, we expect higher (smaller) mean values when $\xi \to 0$ ($\xi \to 1$, respectively). It is evident that skewness of the distribution is governed by $(\frac{1}{2}-\xi)$, and positive (negative) asymmetry occurs when $\xi > \frac{1}{2}$ ($\xi < \frac{1}{2}$), respectively.

- Some word of caution is necessary since we are modelling qualitative (ordinal) phenomena, thus statistical indexes as expectation, mode, variance, and so on, are not immediately interpretable. However, although one should not use quantitative measure to synthesize the responses, these indexes may refer to continuous latent variables that corresponds to the declared qualitative assessment.

- This correspondence is not useful *per se* but may be of interest when comparing ordinal responses given by subgroups of subjects (different in time, space or circumstances).

A different parameterization

➤ From the formula of $\mathbb{E}\,(R)$ we realize that different parameter vectors $\boldsymbol{\theta} = (\pi,\, \xi)'$ may generate the same mean value; thus, for this class of models, it is not adequate to introduce a link among expectation and covariates.

➤ In **CUB** models we assume that uncertainty and perception parameters are related to covariates by a logistic function, that is by means of two *systematic components*:

$$\pi_i = \frac{1}{1 + e^{-\boldsymbol{y}_i\, \boldsymbol{\beta}}}\,; \qquad \xi_i = \frac{1}{1 + e^{-\boldsymbol{w}_i\, \boldsymbol{\gamma}}}\,, \qquad i = 1, 2, \ldots, n;$$

where $\boldsymbol{y}_i$ and $\boldsymbol{w}_i$ are the subjects' covariates for explaining π_i e ξ_i, respectively. In these cases, we will use the notation **CUB** (p, q) for denoting the number of covariates entering in the model for explaining feeling and uncertainty components, respectively. Of course, a **CUB** (0,0) model is a model without covariates.

➤ Indeed, any one-to-one function mapping real numbers into unit interval is adequate; however, in our experience, logistic function has been a convenient solution in any case.

CUB model with covariates

➤ Then, the general formulation of a **CUB** (p, q) model (with p covariates to explain uncertainty and q covariates to explain feeling) is expressed by:

$$Pr(R = r \mid \boldsymbol{y}_i;\, \boldsymbol{w}_i) = \pi_i \binom{m-1}{r-1} \xi_i^{m-r} (1-\xi_i)^{r-1} + (1-\pi_i)\left(\frac{1}{m}\right), \qquad r = 1, 2, \dots, m,$$

and two *systematic components*:

$$\pi_i = \frac{1}{1+e^{-\boldsymbol{y}_i\,\boldsymbol{\beta}}}\,; \qquad \xi_i = \frac{1}{1+e^{-\boldsymbol{w}_i\,\boldsymbol{\gamma}}}\,; \qquad i = 1, 2, \dots, n;$$

where $\boldsymbol{y}_i$ and $\boldsymbol{w}_i$ are the subjects' covariates for explaining π_i e ξ_i, respectively. Notice that this formalization allows that the set of covariates used for explaining the parameters may or may not present some overlapping.

➤ The probability distributions (***U***niform and shifted ***B***inomial) included in the mixture and the presence of ***C***ovariates justify the acronym **CUB** .

Notation for CUB model with covariates

Models	*Covariates*	*Parameters*	*Parameter spaces*
CUB $(0,0)$	no covariates	$\boldsymbol{\theta} = (\pi, \xi)'$	$(0,1] \times [0,1]$
CUB $(p,0)$	only for π	$\boldsymbol{\theta} = (\boldsymbol{\beta}', \xi)'$	$\mathbb{R}^{p+1} \times [0,1]$
CUB $(0,q)$	only for ξ	$\boldsymbol{\theta} = (\pi, \boldsymbol{\gamma}')'$	$(0,1] \times \mathbb{R}^{q+1}$
CUB (p,q)	both for π and ξ	$\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\gamma}')'$	$\mathbb{R}^{p+q+2}$

Parsimony of CUB models, without and with covariates

- Although we move within the *logic* of the GLM models, the **CUB** random variable ***does not belong*** to the exponential family; as a consequence, there is no linear link function between expectation and parameters.

- With respect to the classical GLM approach (where proportional, adjacent or continuation ratio probabilities are introduced for ordinal data), **CUB** models offer a straightforward relationship between a probability statement for ordinal answers and subjects' covariates by means of a monotone function (logistic function, in most cases).

- Moreover, although latent variables are conceptually necessary in order to specify the nature of the mixture components, the inferential procedures are not based upon the knowledge (or estimation) of cut-points.

- As a consequence, when a **CUB** model turns out to be adequate in fitting data, it is usually more parsimonious with respect to models derived by the GLM approach.

Sample data, log-likelihood function and E-M algorithm

- We suppose to collect a sample of ordinal data $\boldsymbol{r} = (r_1, r_2, \dots, r_n)'$, and for each respondent we have information which are able to characterize both perception and uncertainty, respectively.

- Such information are collected by the matrices:

$$\boldsymbol{Y} = \begin{pmatrix} 1 & y_{11} & y_{12} & \dots & y_{1p} \\ 1 & y_{21} & y_{22} & \dots & y_{2p} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & y_{i1} & y_{i2} & \dots & y_{ip} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & y_{n1} & y_{n2} & \dots & y_{np} \end{pmatrix} ; \ \boldsymbol{W} = \begin{pmatrix} 1 & w_{11} & w_{12} & \dots & w_{1q} \\ 1 & w_{21} & w_{22} & \dots & w_{2q} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & w_{i1} & w_{i2} & \dots & w_{iq} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & w_{n1} & w_{n2} & \dots & w_{nq} \end{pmatrix} .$$

- For compactness, we introduce $Y_0 \equiv 1$ e $W_0 \equiv 1$ and they specify the constant baselines of the model.

- Notice that our parameterization allows that information set contained in $\boldsymbol{Y}$ e $\boldsymbol{W}$ may be partially or completely overlapped.

Inference for a CUB model without covariates

➤ Given the observed frequencies vector $(n_1, n_2, \ldots, n_m)'$, where n_r is the frequency of $(R = r)$, and letting $p_r(\pi,\, \xi) = Pr(R = r \mid \pi,\, \xi)$, $r = 1, 2, \ldots, m$, the *log-likelihood function* for the CUB model is:

$$\begin{aligned}\log L(\pi,\, \xi) &= \sum_{r=1}^{m} n_r \, \log\{p_r(\pi,\, \xi)\} \\ &= \sum_{r=1}^{m} n_r \, \log\left[\pi \binom{m-1}{r-1} (1-\xi)^{r-1} \xi^{m-r} + (1-\pi)\, \frac{1}{m}\right]\end{aligned}$$

➤ As it is common for mixture models, the ML estimates of π and ξ can be obtained by means of the E-M algorithm (D'Elia and Piccolo, 2005).

➤ It is also possible to derive the asymptotic standard errors of the ML estimates, and thus to test their significance.

Inference for a CUB model with covariates

➤ A CUB model with covariates is defined for any $i = 1, 2, \ldots, n$ by the following probability distribution:

$$Pr\left(R = r_i \mid \boldsymbol{\beta}, \boldsymbol{\gamma}\right) = \frac{1}{1 + e^{-\boldsymbol{y}_i \boldsymbol{\beta}}} \left[\binom{m-1}{r_i - 1} \frac{\left(e^{-\boldsymbol{w}_i \boldsymbol{\gamma}}\right)^{r_i - 1}}{\left(1 + e^{-\boldsymbol{w}_i \boldsymbol{\gamma}}\right)^{m-1}} - \frac{1}{m} \right] + \frac{1}{m} \,.$$

➤ The related log-likelihood function is:

$$\log\, L\left(\theta\right) = \sum_{i=1}^{n} \log \left[\frac{1}{1 + e^{-\boldsymbol{y}_i \boldsymbol{\beta}}} \left\{ \binom{m-1}{r_i - 1} \frac{e^{(-\boldsymbol{w}_i \boldsymbol{\gamma})(r_i - 1)}}{\left(1 + e^{-\boldsymbol{w}_i \boldsymbol{\gamma}}\right)^{m-1}} - \frac{1}{m} \right\} + \frac{1}{m} \right] .$$

➤ Then, for maximum likelihood (ML) inference, standard results apply, although we are currently using E-M algorithm for estimation and observed information matrix for asymptotic inference: Piccolo (2006).

➤ We are currently working for improving these procedures: E-M algorithm converges slowly but almost surely, whereas scoring algorithm is incredible faster but requires initial values very accurate. Thus, a two-steps procedure is the key solution for an optimal procedure given consistent preliminary estimators.

The E-M algorithm for a CUB (p,q) model

Steps	***E-M algorithm for maximum likelihood estimation***
0	$\boldsymbol{\theta}^{(0)} = \left(\boldsymbol{\beta}'^{(0)}; \boldsymbol{\gamma}'^{(0)}\right)' = (0.1, \dots, 0.1;\ 0.1, \dots, 0.1)'$; $l^{(0)} = \ell\left(\boldsymbol{\theta}^{(0)}\right)$; $\epsilon = 10^{-6}$.
1	$\xi_i^{(k)} = \frac{1}{1+e^{-\boldsymbol{w}_i\boldsymbol{\gamma}^{(k)}}}$; $b\left(r_i; \boldsymbol{\gamma}^{(k)}\right) = \binom{m-1}{r_i-1} \frac{e^{-(r_i-1)\,\boldsymbol{w}_i\boldsymbol{\gamma}^{(k)}}}{\left(1+e^{-\boldsymbol{w}_i\boldsymbol{\gamma}^{(k)}}\right)^{m-1}}$, $i = 1, 2, \dots, n.$
2	$\pi_i^{(k)} = \frac{1}{1+e^{-\boldsymbol{y}_i\boldsymbol{\beta}^{(k)}}}$; $\tau\left(r_i; \boldsymbol{\theta}^{(k)}\right) = \left[1 + \frac{e^{-\boldsymbol{y}_i\boldsymbol{\beta}^{(k)}}}{m\, b\left(r_i; \xi^{(k)}\right)}\right]^{-1}$, $i = 1, 2, \dots, n.$
3	$Q_1\left(\boldsymbol{\beta}^{(k)}\right) = -\sum_{i=1}^{n} \left\{\log\left(1+e^{-\boldsymbol{y}_i\boldsymbol{\beta}^{(k)}}\right) - \left(1-\tau\left(r_i; \boldsymbol{\theta}^{(k)}\right)\right) \boldsymbol{y}_i\boldsymbol{\beta}^{(k)}\right\}.$
4	$Q_2\left(\boldsymbol{\gamma}^{(k)}\right) = -\sum_{i=1}^{n} \tau\left(r_i; \boldsymbol{\theta}^{(k)}\right) \left\{(r_i-1)\,\boldsymbol{w}_i\boldsymbol{\gamma}^{(k)} + (m-1)\log\left[1+e^{-\boldsymbol{w}_i\boldsymbol{\gamma}^{(k)}}\right]\right\}.$
5	$\boldsymbol{\beta}^{(k+1)} = argmax_{\boldsymbol{\beta}}\ Q_1\left(\boldsymbol{\beta}^{(k)}\right)$; $\boldsymbol{\gamma}^{(k+1)} = argmax_{\boldsymbol{\gamma}}\ Q_2\left(\boldsymbol{\gamma}^{(k)}\right)$.
6	$\boldsymbol{\theta}^{(k+1)} = \left(\boldsymbol{\beta}'^{(k+1)}, \boldsymbol{\gamma}'^{(k+1)}\right)'.$
7	$l^{(k+1)} = \ell\left(\boldsymbol{\theta}^{(k+1)}\right).$
8	$\begin{cases} \text{if } l^{(k+1)} - l^{(k)} \geq \epsilon, & k \to k+1; \text{go to } 1; \\ \text{if } l^{(k+1)} - l^{(k)} < \epsilon, & \hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^{(k+1)}; \text{stop}. \end{cases}$

Dissimilarity as an inverse fitting measure

➤ For this kind of data, traditional goodness-of-fit indexes can not be applied usefully. As a matter of fact, the χ^2 test detects significance values even though there is an "almost perfect fit". Thus, when no covariates are present, we prefer a descriptive measure.

➤ The normalized ***dissimilarity index*** $Diss$ is defined by:

$$\boxed{Diss = \frac{1}{2}\sum_{r=1}^{m} \mid f_r - p_r(\hat{\pi}, \hat{\xi}) \mid; \qquad 0 \leq Diss \leq 1.}$$

where f_r and $p_r(\hat{\pi}, \hat{\xi})$ are the observed relative frequencies and the estimated probabilities by **CUB** model, respectively.

➤ *The index $Diss$ measures the quota (=relative frequency) of subjects to move among the cells of the frequency distribution to reach a perfect fit.*

Likelihood based fitting measures

- Generally, we compare differences in deviances obtained by different models by standard asymptotic inference, as in the following table. Degrees of freedom (g) are obtained by the difference between the number of parameters in the corresponding models.

Comparisons	*Difference of deviances*	g
CUB $(p, 0)$ *versus* **CUB** $(0, 0)$	$2\ (\ell_{p0} - \ell_{00})$	p
CUB $(0, q)$ *versus* **CUB** $(0, 0)$	$2\ (\ell_{0q} - \ell_{00})$	q
CUB (p, q) *versus* **CUB** $(0, 0)$	$2\ (\ell_{pq} - \ell_{00})$	$p + q$

- A sort of pseudo-R^2 may be introduced if we compare maximized log-likelihood with the worst attainable (that is: $\ell(0) = -n\ \log(m)$, which is the log-likelihood of a Uniform discrete distribution), and we called it $ICON$:

$$\mathcal{ICON} = 1 + \frac{\ell(\hat{\boldsymbol{\theta}})/n}{\log(m)}\,.$$

- Of course, traditional $AIC, BIC, \ldots$ may add useful information, mostly in comparing non-nested models.

CUB models interpretation . . . . [1]

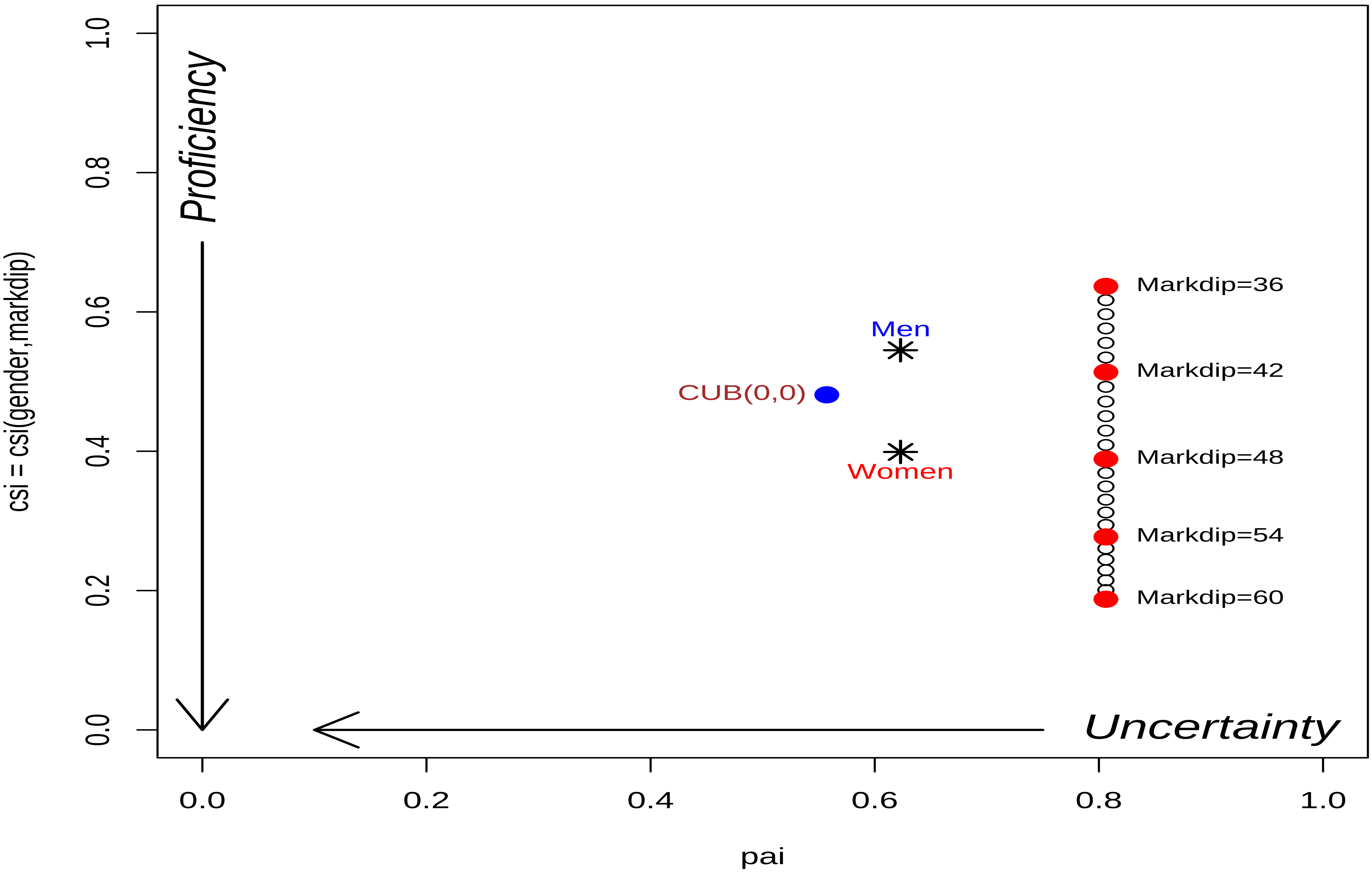
➤ The parameter values help to locate **CUB** models in the parametric space given by the unit square, and we will use this visualization as an interpretative tool.

➤ Since $1 - \pi$ quantifies the *propensity* of respondents to behave in accordance to a completely random choice, the more π is located to the right side of the unit square, the more respondents are inclined to give definite answers (uncertainty is low).

