UNIVERSITA' DEGLI STUDI DI BRESCIA

Dipartimento di Economia e Management - Sede Santa Chiara

Statistical methods and models for ordinal data

An introduction to model-based approaches

Domenico Piccolo

Università degli Studi di Napoli Federico II

domenico.piccolo@unina.it

Some pleasant reasons . . . . . .

➤ I am both ***happy*** and ***proud*** to be here for some nice reasons . . .

1. There is no scientific groups which invest so much in CUB models (2003) and their many variations as Brescia's statisticians:
   - NL-CUB, DK-CUB, CUM,. . . (*Zuccolotto, Manisera, Carpita, Brentari, Golia, Vezzoli*, . . .)
   - **`https://bodai.unibs.it/cub/`**: a substantive archive and a rich mine of CUB models literature
   - young researchers in Brescia are investing on those topics
2. The only mixture alternative to CUB models are BOS (Binary Ordinal Search) models which have been introduced by C. Biernacki and **J. Jacques** (2013).
3. My talk will concern ***the meaning and the role of uncertainty*** in CUB models: a topic regularly discussed with *Rosaria Simone* and which has been stimulated by the constructive criticism of Brescia's statisticians.

At the beginning of the idea .... (2003–. . .)

- During my (long) teaching career, I have observed the behaviour of a huge number of students while they were filling out questionnaires made up of sequences of items.

- The answers often required an ordinal choice over a scale ranging from *"completely unsatisfied"* up to *"completely satisfied"*, as an instance.

- I observed that **firstly** the students used to place their pencils in one direction of the scale (left, center, right). **Then**, before selecting a unique box, they quite often manifested a sort of "swing" with respect to their final selection ....(with varying time: from *well persuaded* to *quite hesitant*).

- Depending on subjects and/or topics and/or context, each of these two "movements" along the scale could be *tiny/broad* to *enhance/obscure* the other.

- Psychologists confirmed me this empirical evidence: indeed, the selection on an ordinal scale may be considered as a mixture of two internal motions, which I called ***feeling*** and ***uncertainty***, just to simplify the discussion.

- As a consequence, *Piccolo* (2003) introduced a parsimonious mixture model to capture such a generating process and explored its main characteristics.

- After a long and troubled refereeing process, *D'Elia and Piccolo* (2005) became the international most quoted standard reference for CUB models (*without covariates*). . . although D'Elia (2003) proposed and applied a general CUB model ***with covariates***.

The ”core” of CUB models specification

➤ Given a random sample of ordinal ratings $R_i \in \{1, \dots, m\}$, $m > 3$, and $i = 1, 2, \dots, n$, a **CUB model** (= ***C***ombination of discrete ***U***niform and shifted ***B***inomial random variables) is defined by:

1. A ***stochastic component***:

$$Pr(R_i = r \,|\, \boldsymbol{x}_i,\, \boldsymbol{w}_i) = \pi_i \underbrace{\left[\binom{m-1}{r-1}(1-\xi_i)^{r-1}\xi_i^{m-r}\right]}_{\textit{feeling}} + (1-\pi_i) \underbrace{\left[\frac{1}{m}\right]}_{\textit{uncertainty}}$$

for $r = 1, 2, \dots, m$, where $\pi_i \in (0, 1]$ and $\xi_i \in [0, 1]$.

2. Two ***systematic components***:

$$\begin{cases} logit(1-\pi_i) & = & -\boldsymbol{x}_i\boldsymbol{\beta} = -\beta_0 - \beta_1\, x_{i1} - \dots - \beta_p\, x_{ip}\,; \\ logit(1-\xi_i) & = & -\boldsymbol{w}_i\boldsymbol{\gamma} = -\gamma_0 - \gamma_1\, w_{i1} - \dots - \gamma_q\, w_{iq}\,. \end{cases}$$

where $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are the parameters to be estimated, and $\boldsymbol{x}_i$ and $\boldsymbol{w}_i$ are the *row vectors* containing the values of the covariates of the i-th subject, suitable to explain π_i and ξ_i, respectively.

Meaning and interpretation of the parameters

- Some people argue that CUB models are a mixture of two sub-populations ("motivated" and "random" respondents, respectively): this may be correct but **this circumstance does not motivate my proposal**.

- Each respondent acts with a ***propensity*** to adhere to a thoughtful and to a completely uncertain choice, whose weight is measured by (π_i) and $(1-\pi_i)$, respectively.

- In case of a rating question/item with positive wording:
  - $(1-\xi_i)$ may be interpreted as a ***measure of preference*** towards the item.
  - $(1-\pi_i)$ is a ***weight of the uncertainty*** included in the responses.