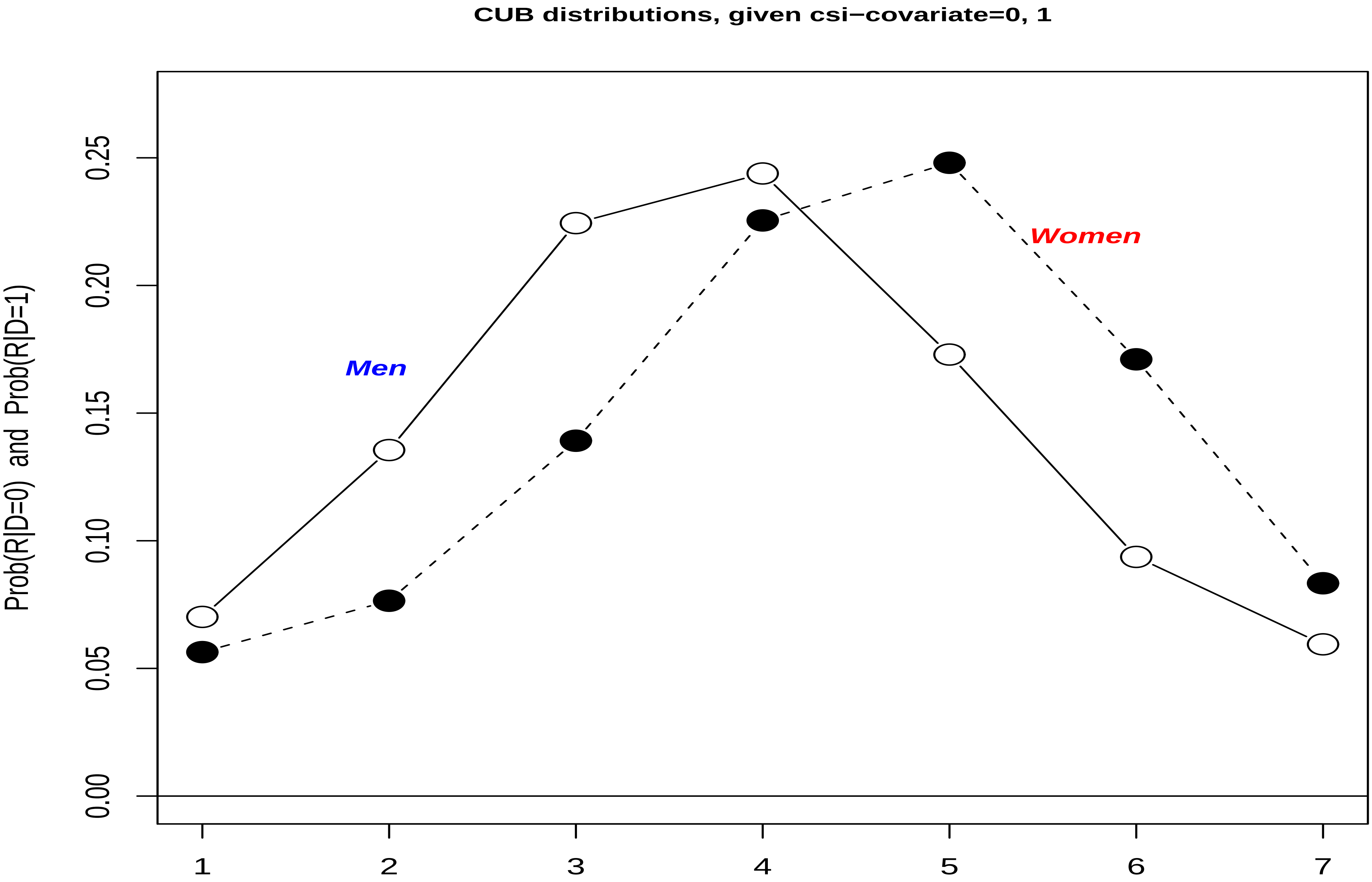
➤ Similarly, since $1 - \xi$ measures the *strength of feeling* of the subjects for a direct and positive evaluation of the object, the closer ξ is located to the border of the upper region of the unit square the less the item has been preferred.

➤ In this way, it is immediate to see how the introduction of covariates and/or the analysis of subroups modify the behavior of respondents, as in the following case studies.

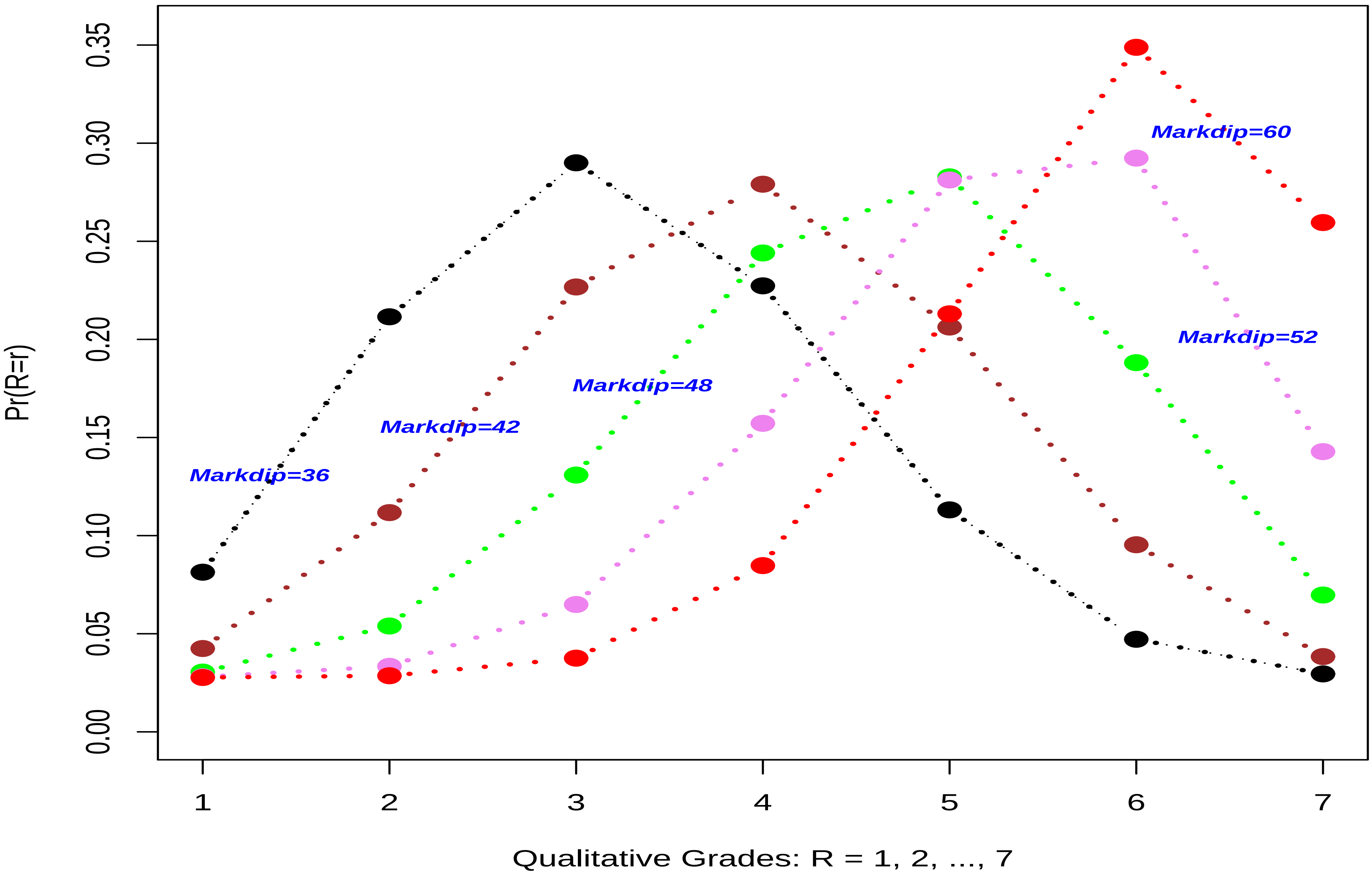
CUB models interpretation .... [2]



CUB models interpretation ....[3]



CUB models interpretation ....[4]



Main fields of application of CUB models

➤ Parsimonious and consistent specification of these models have been successfully applied in the following fields:

- ***Marketing researches***
- ***Psychological behavior***
- ***Sociological surveys***
- ***Policy analysis***
- ***Services evaluation***
- ***Sensometrics***
- ***Linguistic***
- ***Medicine***
- . . . .

Real experiences with CUB models (*rating* and *ranking*)

Preferences ⟹

- *Colors (young people, children, air force cadets)*
- ***Cities where to live***
- *Professions for students of Political Sciences graduates*
- *Olive oils*
- *South of Italy typical products*

Evaluations ⟹

- *Orientation services*
- *University teaching and structures*
- *Services for E-bay users*
- *Characteristics of transports to a metropolitan area*
- *Degree of preference for buying equo-solidal agricoltural products*
- *Quality of services in a protected area*
- *Customers' satisfaction of European consumers towards salmon*
- *Final degree of University graduates*

Perceptions ⟹

- ***Urban audit surveys about city emergencies***
- *Perceived risk in a printing factory*
- *Chronic pain threshold in TMD*
- *Synonimy and semantic space of words*
- *Ethnical identity of immigrants by cohorts*
- *European Union objectives and policies*
- ***Subjective survival probability to 75 and 90 years***

Available software for CUB models inference

- Based on a previous efficient code, written in ***GAUSS***© language, a software in ***R*** is now ***freely*** available.

- The current versione (1.1) may be downloaded from:

`http://www.dipstat.unina.it/cub/cubmodels.htm`

- Software is released with a paper (available in pdf) where essential features of **CUB** models are explained. There is a continuously updated bibliography about this class of models and step-by-step commands to implement immediate applications of the software.

- Current version might be further improved in order to make efficient the numerical aspects of the procedure.

Part III

Experiences on real data with CUB models

Experiences on real data sets

- We will present some results about modelling ordinal data set with **CUB** models:

1. Preferences for the cities where to live $(m = 12)$

2. Urban audit surveys about city emergencies $(m = 9)$

3. Subjective survival probability to 75 and 90 years $(m = 7)$

Preferences for the cities where to live

Preferences for the cities where to live ....[1]

- ➤ A study has been pursued for studying the preference of young people (living in Naples) towards several italian cities by collecting ranking of a sample of $n = 214$ observations.

- ➤ Each rater was asked to rank $m = 12$ cities from the most preferred as living place, until th eleast preferred on the basis of the same criterion.

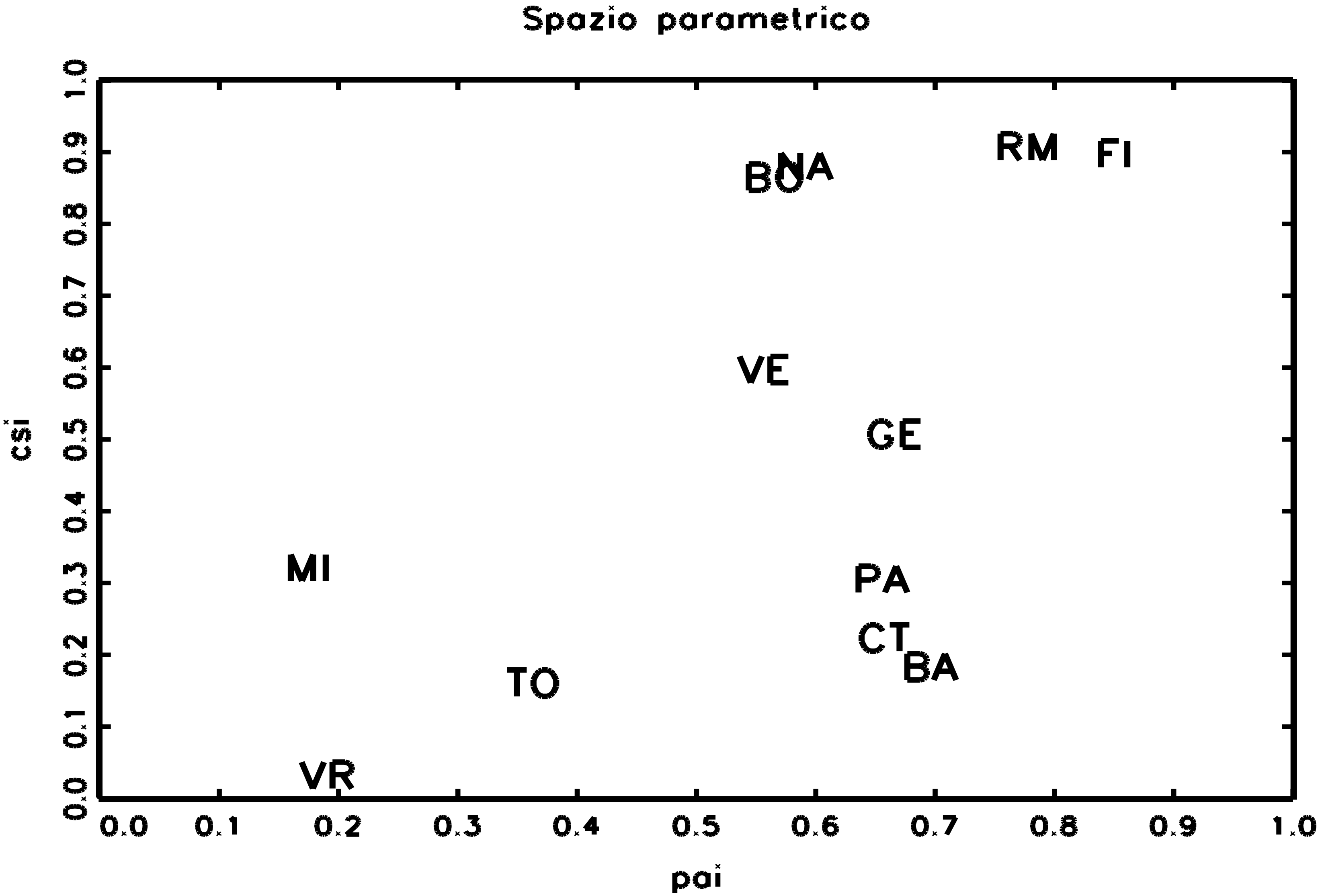
- ➤ In the next table estimated parameters for **CUB** $(0, 0)$ models are shown and listed according to the average rank expressed by respondents.

- ➤ The related representation in the parametric space of (π, ξ) helps effectively in characterizing the survey and in synthesizing hundreds of responses.

Preferences for the cities where to live .... [2]

Città	$\hat{\pi}$	$es(\pi)$	$\hat{\xi}$	$es(\pi)$
Firenze	0.834	(0.035)	0.879	(0.008)
Roma	0.750	(0.041)	0.889	(0.009)
Napoli	0.566	(0.051)	0.862	(0.013)
Bologna	0.539	(0.053)	0.847	(0.014)
Venezia	0.535	(0.064)	0.579	(0.020)
Genova	0.642	(0.060)	0.489	(0.017)
Milano	0.155	(0.065)	0.303	(0.060)
Verona	0.168	(0.040)	0.014	(0.013)
Palermo	0.631	(0.057)	0.287	(0.015)
Torino	0.340	(0.055)	0.142	(0.020)
Catania	0.635	(0.053)	0.205	(0.013)
Bari	0.672	(0.048)	0.165	(0.012)

Preferences for the cities where to live . . . . [3]



Emergencies of a metropolitan area

Main city problems in Naples, Italy . . . . [1]

➤ In December 2004, several subjects in Naples were asked to rank the main city problems (=*metropolitan emergencies*) of the city where they lived:

- ***Political patronage and corruption***
- ***Organized crime***
- ***Unemployment***
- ***Environmental pollution***
- ***Public health shortcomings***
- ***Petty crimes***
- ***Immigration***
- ***Streets cleanness and waste disposal***
- ***Traffic and local transport***

Main city problems in Naples, Italy . . . . [2]

- ➤ They should give rank 1 to the most serious problem, and so on, until rank 9 to the least serious one.

- ➤ Several information related to the subject were also collected: Gender, Age, Instruction, Residence, Working condition, etc.

- ➤ At the end, $n = 354$ questionnaires formed the basis of the analyses.

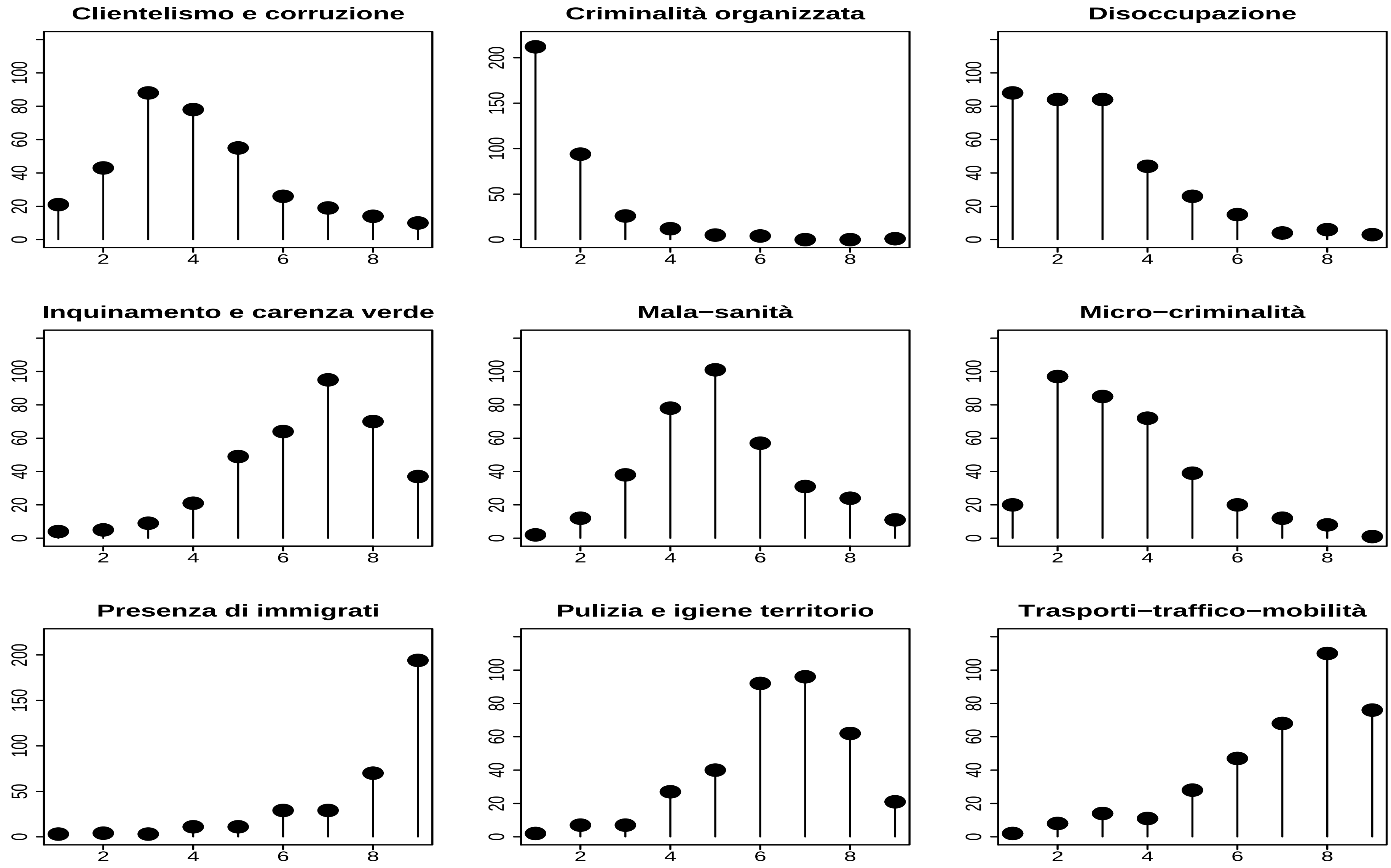
- ➤ Notice that, in this case study, the *feeling* is instead the ***concern*** about a problem.

Main city problems in Naples, Italy .... [3]

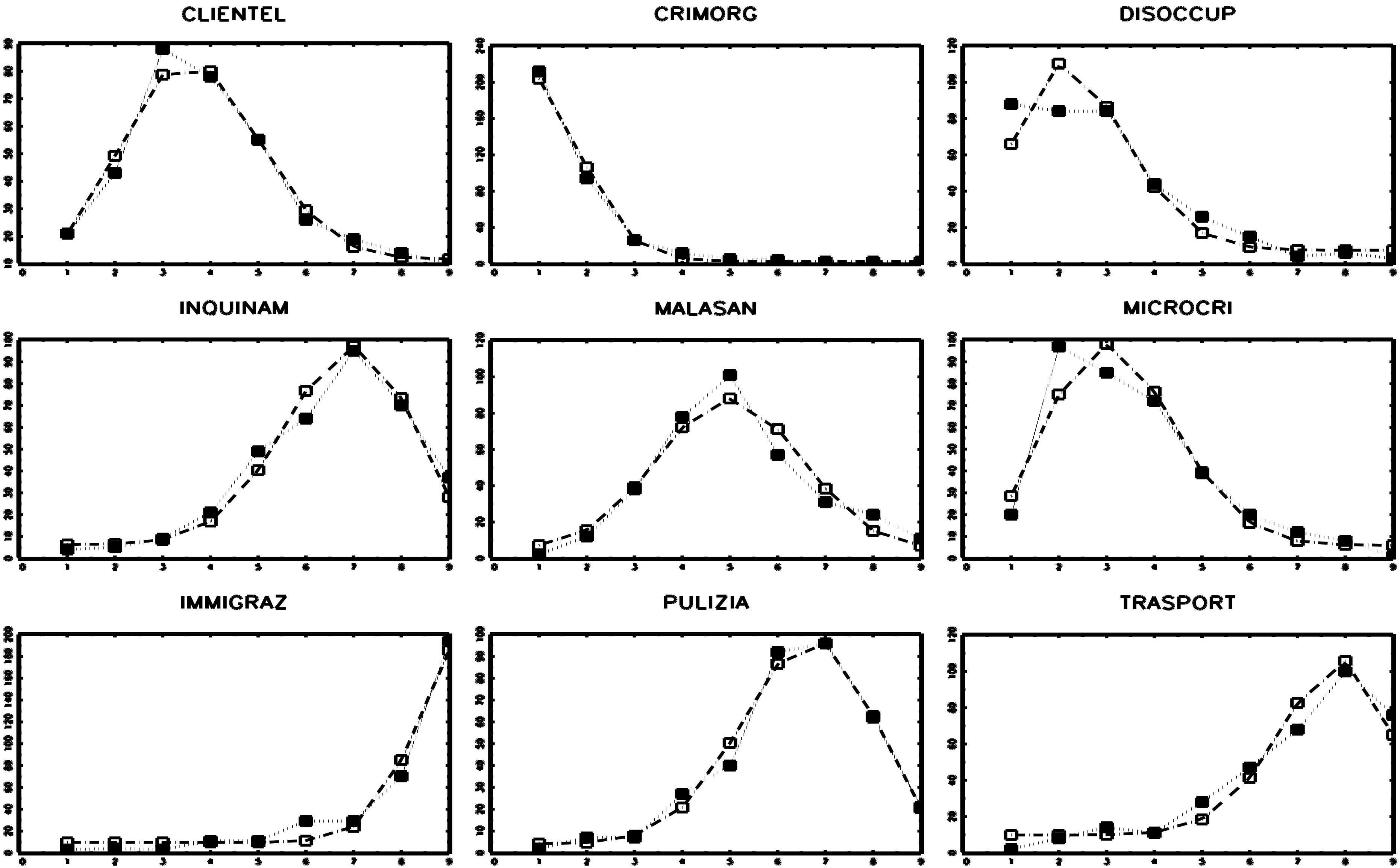
➤ For each problem, we obtained the following frequency distributions.

Main problems *Ranks* →	1	2	3	4	5	6	7	8	9
Political patronage and corruption	21	43	**88**	78	55	26	19	14	10
Organized crime	**212**	94	26	12	5	4	0	0	1
Unemployment	**88**	84	84	44	26	15	4	6	3
Environmental pollution	4	5	9	21	49	64	**95**	70	37
Public health shortcomings	2	12	38	78	**101**	57	31	24	11
Petty crimes	20	**97**	85	72	39	20	12	8	1
Immigration	3	4	3	11	11	29	29	70	**194**
Streets cleanness and waste disposal	2	7	7	27	40	92	**96**	62	21
Traffic and local transport	2	8	14	11	28	47	68	**100**	76

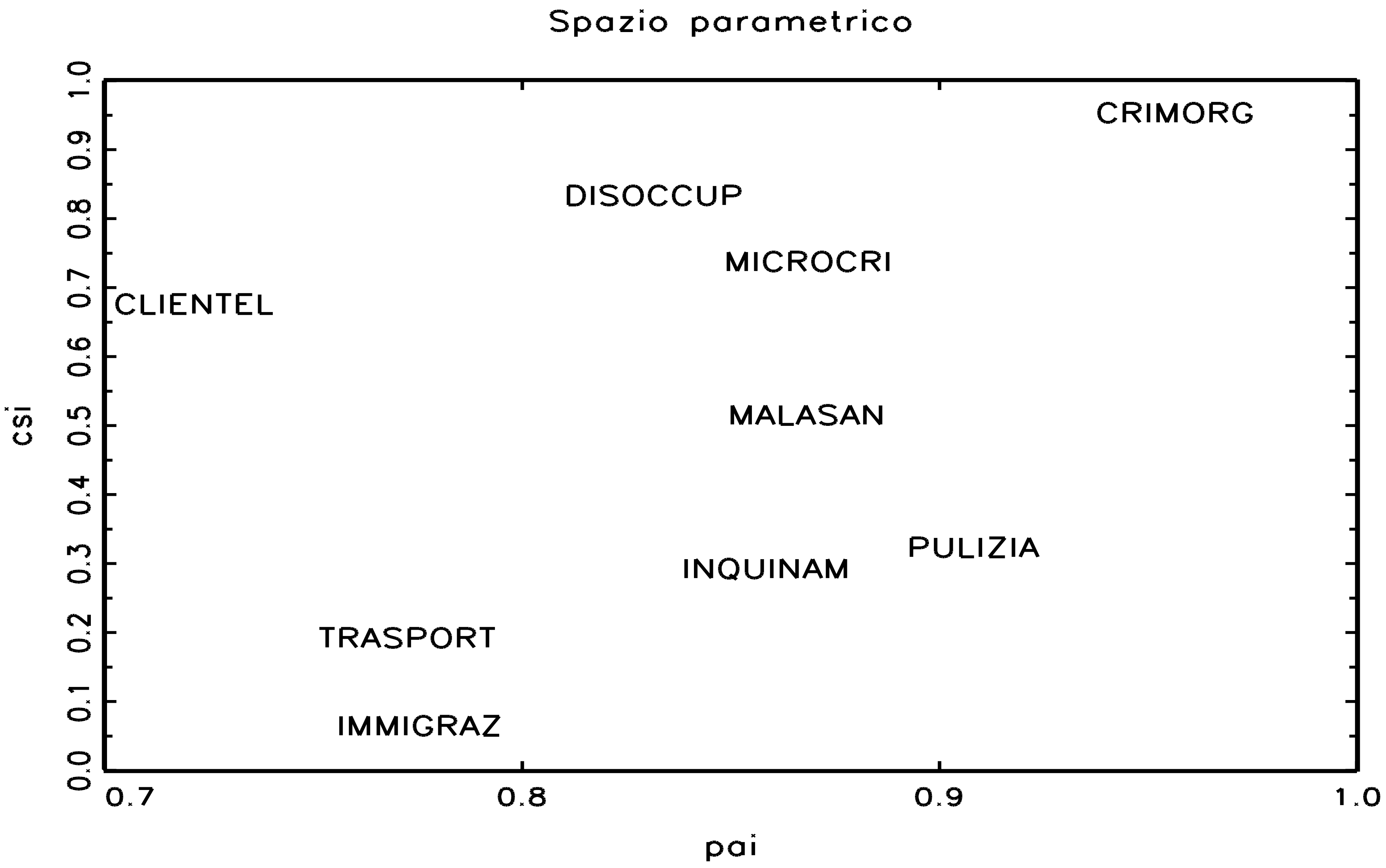
Main city problems in Naples, Italy . . . . [4]



Main city problems in Naples, Italy . . . . [5]



Main city problems in Naples, Italy .... [6]



Main city problems in Naples, Italy . . . . [8]

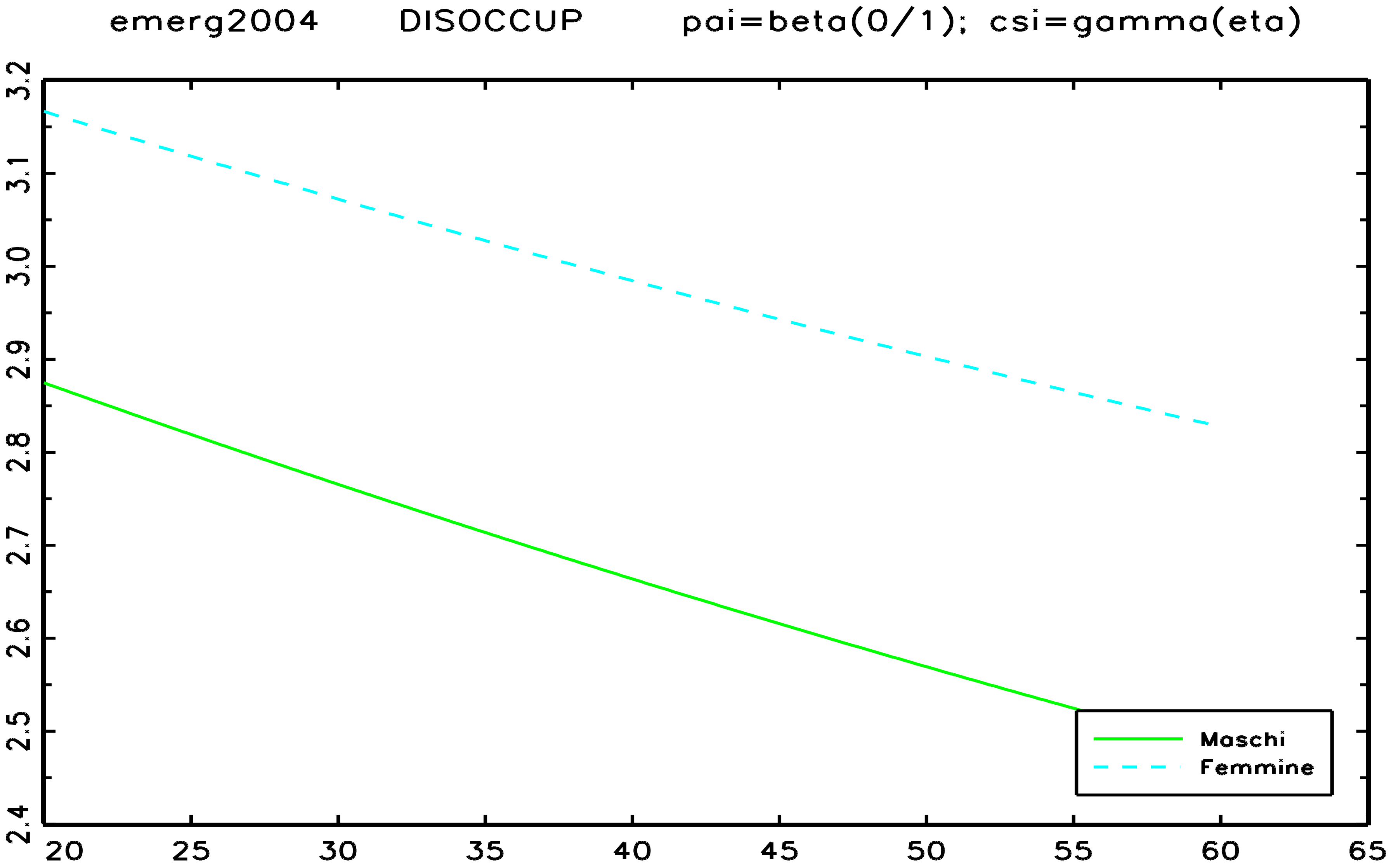
- We present a **CUB** model with covariates, with reference to the concerns for the "Unemployment" problem.

- This problem is one of the most serious in a city like Naples, mainly for young people: the unemployment rate is currently over 25%.

- We found that the ξ parameter requires the *Age* as a significant (continuous) covariate, while the uncertainty is better explained if we include the (dichotomous) covariate *Gender*.

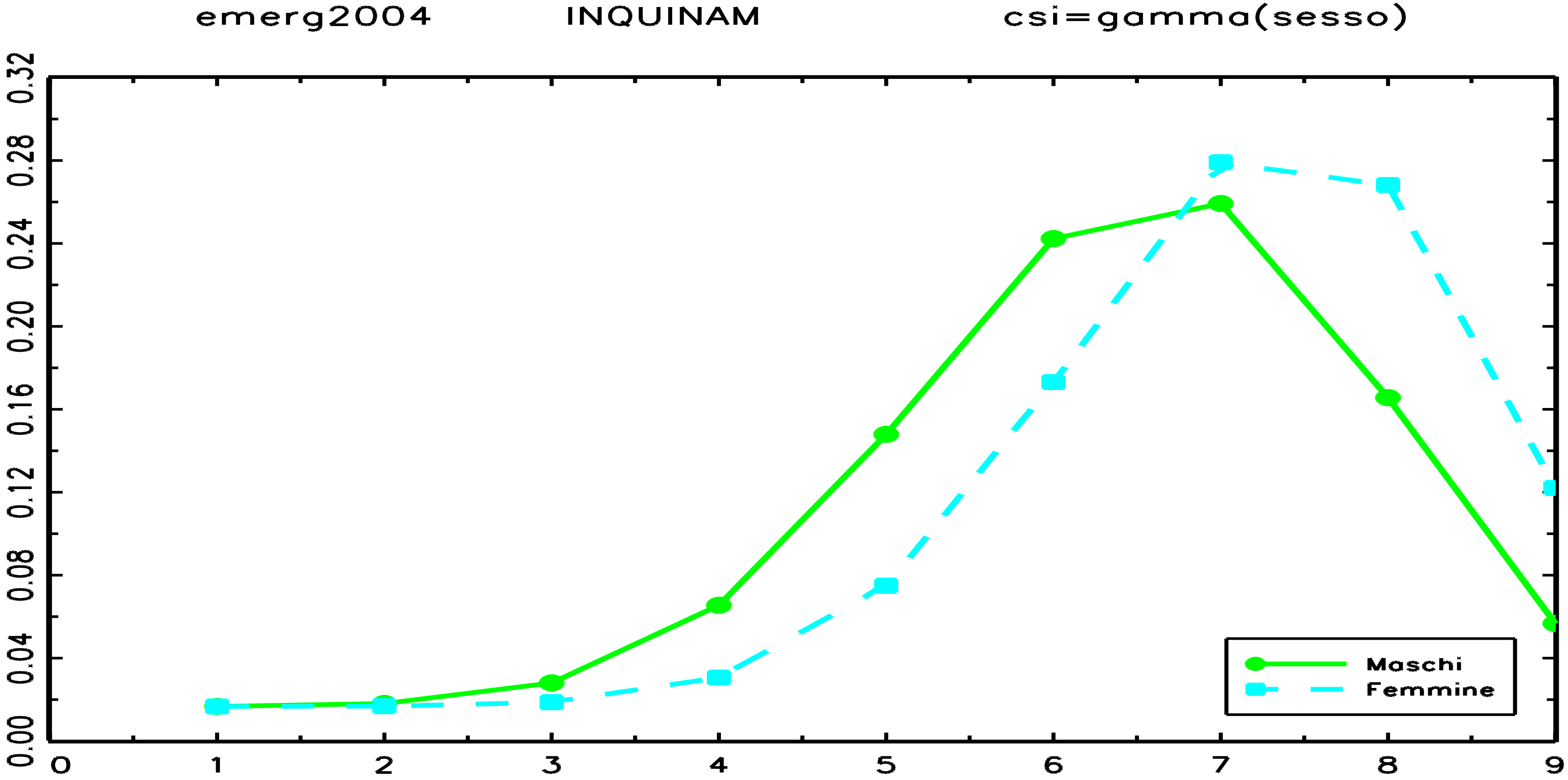
- In these cases, the evolution of the expected rank (that is, the *expected concern* for the Unemployment), given the dichotomous variable, is the most significante graph.

Main city problems in Naples, Italy .... [9]



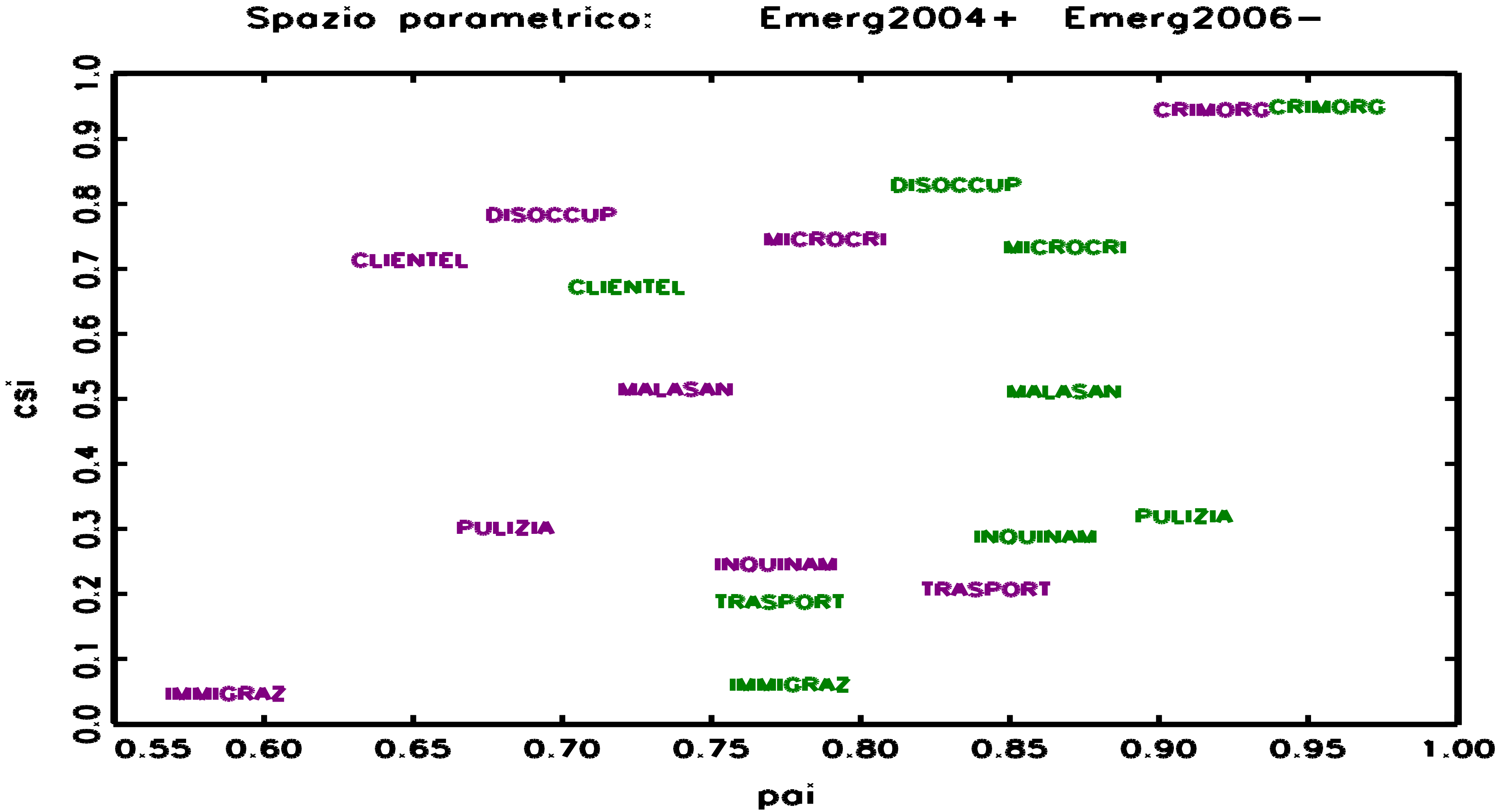
Main city problems in Naples, Italy .... [10]

➤ According to our sample, “Environmental pollution” is not a serious problem in Naples, since mode and median are 7, with an average rank of 6.5. However, there is some difference between the genders, with men more sensible to this issue.