- A noticeable aspect of CUB models is the ***direct link*** between ***subjects' covariates*** and ***parameters*** $(\pi_i \, \xi_i)$.

- Thus, the link is **not** derived by *expectations*; then, **CUB** ***models are not based on numerical scores for the responses***.

CUB model as a discrete random variable

➤ *Unlike cumulative models*, for a given item, CUB models are able to summarize the n ordinal responses without the need to specify covariates.

➤ Thus, a CUB model (**without** covariates *or* **conditioned** to some values of the covariates) is a discrete random variable with probability mass function:

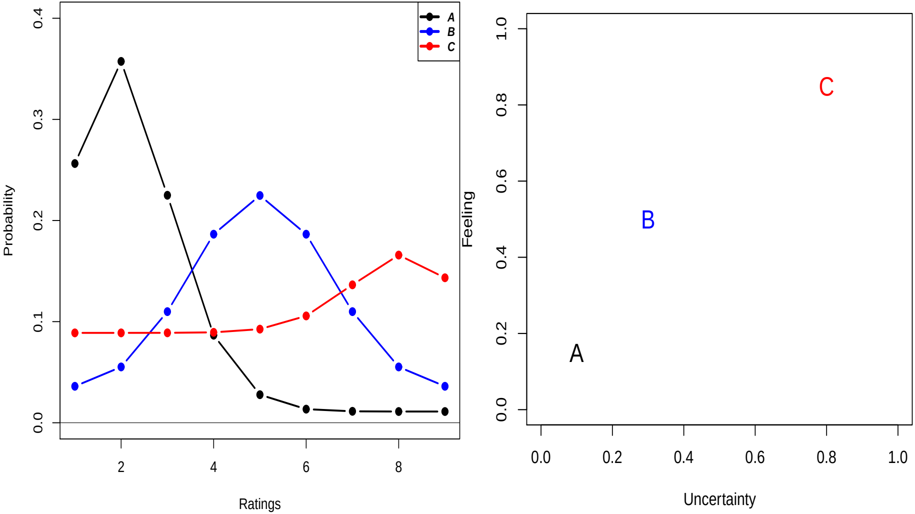
$$\boldsymbol{Pr\,(R=r)\;=\;\pi\binom{m-1}{r-1}(1-\xi)^{r-1}\xi^{m-r}\;+\;(1-\pi)\,\frac{1}{m}\,,\qquad r=1,\dots,m\,.}$$

where $\pi \in (0, 1]$ and $\xi \in [0, 1]$ specify the *parameter space* as a left-opened unit square.

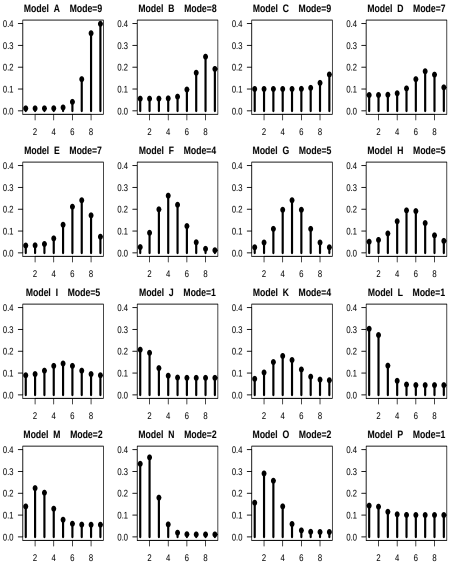
➤ Some useful statistical consequences:

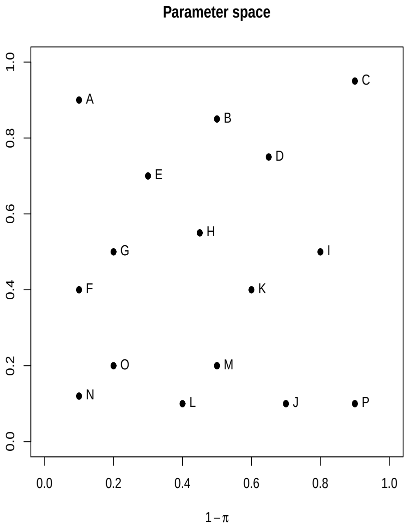
- comparing and ranking items
- detecting latent classes
- managing/imputing missing values in ordinal data
- model-based composite indicators of ratings and preferences
- dynamic and/or spatial differences of subjects' behaviour
- immediate usage in *machine learning algorithms*