Emergencies of a metropolitan area .... [10]

➤ We present the main result by repeating the same survey on different samples after two years: December 2004 (verde, $n_{2004} = 354$); December 2006 (viola, $n_{2006} = 419$)



Subjective survival probability to age 75 and 90 years

Subjective survival probability

- Subjective survival probability plays a fundamental role as a key variabile for making relevant decisions about consumption, saving, insurance, and so on.

- Analysis of subjective probability levels and investigation about covariates explaining their pattern are important issues for interpretation and prediction of human behaviour.

- Our focus is to model subjective survival probability expressed by respondents as expressed perception of some latent variable, and we discuss how respondents' covariates empirically affect subjective life expectancy.

- We do not argue with other methods but we will maintain the idea that an alternative approach deserve some consideration as an added value to the existing ones.

Subjective survival probability: PLUS survey

- We refer to PLUS, a cross-sectional survey carried out by ISFOL by means of a well structured questionnaire, obtained by CATI method and described by Peracchi and Perotti (2008).

- We refer to this paper for several considerations related to this topic, for the experimental design and related aspects of this sampling survey. Moreover, we will not discuss the comparison of subjective probabilities (produced by an extensive sampling of Italian population) and life-table data (prepared by ISTAT).

- Data set includes several variables related to subjects and information about gender, age, geographical region and town size, educational attainments, marital status, presence of children, household head status, activity status, and nationality.

- Self-reported health status, perceived quality of public health services and emergency and, for people currently working, several questions (earning, satisfaction, risk, etc.) related to job are also available for a subset of respondents.

- We limit ourselves to consider complete data (no missing values) and also validated data by removing inconsistencies in the answers.

Subjective survival probability: PLUS survey

➤ The fundamental issue for this talk is the quantitative expression of subjective survival probability which has been collected in the following way:

> ***For scientific purposes only, we would like to ask you:***
>
> ***“In your opinion, what is the probability that you will reach age 75, and age 90?”***
>
> ***Please provide a value between 100 (certain event) and 0 (impossible event).***
>
> 1. ***Probability of reaching age 75: ......out of 100***
> 2. ***Probability of reaching age 90: ......out of 100***

➤ All subsequent analyses are based on validated and consistent responses of $n = 20184$ people of age 15-64 years to both questions and their relevant covariates.

Interpretation of the given percentages

- We interpret responses as the final result of a hierarchical two-steps process:
  - First, one reacts to assess the perception of the problem and locate the answer in a broad sector of the admissible range.
  - Then, respondent expresses this perception on a numeric scale following standard rules of approximation and rounding.

- As a consequence, numbers we are going to study are the result of both:
  - ***psychological factors***, and
  - ***arithmetical effects***.

- Effective statistical models for the responses should consider both causes.

From numerical to qualitative assessments

➤ Expressed subjective survival probabilities are qualitative judgment about the occurrence of an event and they cannot be considered as numbers in a strict sense.

➤ We motivate this opinion by the following arguments:

- **`Rounding effects:`** people asked to choose a real number within a range like $[0, 100]$ mostly prefer a multiple of 5, with focal responses at multiples of $1/4$, and large preference for extreme values as 0 (impossible event) or 100 (certain choice).
- **`Selective effects:`** it seems hard to say that people realize that 0.15, 0.18, 0.20 are three distinct assessments of probability; thus, they perhaps cannot be perceived as different responses.
- **`Miscellanea effects:`** education, job, sport practices, numeracy ability, mass-media, etc. modify in a significant way the consideration of such scales.

➤ We will study the expressed evaluation of survival probability on a qualitative ordinal scale, and we choose a a 7-points Likert scale.

➤ Of course, any splitting of a real interval into a finite number of bins is arbitrary and different alternatives are legitimate. Thus, there is ***no correct*** solution to this problem and we are choosing a ***useful*** one, mostly motivated by a compromise between sensitivity and sparseness.

Qualitative assessment of subjective survival probability

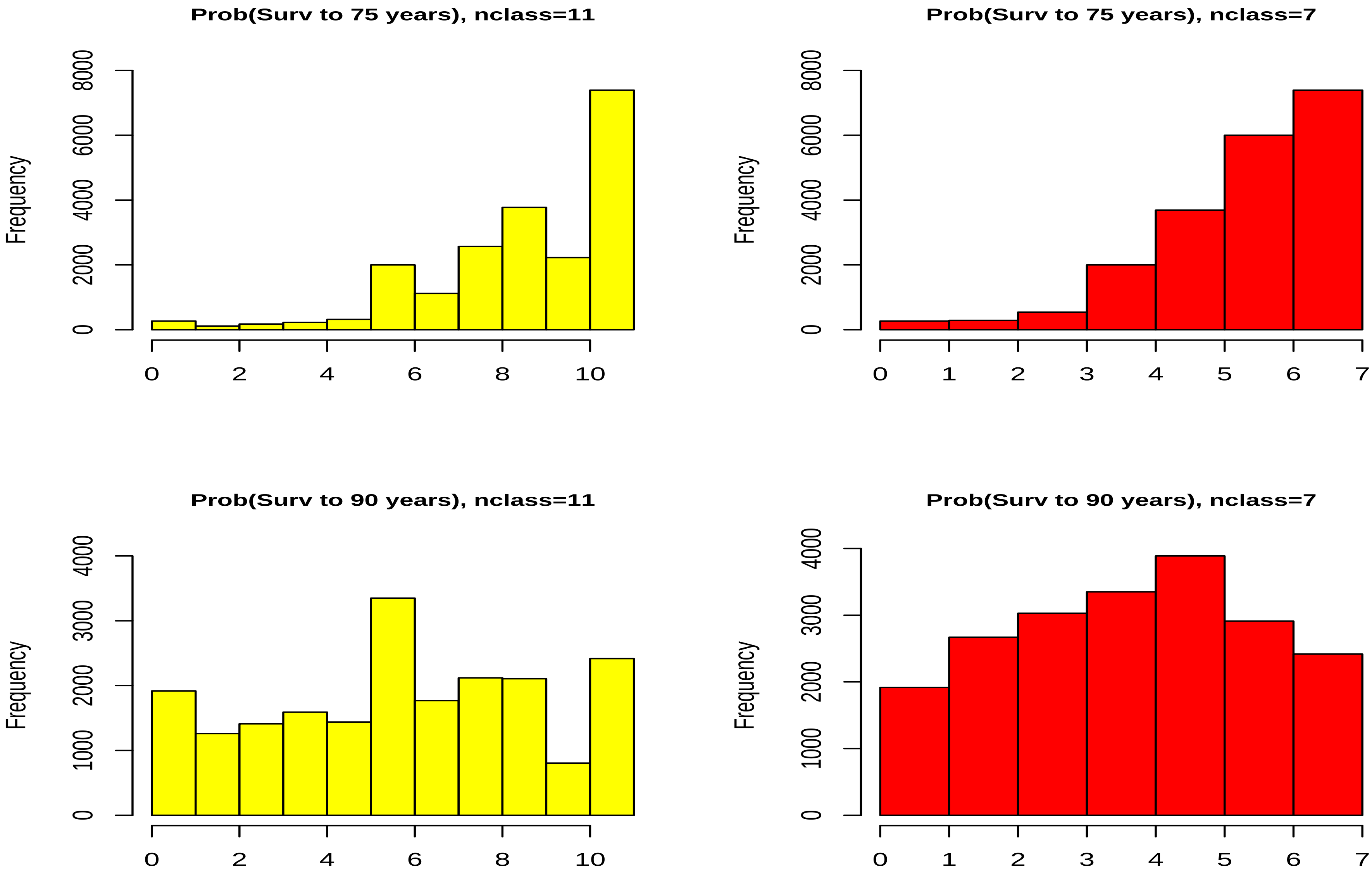
Class	*Subjective survival probability*	*Interpretation of the perception*
1	$0.00 \leq Pr(S) \leq 0.05$	**IMPOSSIBLE/Almost IMPOSSIBLE**
2	$0.05 < Pr(S) \leq 0.25$	**LOW**
3	$0.25 < Pr(S) \leq 0.45$	**Moderately LOW**
4	$0.45 < Pr(S) \leq 0.55$	**About FIFTY/FIFTY**
5	$0.55 < Pr(S) \leq 0.75$	**Moderately HIGH**
6	$0.75 < Pr(S) \leq 0.95$	**HIGH**
7	$0.95 < Pr(S) \leq 1.00$	**SURE/Almost SURE**

Rationale for a 7-points scale

The rationale for the previous classification stems from the following considerations:

- First of all, we realize that people cannot distinguish between small interval of probability when they express subjective probability; thus, although there are logical reasons for distinguishing between a 0 probability and a probability declared as small as 0.05, we think that in this range people are expressing an evaluation almost impossible, and we put this perception in class 1.
- A dual argument for almost sure perception produces a specular result, and in this way we define class 7.
- Then, we are considering as “low” and “high” the evaluations that are less than $1/4$ and more than $3/4$, respectively, and this seems reasonable as far as the common use of this quantities is considered.
- Subjective probabilities centred at $1/2$, that is in the range $(0.45, 0.55]$, deserve special consideration since they are just expressing a sound uncertainty in the evaluation, that is an *epistemic uncertainty* (Bruine de Bruin et al., 2002).
- Finally, classes 3 and 5 are uniquely determined given the previous ones.

Comparing 11-points and 7-points scales



CUB models for subjective survival probability

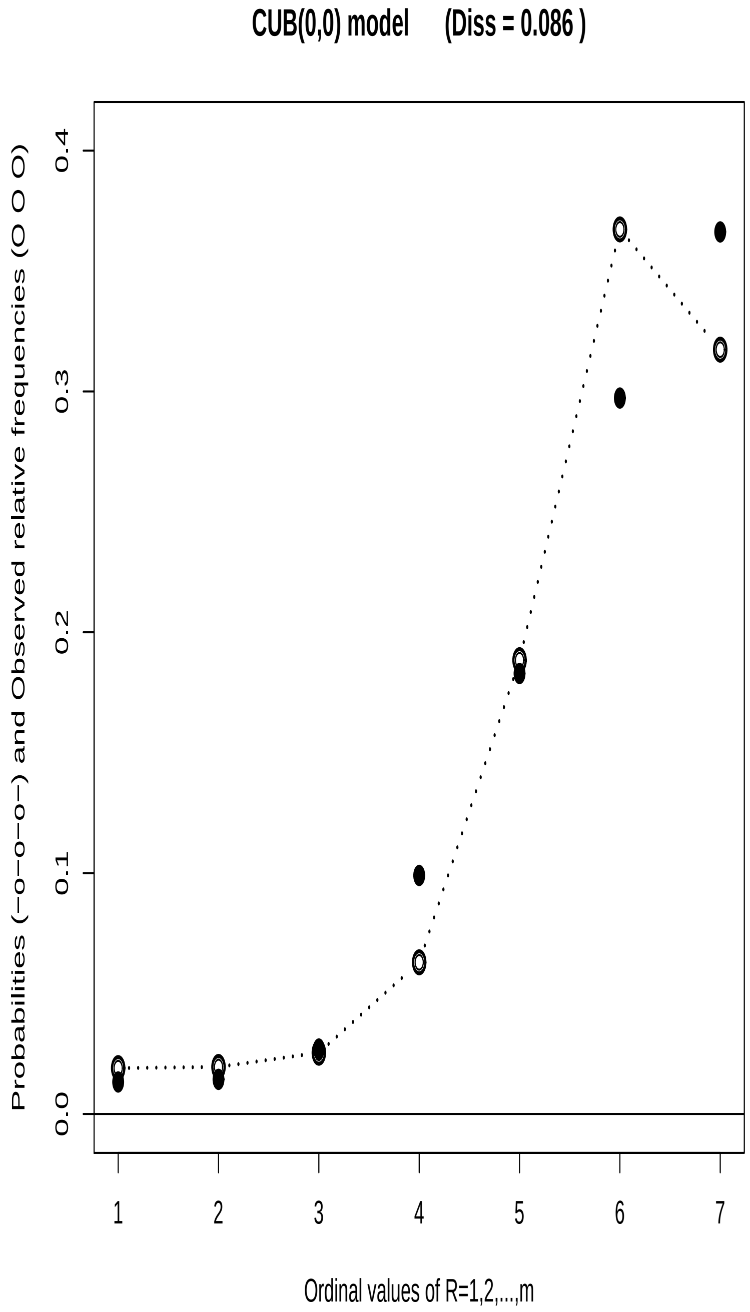
- We will specify and estimate **CUB** models for perceived survival probability to 75 and 90 years, respectively, and will discuss the main results in terms of interpretations and visualizations in their parameter space.

- Estimation will be pursued without and with the introduction of sensible covariates (gender, marital status, work position, age), which will be maintained in the models if and only if they are significant.

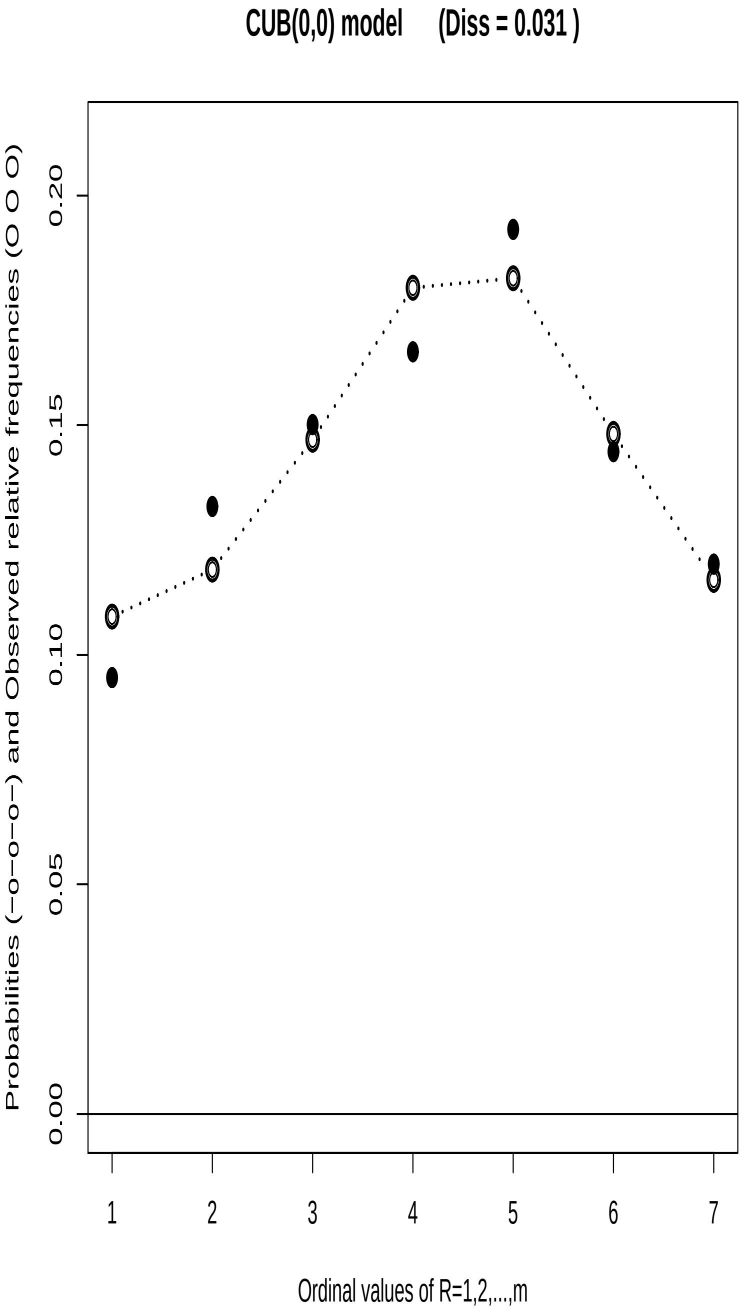
- Firstly, we estimate **CUB** $(0, 0)$ and obtain global results for both probabilities as in the following table.

Survival probability	$\hat{\pi}$	$\hat{\xi}$	$\ell(\hat{\boldsymbol{\theta}})/n$	$Diss$
to age 75 years	0.867 *(0.005)*	0.163 *(0.001)*	-1.505	0.087
to age 90 years	0.252 *(0.009)*	0.422 *(0.006)*	-1.927	0.031

Modelling subjective survival probability

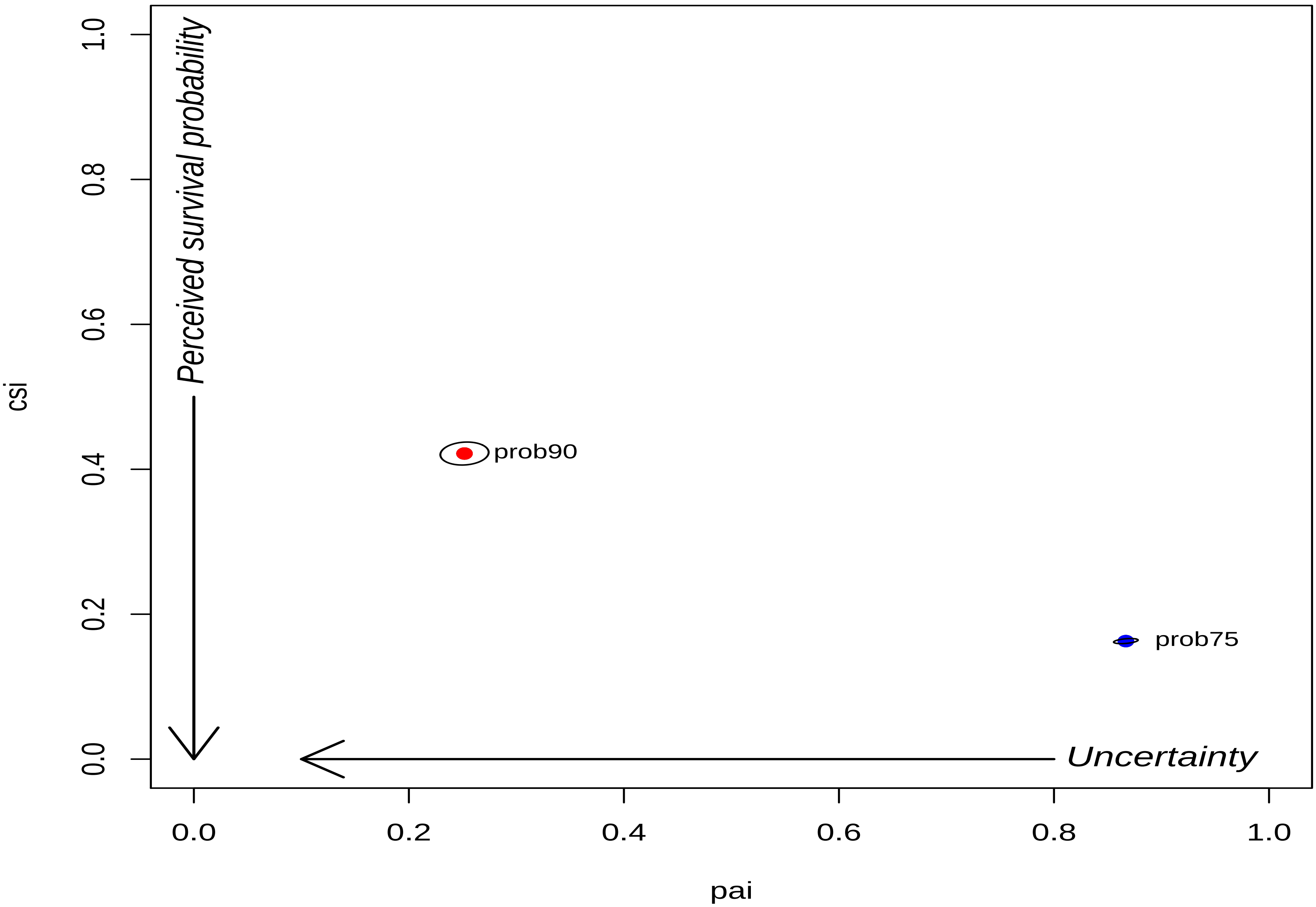


CUB(0,0) model (Diss = 0.086)



CUB(0,0) model (Diss = 0.031)

Modelling subjective survival probability



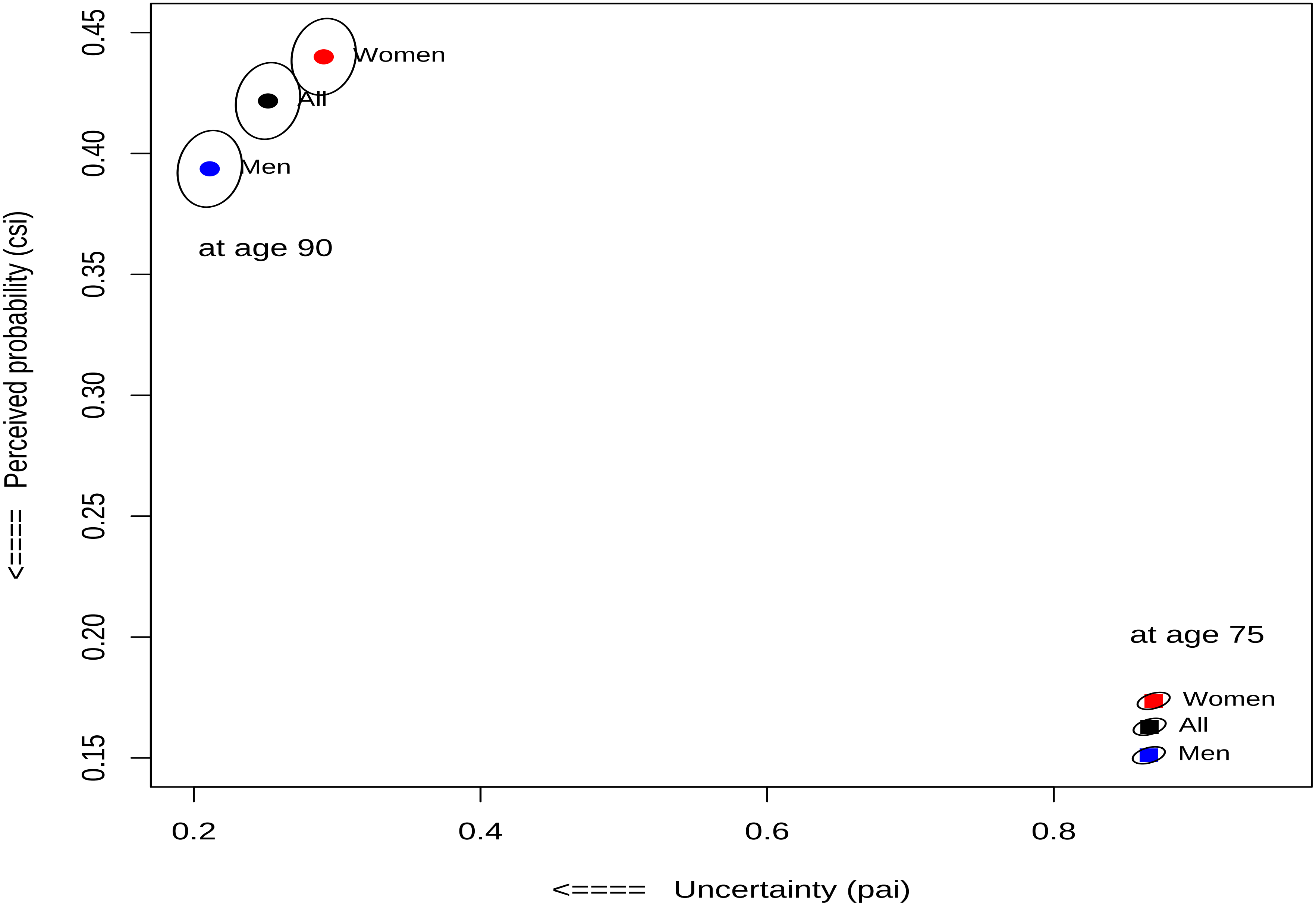
Estimated CUB models with gender effect (age 75)

Models	Parameters estimates		$\ell(\hat{\boldsymbol{\theta}})$
	Subjective survival probability to age 75		
CUB $(0,0)$ Women	$\hat{\pi} = 0.870$ *(0.006)*	$\hat{\xi} = 0.174$ *(0.002)*	-16605
CUB $(0,0)$ Men	$\hat{\pi} = 0.866$ *(0.007)*	$\hat{\xi} = 0.151$ *(0.001)*	-13742
CUB $(0,0)$	$\hat{\pi} = 0.867$ *(0.005)*	$\hat{\xi} = 0.163$ *(0.001)*	-30383
CUB $(1,0)$	$\hat{\beta}_0 = 1.985$ *(0.056)* $\hat{\beta}_1 = -0.218$ *(0.073)*	$\hat{\xi} = 0.163$ *(0.001)*	-30378
CUB $(0,1)$	$\hat{\pi} = 0.868$*(0.005)*	$\hat{\gamma}_0 = -1.724$ *(0.015)* $\hat{\gamma}_1 = 0.163$ *(0.019)*	-30347
CUB $(1,1)$	$\hat{\beta}_0 = 1.868$ *(0.056)* $\hat{\beta}_1 = 0.029$ *(0.080)*	$\hat{\gamma}_0 = -1.726$ *(0.016)* $\hat{\gamma}_1 = 0.166$ *(0.021)*	-30347

Estimated CUB models with gender effect (age 90)

Models	Parameters estimates		$\ell(\hat{\boldsymbol{\theta}})$
Subjective survival probability to age 90			
CUB $(0, 0)$ Women	$\hat{\pi} = 0.291$ *(0.013)*	$\hat{\xi} = 0.440$ *(0.008)*	-20870
CUB $(0, 0)$ Men	$\hat{\pi} = 0.211$ *(0.013)*	$\hat{\xi} = 0.394$ *(0.011)*	-18016
CUB $(0, 0)$	$\hat{\pi} = 0.252$ *(0.009)*	$\hat{\xi} = 0.422$ *(0.006)*	-38900
CUB $(1, 0)$	$\hat{\beta}_0 = -1.313$ *(0.080)* $\hat{\beta}_1 = 0.404$ *(0.102)*	$\hat{\xi} = 0.425$ *(0.006)*	-38892
CUB $(0, 1)$	$\hat{\pi} = 0.253$*(0.009)*	$\hat{\gamma}_0 = -0.414$ *(0.042)* $\hat{\gamma}_1 = 0.165$ *(0.054)*	-38896
CUB $(1, 1)$	$\hat{\beta}_0 = -1.318$ *(0.080)* $\hat{\beta}_1 = 0.426$ *(0.100)*	$\hat{\gamma}_0 = -0.432$ *(0.047)* $\hat{\gamma}_1 = 0.191$ *(0.056)*	-38886

Modelling subjective survival probability (gender)



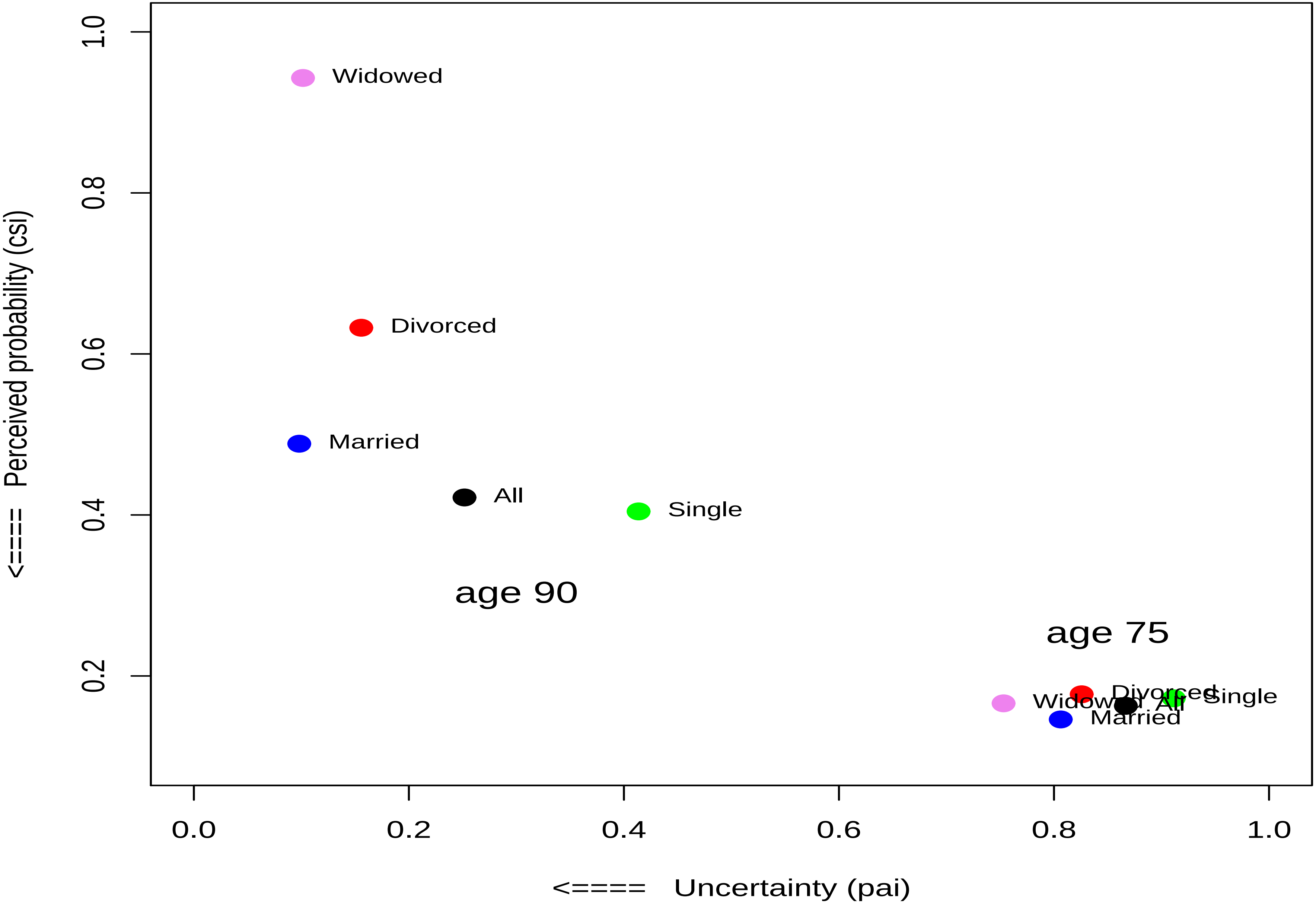
Estimated CUB $(1,4)$ model (gender and marital status)

- We realize that uncertainty parameter is significant only for single with respect to married, whereas there is a sensible effect of both gender and marital status as far as perceived probability is concerned.

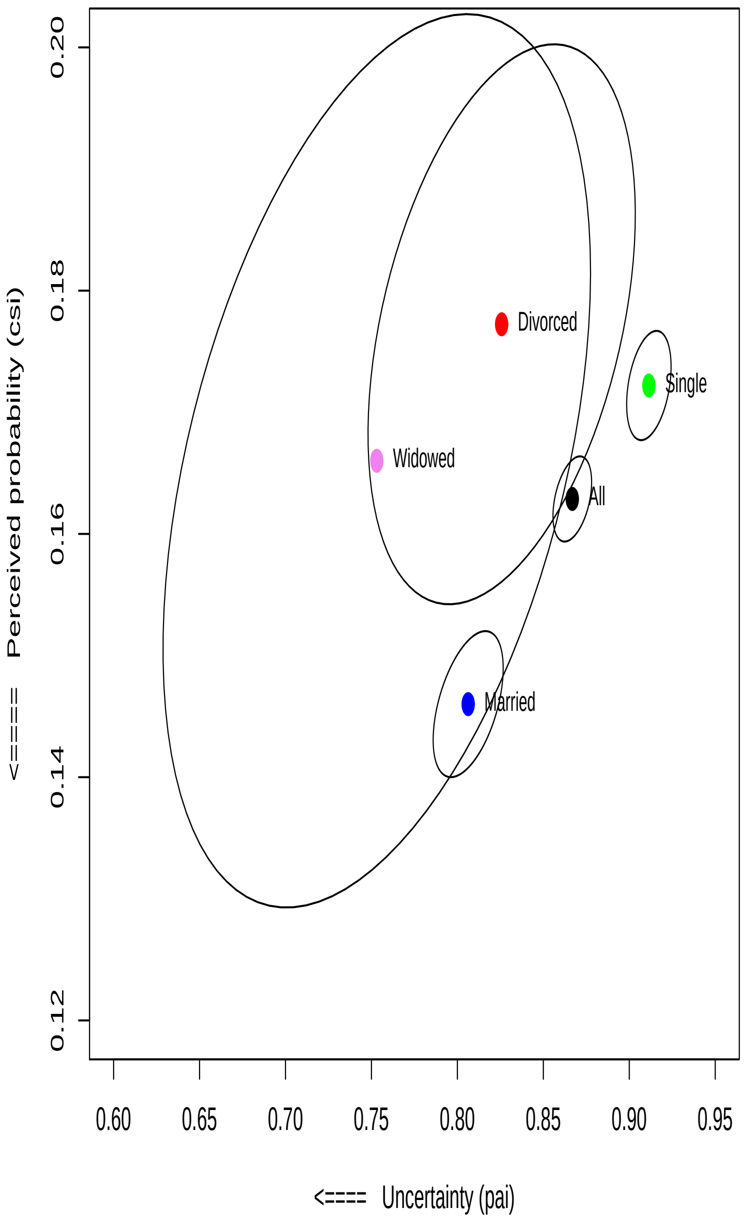
Dummies	Parameters estimates	
Constant	$\hat{\beta}_0 = 1.447$ *(0.051)*	$\hat{\gamma}_0 = -1.841$ *(0.022)*
Single	$\hat{\beta}_1 = 0.877$ *(0.083)*	$\hat{\gamma}_1 = 0.188$ *(0.023)*
Divorced		$\hat{\gamma}_2 = 0.213$ *(0.063)*
Widowed		$\hat{\gamma}_3 = 0.163$ *(0.093)*
Female		$\hat{\gamma}_4 = 0.151$ *(0.020)*

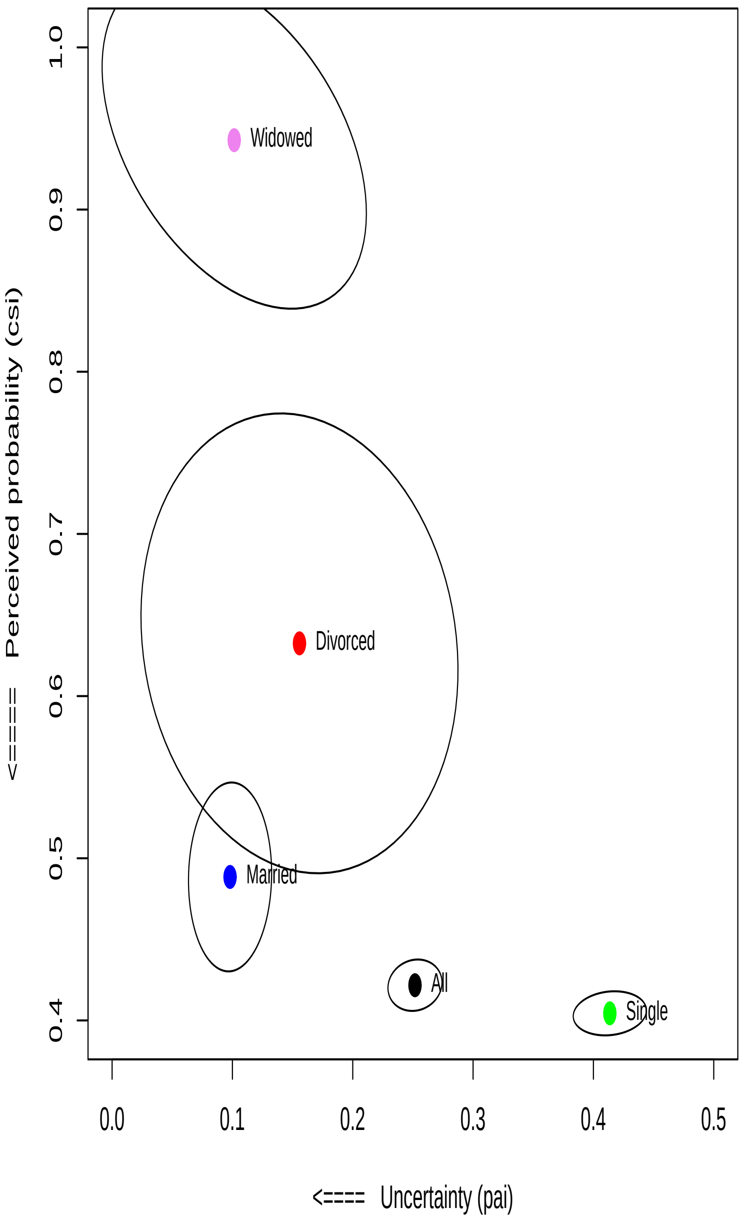
- For this data, the best CUB $(1,4)$ model improves by 23% log-likelihood with respect to the worst model (Uniform distribution of responses). Moreover, it reduces the deviance with respect to a CUB $(0,0)$, a model without covariates, as confirmed by a likelihood test of $2*(-30275-(-30383)) = 216$, which is highly significant with 4 degrees of freedom.