One of the most attractive feature: CUB models visualization

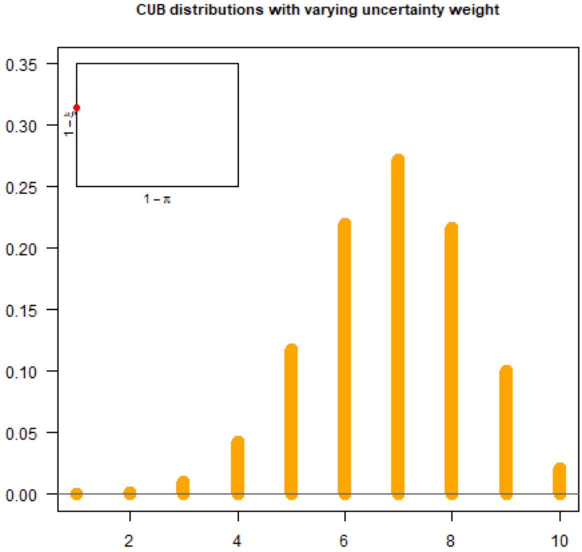


CUB models and parameter space

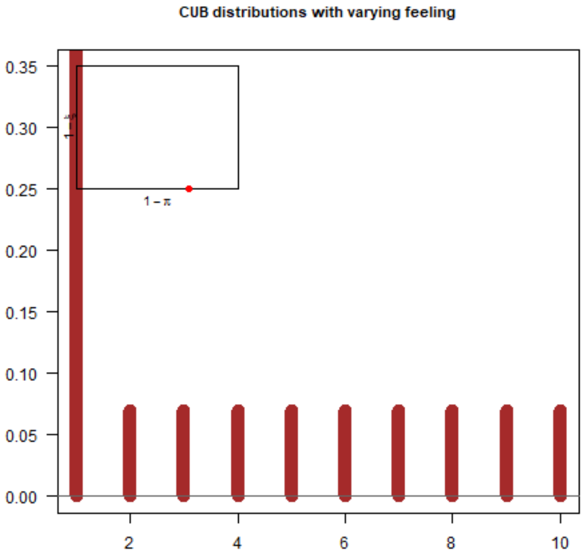




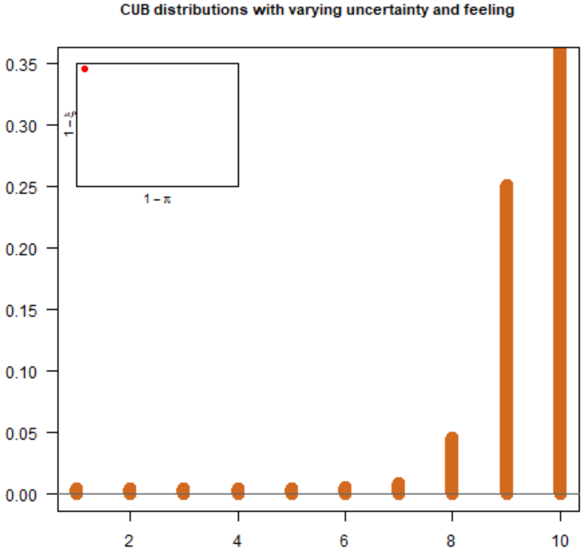
CUB distributions with $\xi = 0.35$ and **varying uncertainty**



CUB distributions with $\pi = 0.30$ and **varying feeling**



CUB distributions for **varying uncertainty** ***and*** **feeling**



CUBE models with covariates

- A **CUBE *model with covariates*** (Piccolo, 2015) is defined by:

$$\begin{cases} Pr\,(R=r_i) = \pi_i\ \beta e(\xi_i, \phi_i)\ +\ (1-\pi_i)\,\dfrac{1}{m}\,; \\[2ex] \pi_i = \dfrac{1}{1+e^{-\boldsymbol{x}_i\,\boldsymbol{\beta}}}\,; \quad \xi_i = \dfrac{1}{1+e^{-\boldsymbol{w}_i\,\boldsymbol{\gamma}}}\,; \quad \phi_i = \log\left(\boldsymbol{z}_i\,\boldsymbol{\alpha}\right)\,; \end{cases} \qquad i=1,2,\dots,n$$

where the Beta-Binomial distribution is:

$$\beta e(\xi_i, \phi_i) = \binom{m-1}{r_i-1}\,\frac{\displaystyle\prod_{k=1}^{r_i}\left[1-\xi_i+\phi_i\,(k-1)\right]\ \prod_{k=1}^{m-r_i+1}\left[\xi_i+\phi_i\,(k-1)\right]}{\left[1-\xi_i+\phi_i\,(r_i-1)\right]\ \left[\xi_i+\phi_i(m-r_i)\right]\ \displaystyle\prod_{k=1}^{m-1}\left[1+\phi_i\,(k-1)\right]}\,.$$

- If $\phi_i \to 0$ the (shifted) Beta-binomial tends to the (shifted) Binomial distribution.