CUB models and marital status

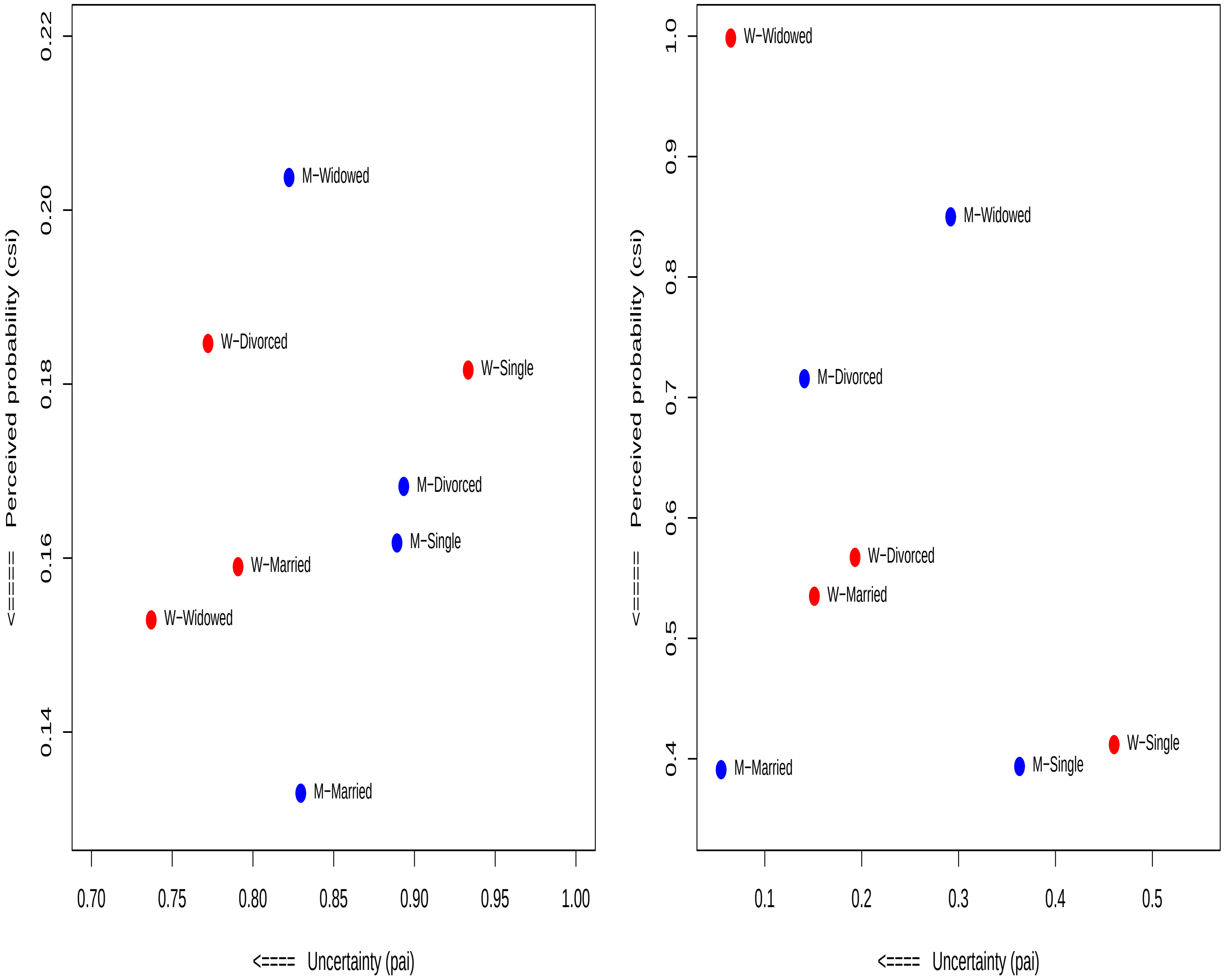


CUB models with marital status (enlarged)





CUB models with gender and marital status

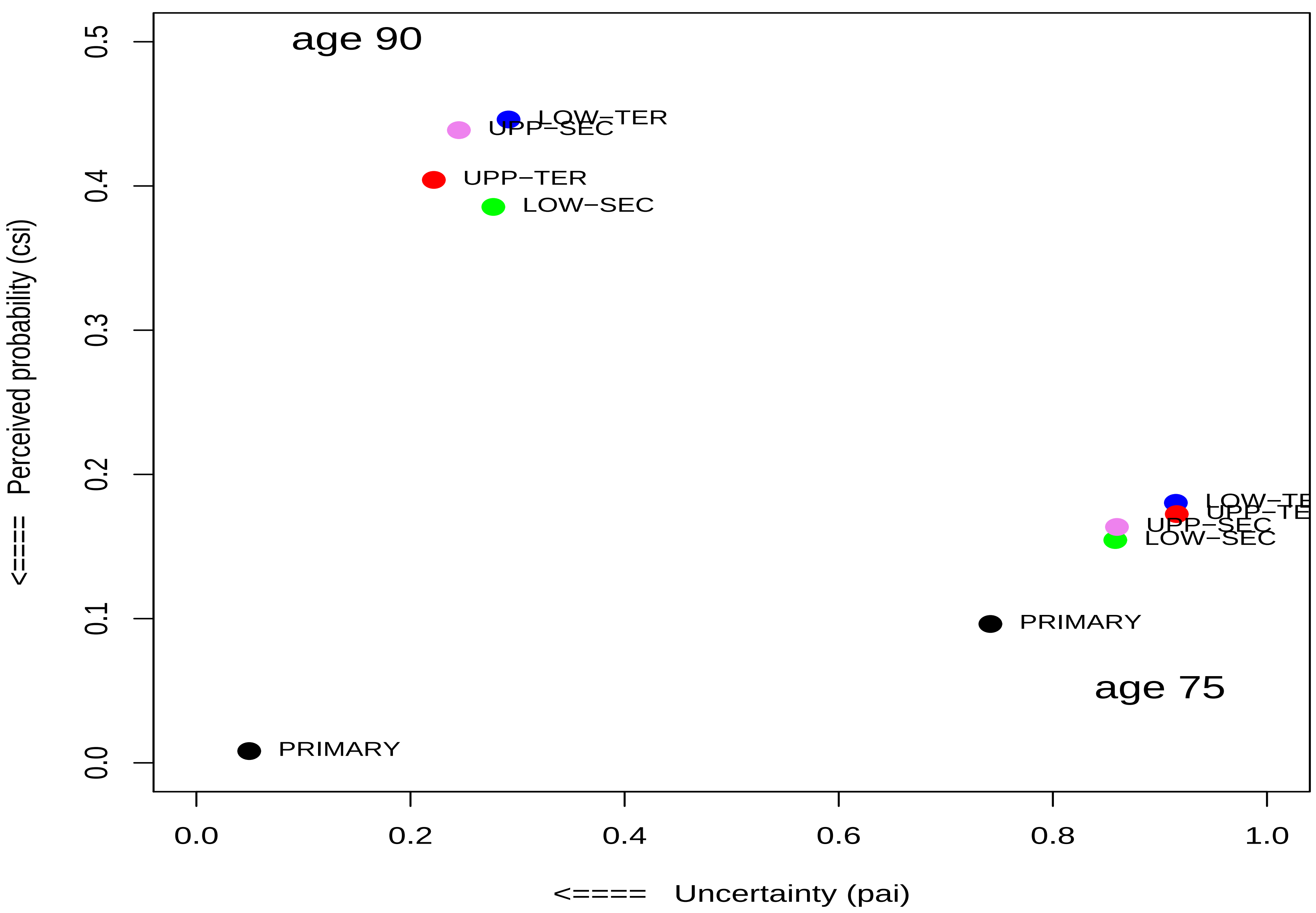


Education and subjective survival probability

➤ Respondents are classified with respect to education in the following five classes (in parenthesis, we report the short form and the size of each subgroup).

Educational achievement	*Sample size*
Primary education or less *(PRIMARY)*	$n = 580$
Lower secondary education (LOW-SEC)	$n = 5126$
Upper secondary education (UPP-SEC)	$n = 10603$
Lower tertiary education (LOW-TER)	$n = 3605$
Upper tertiary education (UPP-TER)	$n = 270$

Education and subjective survival probability



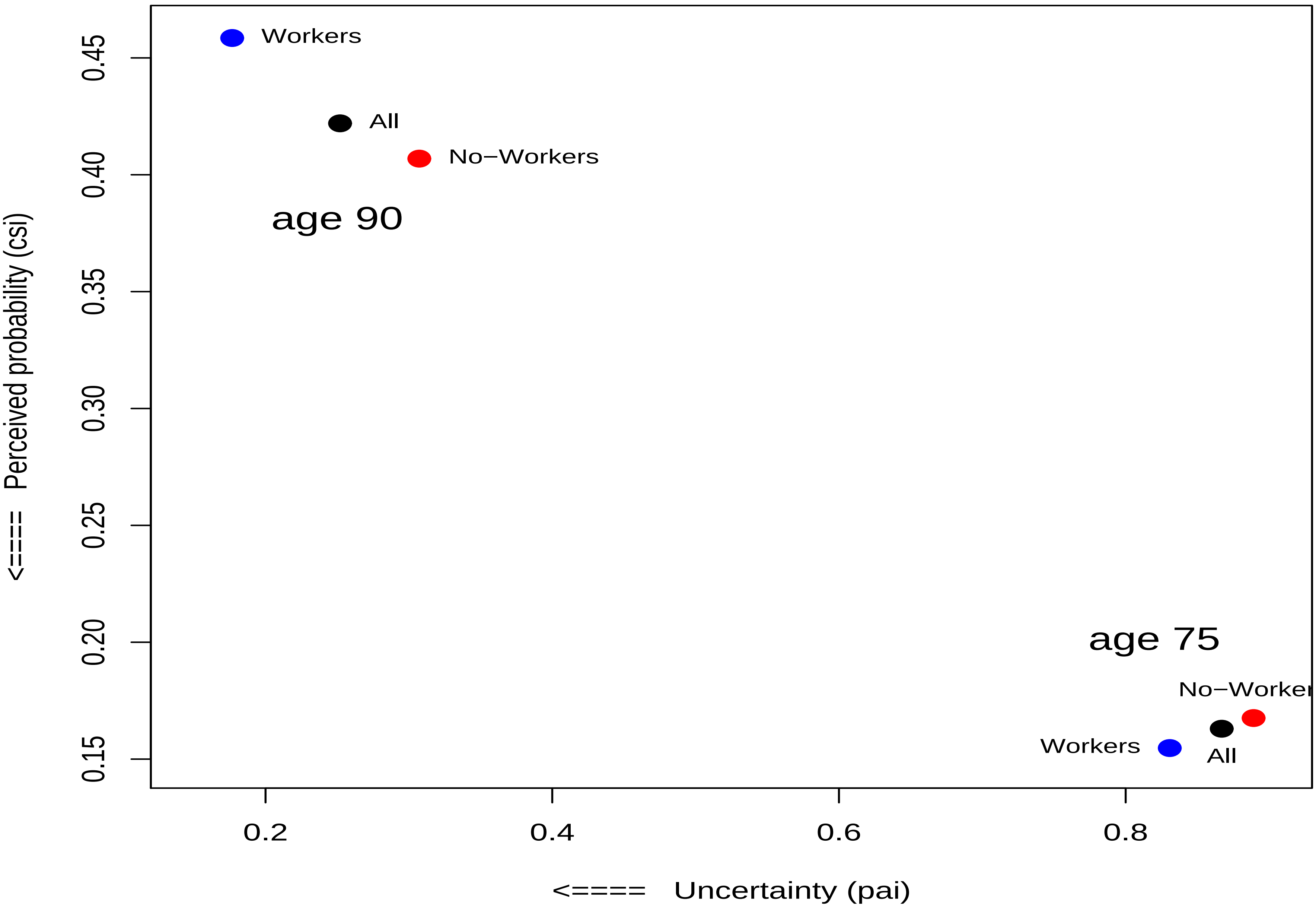
Education and subjective survival probability

- Plot enhances that a low education level (as primary) lowers also the perception of subjective probability of survival to both ages; moreover, in this subgroup, a great indecision is evident as far as responses to age 90 are concerned since the uncertainty parameter is close to 0.

- On the contrary, more educated people tend to give higher evaluation of probability according to the years of study with the exception of upper tertiary education subgroup which lower a bit (but systematically) this evaluation with respect to trend.

- Estimated model confirms that people with higher level of education are more pessimistic about the probability of survival perhaps as a consequences of having a thorough knowledge about their effective wealth status.

Working status and subjective survival probability



CUB models for age effects

Models	Parameters estimates		$\ell(\hat{\boldsymbol{\theta}})$
	Subjective survival probability to age 75		
CUB $(0,0)$	$\hat{\pi} = 0.867$ *(0.005)*	$\hat{\xi} = 0.163$ *(0.001)*	-30383
CUB $(2,2)$	$\hat{\beta}_0 = 1.518$ *(0.059)*	$\hat{\gamma}_0 = -1.559$ *(0.016)*	-30221
$log(age)$	$\hat{\beta}_1 = -1.338$ *(0.126)*	$\hat{\gamma}_1 = -0.233$ *(0.027)*	
$[log(age)]^2$	$\hat{\beta}_2 = 2.568$ *(0.327)*	$\hat{\gamma}_2 = -0.538$ *(0.071)*	
	Subjective survival probability to age 90		
CUB $(0,0)$	$\hat{\pi} = 0.252$ *(0.009)*	$\hat{\xi} = 0.422$ *(0.006)*	-38900
CUB $(1,2)$	$\hat{\beta}_0 = -1.255$ *(0.072)*	$\hat{\gamma}_0 = 0.142$ *(0.040)*	-38656
$log(age)$	$\hat{\beta}_1 = -2.636$ *(0.174)*	$\hat{\gamma}_1 = 0.389$ *(0.144)*	
$[log(age)]^2$		$\hat{\gamma}_2 = -0.369$ *(0.219)*	

Cohort and age effects

➤ We consider cohort effects in the response by introducing age of respondents as a continuous covariate.

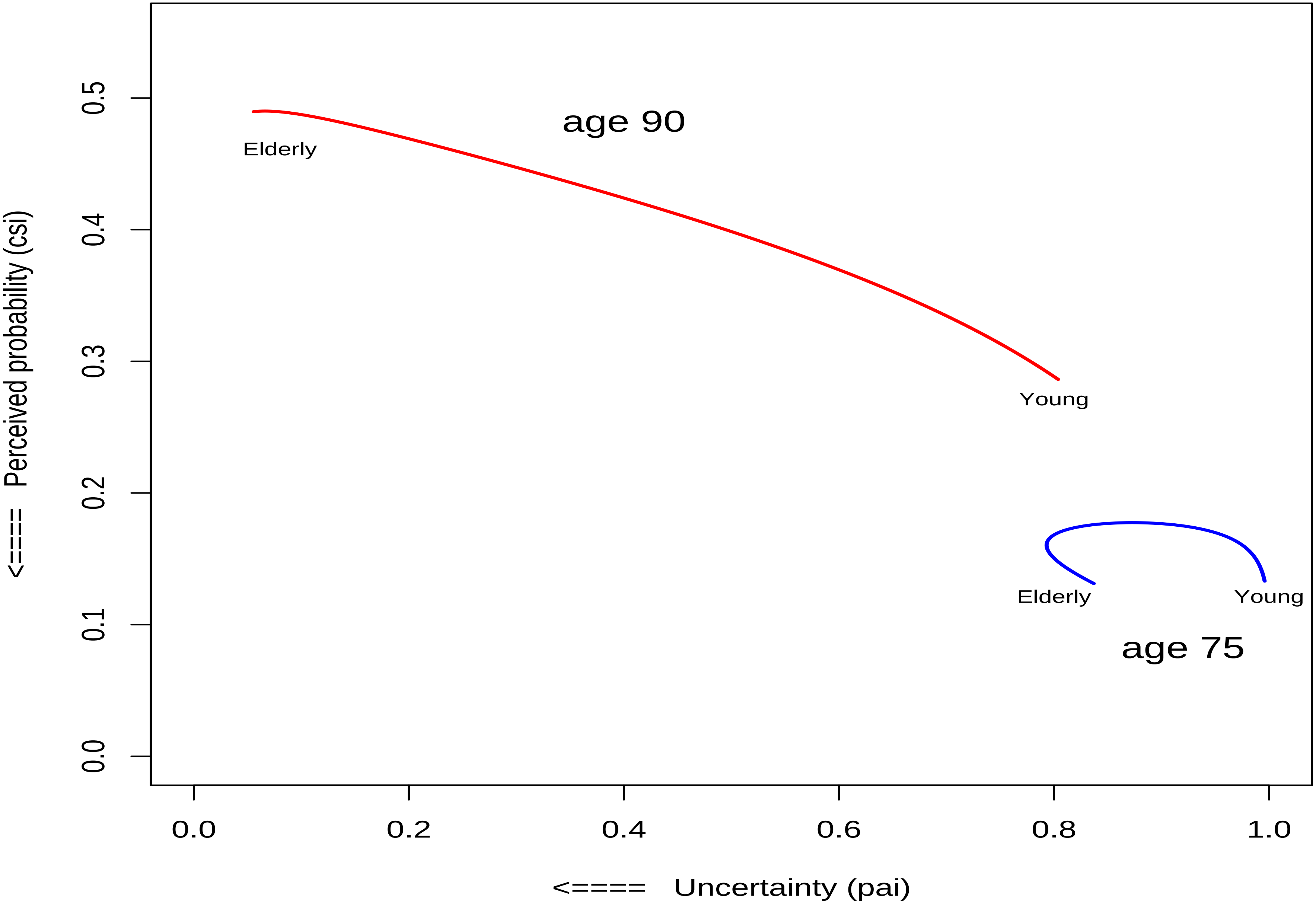
➤ Specifically, we use the natural logarithm of declared age in years and we introduce in the model deviations and squared deviations of this variable. Motivations for this transformations are standard in statistical modelling literature:

- logarithm accelerates the convergence of maximum likelihood estimators;
- deviations reduce collinearity problems in variance-covariance matrix of estimators;
- squared variables are important to fit some observed inversion of pattern with increasing ages.

➤ Now, we plot in the parametric space uncertainty and perception parameters as functions of age to visualize ***dynamically*** how subjective survival probabilities are modified by different cohorts of respondents. Indeed, we are graphing the parametric functions:

$$\pi = \pi(t); \qquad \xi = \xi(t); \quad t \in [15, 75].$$

Dynamic modelling subjective survival probability



Comprehensive CUB models

- ➤ If we limit the analysis only to covariates previously examined, we found significant result by including as relevant covariates:
  - ***to age 75:***
    - for explaining `uncertainty`: gender, single, primary education, working status, *ln(age)* and *ln(age)-squared*;
    - for explaining `perception`: gender, single, divorced, widowed, working status, *ln(age)* and *ln(age)-squared*;
  - ***to age 90:***
    - for explaining `uncertainty`: gender, single, primary education, working status, *ln(age)*;
    - for explaining `perception`: gender, single, divorced, widowed, primary education, working status, *ln(age)* and *ln(age)-squared*;

- ➤ It should be observed that when we introduced the same covariates for both parameters, in some cases, one or both of them drop as only one effect may result significant or because there is some collinearity among them. For instance, primary education (in the first response) turns out to be not significant as a covariate for explaining uncertainty and perception whereas single, divorced and working status are not significant for explaining perception. Similarly, this happens with primary education (strongly related to age of respondents, and thus to a cohort effect).

Omnibus CUB model (survival probability to age 75)

➤ Given the selected covariates, we present the best **CUB** model we obtained after a stepwise procedure aimed at reaching a model with all parameters significant.

Covariates	Parameters estimates	
Constant	$\hat{\beta}_0 = 1.510$ *(0.077)*	$\hat{\gamma}_0 = -1.603$ *(0.025)*
Gender	$\hat{\beta}_1 = 0.032$ *(0.081)*	$\hat{\gamma}_1 = 0.104$ *(0.022)*
Divorced		$\hat{\gamma}_2 = 0.246$ *(0.062)*
Widowed		$\hat{\gamma}_3 = 0.290$ *(0.091)*
Work		$\hat{\gamma}_4 = -0.047$ *(0.022)*
$ln(age)$	$\hat{\beta}_5 = -1.310$ *(0.126)*	$\hat{\gamma}_5 = -0.226$ *(0.028)*
$[ln(age)]^2$	$\hat{\beta}_6 = 2.534$ *(0.382)*	$\hat{\gamma}_6 = -0.559$ *(0.074)*

Omnibus CUB model (survival probability to age 90)

➤ Given the selected covariates, we present the best CUB model we obtained after a stepwise procedure aimed at reaching a model with all parameters significant.

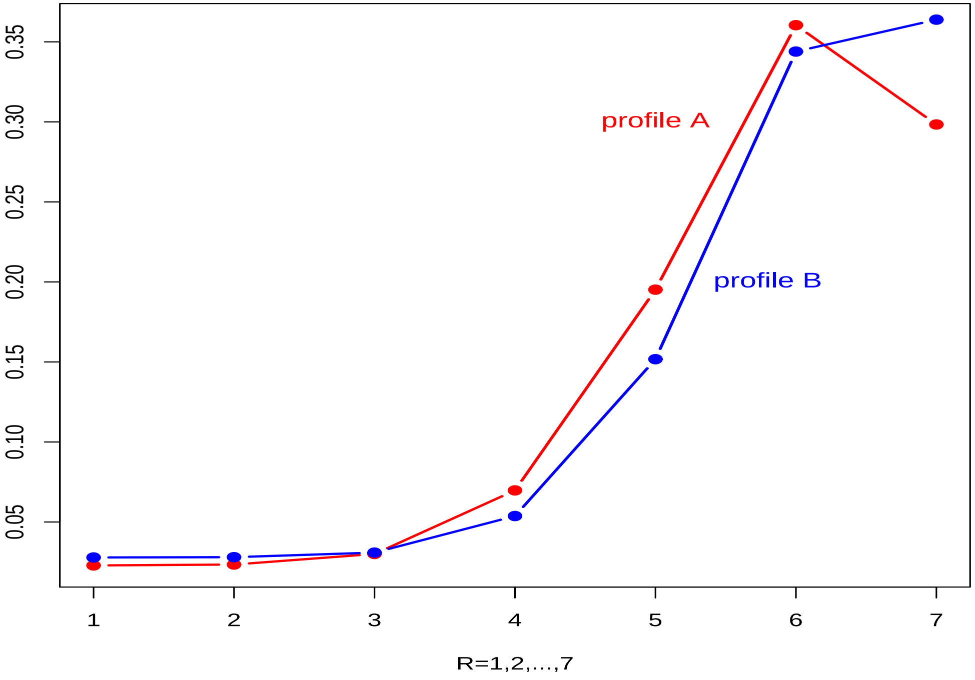
Covariates	Parameters estimates	
Constant	$\hat{\beta}_0 = -1.868$ *(0.169)*	$\hat{\gamma}_0 = -0.201$ *(0.051)*
Gender	$\hat{\beta}_1 = 0.384$ *(0.192)*	$\hat{\gamma}_1 = 0.081$ *(0.044)*
Single	$\hat{\beta}_2 = 0.949$ *(0.192)*	
Widowed		$\hat{\gamma}_3 = 2.231$ *(0.575)*
Work	$\hat{\beta}_4 = -0.345$ *(0.119)*	
$ln(age)$	$\hat{\beta}_5 = -1.676$ *(0.221)*	$\hat{\gamma}_5 = 0.591$ *(0.085)*

Subjects' profiles and estimated probability (age 75)

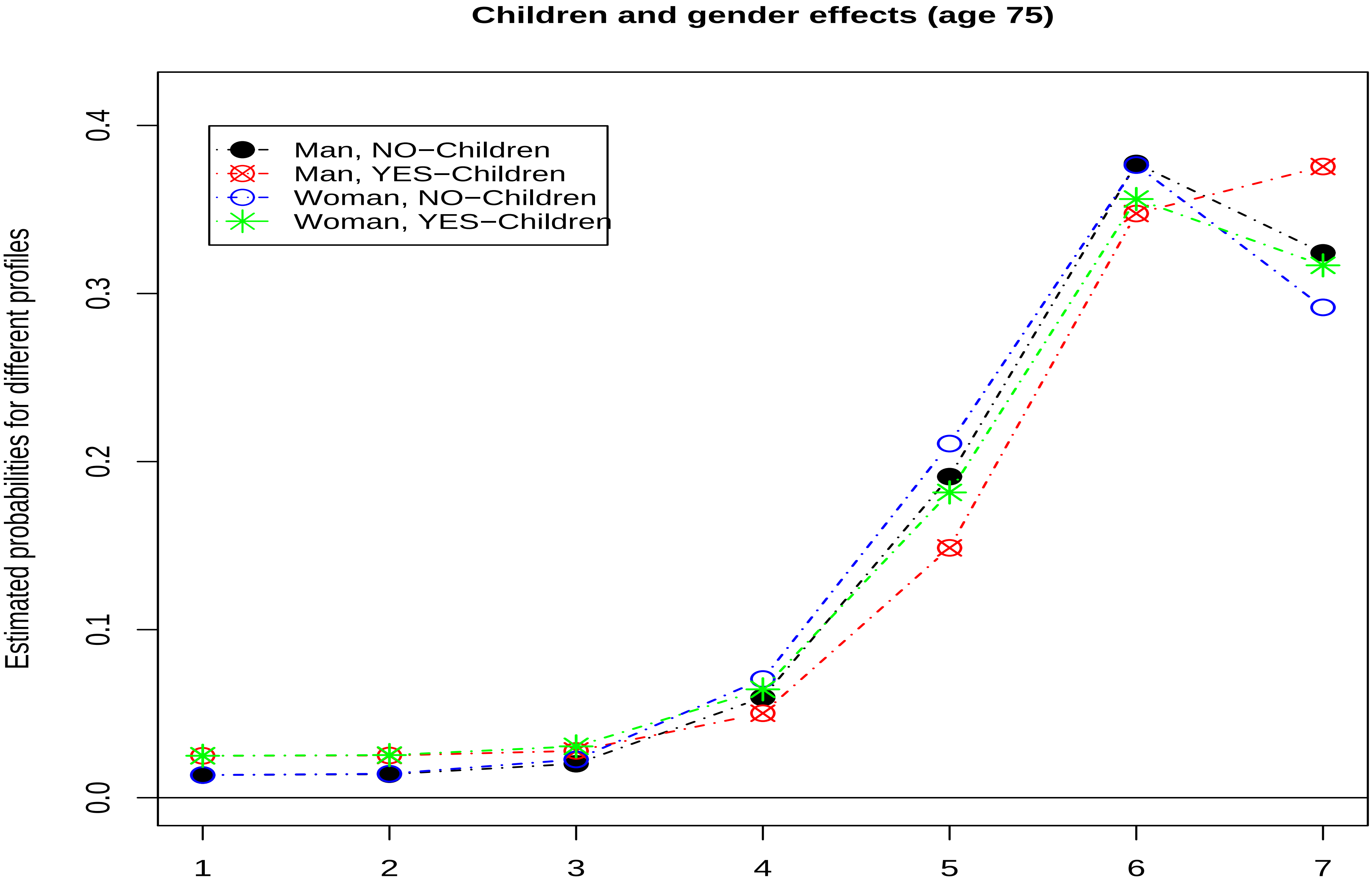
➤ Such models make possible to generate predictions of responses for given profiles of respondents. Indeed, similar considerations may be deduced by comparing the parameters but they become self-evident if one observes the whole probability distributions.

Profiles	Covariates	Parameters estimates
$\mathcal{A}$	(woman, single, no-worker, 30 years old)	$\pi_{\mathcal{A}} = 0.840, \ \xi_{\mathcal{A}} = 0.185$
$\mathcal{B}$	(man, married, worker, 50 years old)	$\pi_{\mathcal{B}} = 0.805, \ \xi_{\mathcal{B}} = 0.167$

$Pr\,(R = r)$	Profile $\mathcal{A}$	Profile $\mathcal{B}$
$(R = 1)$	0.023	0.028
$(R = 2)$	0.023	0.028
$(R = 3)$	0.030	0.031
$(R = 4)$	0.070	0.054
$(R = 5)$	0.195	0.152
$(R = 6)$	0.360	0.344
$(R = 7)$	0.298	0.364



Interaction effects of gender and children (age 75)



Part IV

Further developments

Extended CUB models

➤ CUB models have been further generalized for taking the possible effect of atypical situations into account (Iannario, 2008). Sometimes, these are derived by *shelter choices*, which represent categories frequently selected by respondents in order to avoid more elaborate decisions.

➤ Specifically, for given m and $c \in \{1, 2, \dots, m\}$, an *extended CUB model* is defined by:

$$p_r(\boldsymbol{\theta}) = \pi_1 \binom{m-1}{r-1} \xi^{m-r}(1-\xi)^{r-1} + \pi_2 \frac{1}{m} + (1-\pi_1-\pi_2)\, D_r^{(c)}\,, \quad r = 1, 2, \dots, m,$$

where $\boldsymbol{\theta} = (\pi_1,\ \pi_2,\ \xi)'$ is the parameter vector characterizing this distribution and $D_r^{(c)}$ is a degenerate random variable whose probability mass is concentrated at $r = c$, that is:

$$D_r^{(c)} = \begin{cases} 1, & \text{if } r = c; \\ 0, & \text{otherwise.} \end{cases}$$

➤ Identifiability of extended CUB models requires $m > 4$.

Extended models and covariates for *shelter choices*

- It should be evident that, in some circumstances –if one avoids considering this component– **CUB** model estimates are biased and inefficient; thus, fitting and predictions are not satisfactory.

- Instead, a remarkable feature of the extended model is that parameter $\delta = 1 - \pi_1 - \pi_2$ measures the added relative contribution of the *shelter choice* at $y = c$ with respect to the standard version of the model.

- Since its significance may be tested via standard asymptotic inference, extended **CUB** models may check the effective relevance of the presence of a *shelter choice*.

- Finally, a recent generalization includes also covariates in models with *shelter choices*.

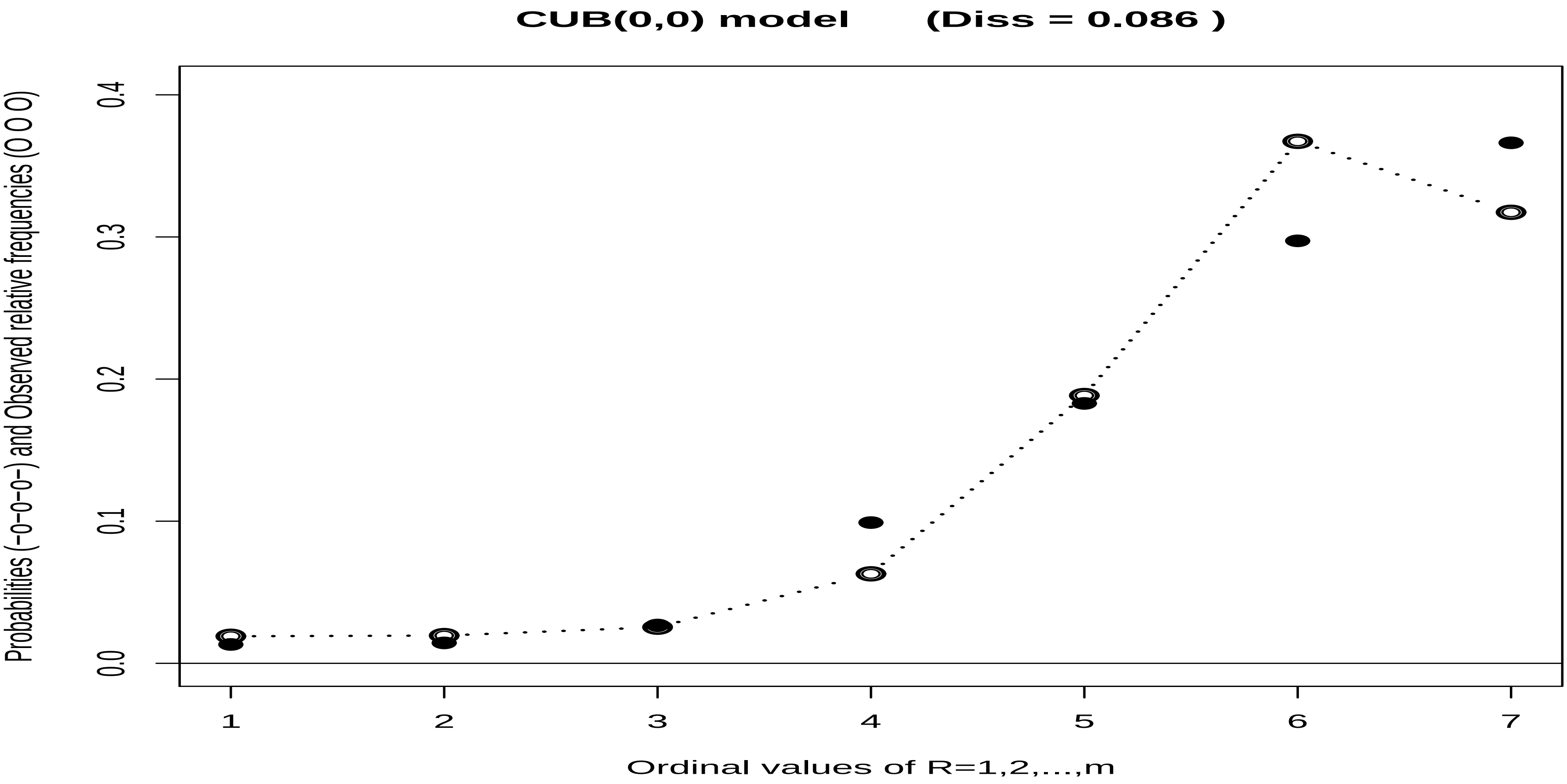
Subjective survival probability and *shelter choices*

- ➤ Subjective survival probabilities to age 75 suggest a modal value at the level “high” (the last but one class) whereas observed distributions are peaked at the level “almost sure” (the last class); notice that dissimilarity measure is $Diss = 0.087$.

- ➤ The last class is the equivalent of (almost) certainty to be alive to 75 years, and thus it exerts a peculiar appeal in the perception of respondents.

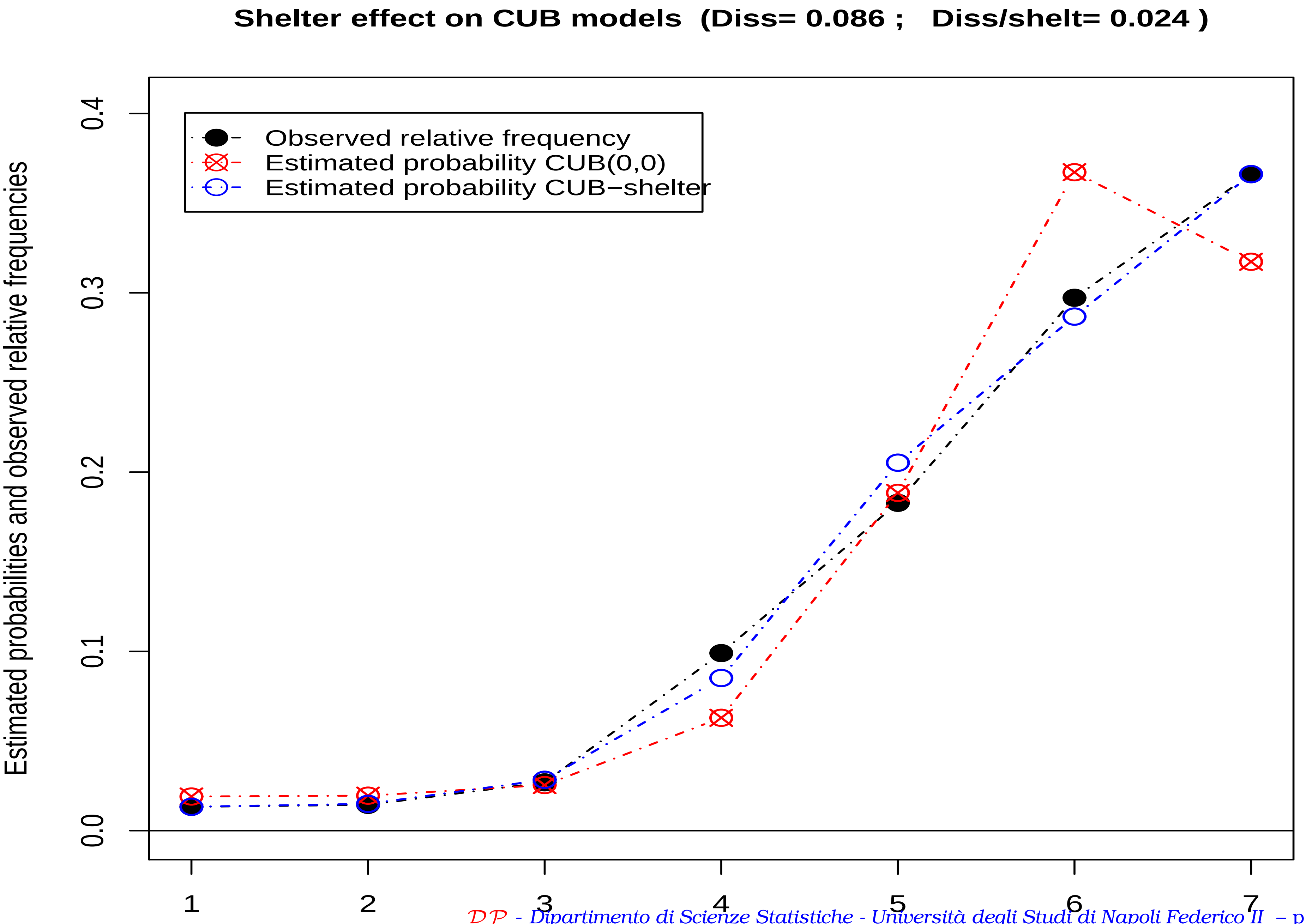
- ➤ Then, this class may be considered a *shelter choice* and extended **CUB** models with $c = 7$ may be fitted in order to estimate their importance.

Subjective survival probability and *shelter choices*

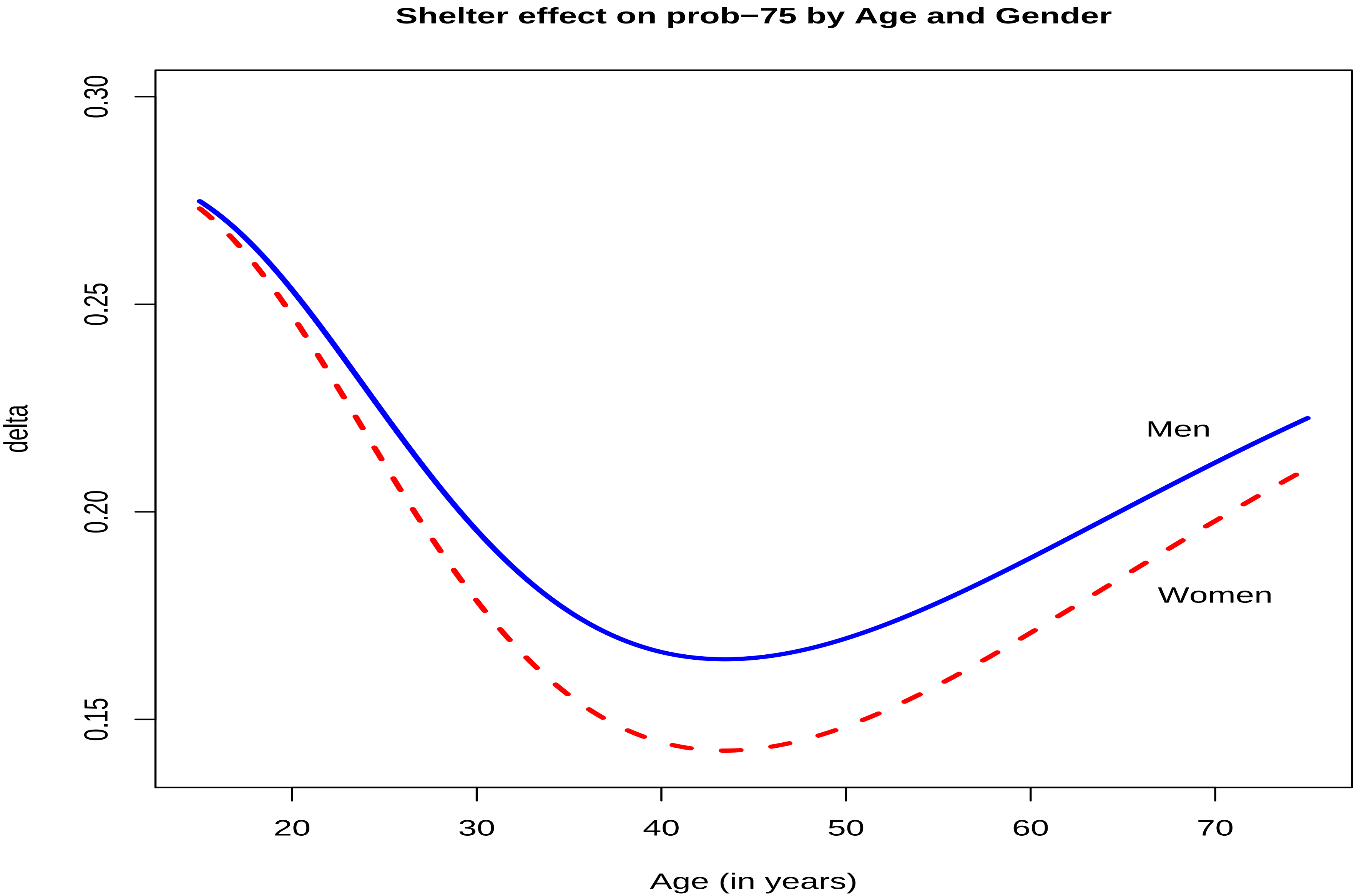


Clusters	n	$\hat{\delta}$	$\hat{\xi}$	*Diss*
All	20184	0.191 *(0.006)*	0.219 *(0.002)*	0.024
Female	10860	0.185 *(0.007)*	0.231 *(0.003)*	0.031
Male	9324	0.194 *(0.009)*	0.205 *(0.004)*	0.016

Improved fit by using CUB models with shelter effect



Shelter effect with respect to age and gender



CUB models with subjects’ and objects’ covariates

- In addition, objects’ covariates may be introduced and thus, similarly to other contexts, CUB models may include *choices’ covariates* and *choosers’ covariates*.

- We denote by $\boldsymbol{z}_k = (z_{k1}, z_{k2}, \dots, z_{kH})$ the row vector of available characteristics for a given k-th object, $k = 1, 2, \dots, K$, and list the H categories of K objects in the matrix $\boldsymbol{Z} = \{z_{kh},\, k = 1, 2, \dots, K;\; h = 1, 2, \dots, H\}$.

- Then, we modify the systematic components:

$$
\begin{aligned}
\pi_{ik} = (\pi \mid \boldsymbol{y}_i, \boldsymbol{z}_k) &= \frac{1}{1 + \exp(-(\boldsymbol{y}_i\, \boldsymbol{\beta} - \boldsymbol{z}_k\, \boldsymbol{\nu}))}\,; \\
\xi_{ik} = (\xi \mid \boldsymbol{w}_i, \boldsymbol{z}_k) &= \frac{1}{1 + \exp(-(\boldsymbol{w}_i\, \boldsymbol{\gamma} - \boldsymbol{z}_k\, \boldsymbol{\eta}))}\,;
\end{aligned}
$$

and $\boldsymbol{\nu} = (\nu_1, \nu_2, \dots, \nu_H)'$ and $\boldsymbol{\eta} = (\eta_1, \eta_2, \dots, \eta_H)'$ are further parameters to be estimated.

- In this way, π_{ik} (ξ_{ik}) is related to ***uncertainty*** (***feeling***) expressed by the i-th subject, whose profile is specified by $\boldsymbol{y}_i$ (by $\boldsymbol{w}_i$) when faced to the k-th object, whose characteristics are specified by $\boldsymbol{z}_k$.

- This approach is more demanding since the dimension of responses is now increased from n to $n\,K$.

Models for Item Response Theory

- ***Item Response Theory*** (IRT) is a well known and consolidated approach in Psychology, Sociology, Marketing and Medicine applied when a set of K items are submitted to a sample of n respondents. It is has been originated by the (dichotomous) basic Rasch model:

$$P_r\left(R_{ik}=1\right)=\frac{1}{1+\exp\{-(\gamma_i-\nu_k)\}}\,,$$

where γ_i (=*subject's ability*) and ν_k (=*item's difficulty*) are estimated by means of the sample data set $\boldsymbol{R}=\{r_{ik},\, i=1,2,\ldots,n;\; k=1,2,\ldots,K\}$.

- Of course: $P_r\left(R_{ik}=0\right)=1-P_r\left(R_{ik}=1\right)$.

- Several variants and extension for ordinal data are now available in the literature in order to generalize Rasch models for ordinal data and for including covariates.

CUB models and IRT paradigm

- If we consider **the item as an object**, then the following **CUB** model:

$$Pr(R = r_{ik} \mid \boldsymbol{y}_i;\, \boldsymbol{w}_i) = \pi \binom{m-1}{r_{ik}-1} \xi_{ik}^{m-r_{ik}} (1-\xi_{ik})^{r_{ik}-1} + (1-\pi)\left(\frac{1}{m}\right),$$

for $i = 1, 2, \ldots, n;\ k = 1, 2, \ldots, K$ and with a *systematic component*:

$$\xi_{ik} = \frac{1}{1+\exp\{-(\boldsymbol{w}_i\,\boldsymbol{\gamma} - \delta_k)\}} = \frac{1}{1+\exp\{-(\gamma_i - \nu_k)\}},$$

may be seen as a generalization of IRT paradigm.

- Specifically, Rasch model is a *special case* of the previous **CUB** model when $m = 2$ and $\pi = 0$.

- Actually, in this way, we get a multivariate version of **CUB** modelling approach.

- Notice that **CUB** model paradigm adds further visual aids to standard representations of IRT.

Generalized CUB models

- We are currently working to estimation routines and statistical inference for the class of ***Generalized CUB models***. For a given m and $c \in \{1, 2, \ldots, m\}$, they are defined, for $r = 1, 2, \ldots, m$, by:

$$\boxed{Pr(R=r\mid \boldsymbol{x}_i) = \pi_{1i}\underbrace{\left[\binom{m-1}{r-1}(1-\xi_i)^{r-1}\,\xi_i^{m-r}\right]}_{\textit{feeling}} + \pi_{2i}\underbrace{\left[\,\frac{1}{m}\,\right]}_{\textit{uncertainty}} + (1-\pi_{1i}-\pi_{2i})\underbrace{\left[\,D_r^{(c)}\,\right]}_{\textit{shelter effect}}}$$

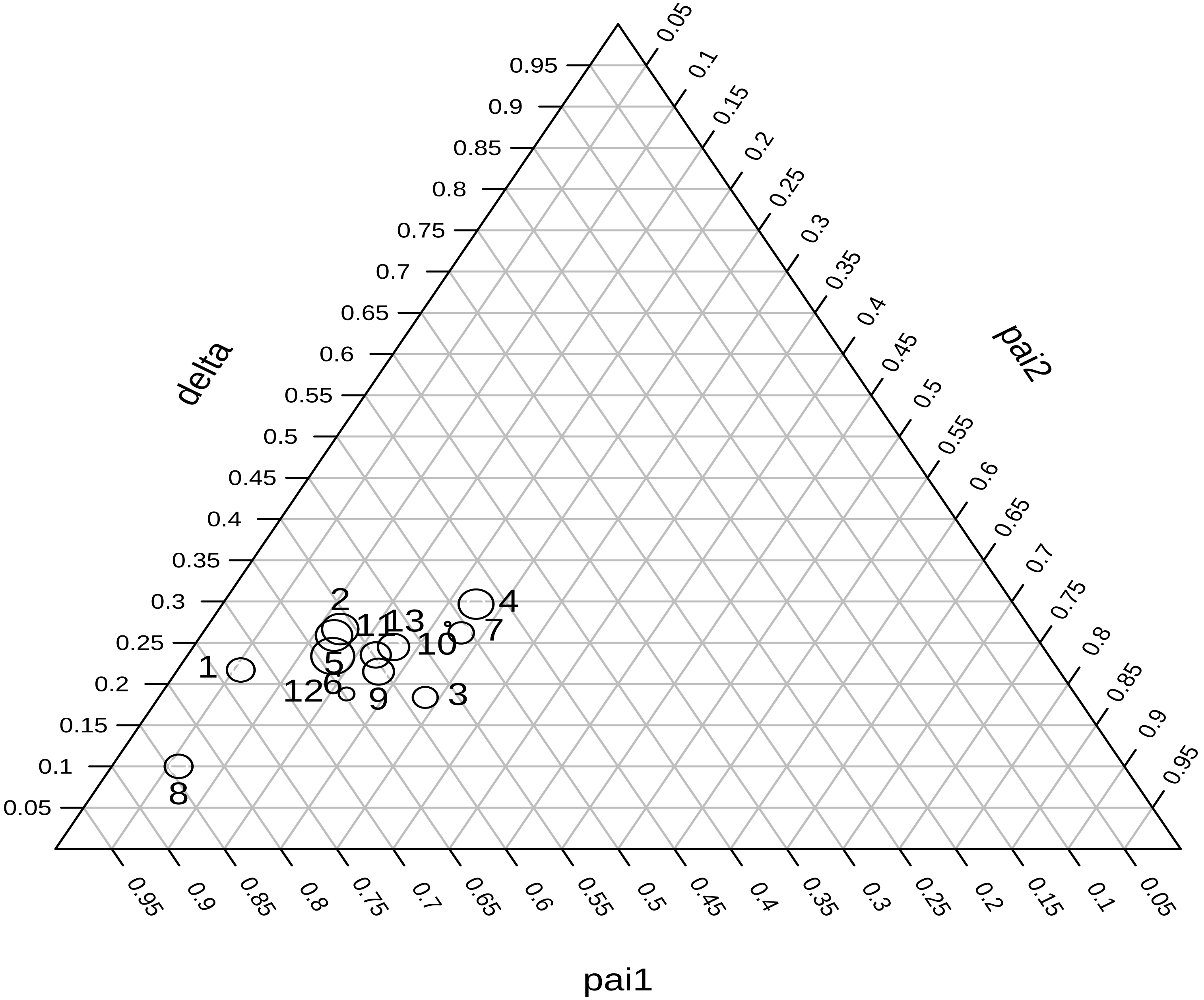
- The *systematic components*, for any $i = 1, 2, \ldots, n$, are:

$$\pi_{1i}=\pi_1(\beta;\boldsymbol{y}_i)=\frac{1}{1+e^{-\boldsymbol{y}_i\,\boldsymbol{\beta}}}\;;\xi_i=\xi(\gamma;\boldsymbol{w}_i)=\frac{1}{1+e^{-\boldsymbol{w}_i\,\boldsymbol{\gamma}}}\;;\pi_{2i}=\pi_2(\omega;\boldsymbol{z}_i)=\frac{1}{1+e^{-\boldsymbol{z}_i\,\boldsymbol{\omega}}}\,,$$

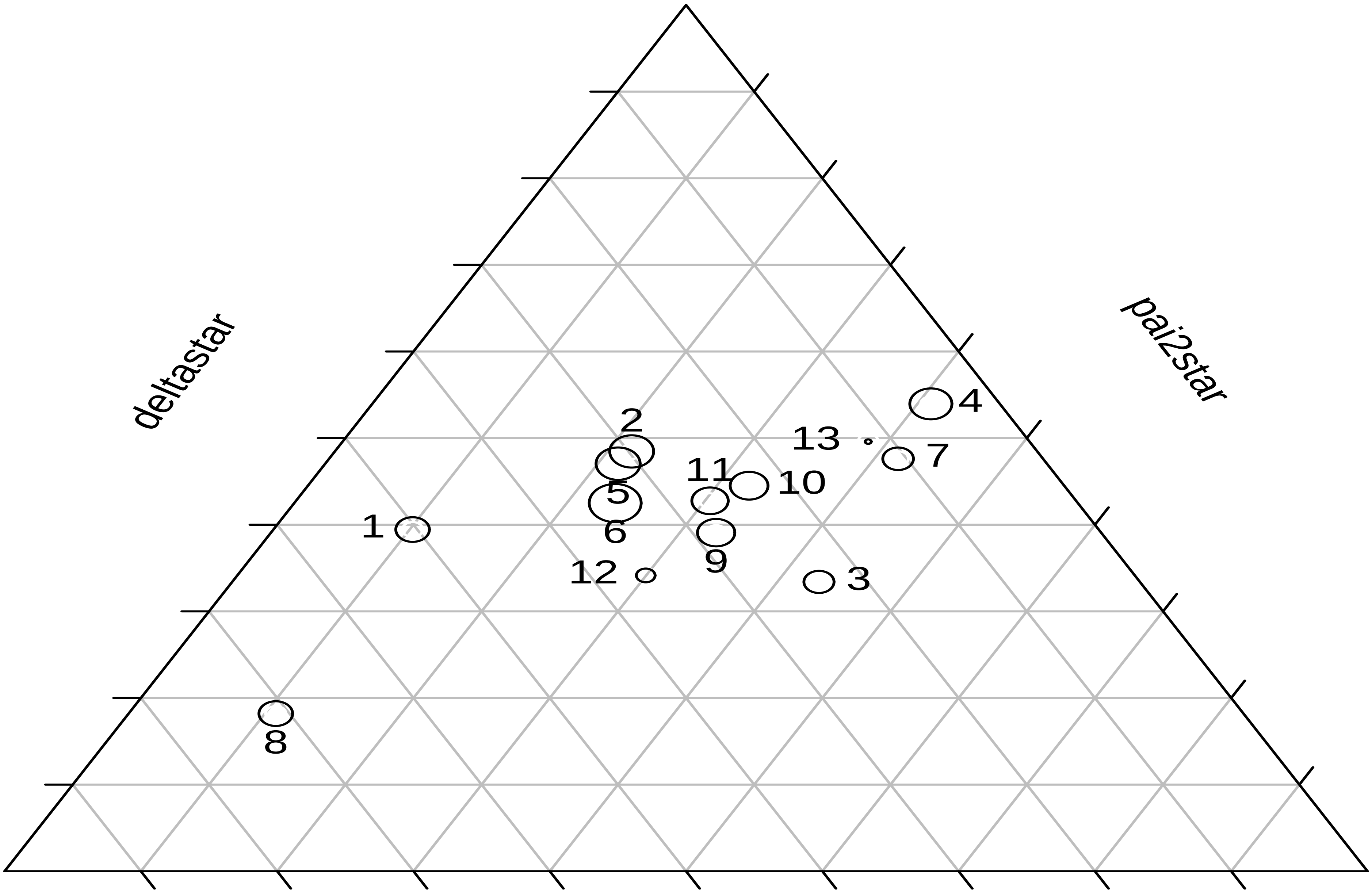
and $\boldsymbol{x}_i = (\boldsymbol{y}_i, \boldsymbol{w}_i, \boldsymbol{z}_i)'$, for $i = 1, 2, \ldots, n$, are the rows of a convenient set of covariates explaining *uncertainty, feeling and shelter choices*, respectively.

- We will denote this class as `GeCUB (p,q,s)` models, where $p = 0, 1, \ldots; q = 0, 1, \ldots;$ $s = -1, 0, 1, \ldots$. This class includes all previous CUB models, it requires the estimation of $(p + q + s + 3)$ parameters, and if $s = -1$ there is no *shelter effect*. Moreover, identifiability implies that $\boldsymbol{Y}$ and $\boldsymbol{Z}$ matrices must not coincide.

Visualization of generalized CUB models



Visualization of generalized CUB models (enlarged)



Open issues and concluding remarks

- Optimize numerical procedures for efficient computations.
- Introduce multilevel **CUB** models.
- Extend **CUB** models in a multivariate setting.
- Develop fitting measures and residual analyses.
- Detect quick methods for selection of significant covariates.
- Improve visual representations in the parametric space.
- . . . .