- **CUB models are nested into CUBE models**.

An inclusive perspective (2016)

- A *GEneralized Mixture model with uncertainty* (GEM) is defined as follows:

$$Pr\,(R_i = j \mid \boldsymbol{\theta}) = \pi_i\, Pr\left(Y_i = j \mid \boldsymbol{t}_i^{(\gamma)}, \boldsymbol{\Psi}\right) + (1-\pi_i)\, Pr\,(V_i = j)\ ,$$

for $i = 1, \dots, n$ and $j = 1, \dots, m$, where $\pi_i = \pi(\boldsymbol{t}_i^{(\boldsymbol{\beta})}, \boldsymbol{\beta}) \in (0,\ 1]$ are introduced to weight the two components and $\boldsymbol{t}_i^{(\gamma)} \in \boldsymbol{T}^{(\gamma)}$ and $\boldsymbol{t}_i^{(\boldsymbol{\beta})} \in \boldsymbol{T}^{(\boldsymbol{\pi})}$ include the values of the selected covariates for the i-th subject.

- The probability distribution of the *feeling component* Y_i is

$$\begin{cases} Pr\left(Y_i = j \mid \boldsymbol{\gamma},\ \boldsymbol{t}_i^{(\gamma)}\right)\ , & \text{if specified via a } \textbf{discrete distribution}; \\ F_{Y_i^*}(\tau_j; \boldsymbol{\gamma},\ \boldsymbol{t}_i^{(\gamma)}) - F_{Y_i^*}(\tau_{j-1}; \boldsymbol{\gamma},\ \boldsymbol{t}_i^{(\gamma)})\,, & \text{if specified via a } \textbf{latent variable distribution}; \end{cases}$$

where $F_{Y_i^*}(\tau_j; \boldsymbol{\gamma},\ \boldsymbol{t}_i^{(\gamma)}) = Pr\left(Y_i^* \le \tau_j \mid \boldsymbol{\gamma},\ \boldsymbol{t}_i^{(\gamma)}\right)$ is the distribution function of the latent variable Y_i^*.

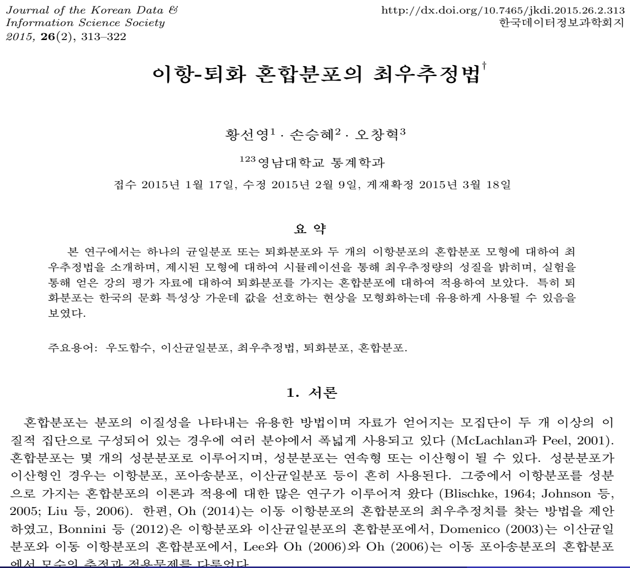
Current applications of CUB models

- ***Nature of responses***: Ratings, marginal rankings, multivariate ratings
- ***Content of the item***: Preference, mood, agreement, likeness, agreeableness, judgements, perception, cognition, priority, assessment, similarity, changeability, attraction, qualitative distance, fear, discrimination, worry, anxiety, pain, distress, awkwardness, ...
- ***Fields of interest***: Marketing surveys, Sensory analysis, Food packaging, Tourism sustainability, Drug effectiveness, Severity of a disease, Consumer preferences, Service evaluations, Political position on a left-right ideological scale, Words synonymy, Quality of life, Job satisfaction, Stress analysis, Work discrimination, Social media reliability, Advertisement efficacy, Politicians approval, Company climate, Video recommendation, Privacy intrusion, Pharmacokinetics, Team ability, Crowd sourcing, Risk perception, Adolescent abuse substances, Cognitive dissonance, ...
- ***Widespread***: **Italy** (Naples, Brescia, Cosenza, Bergamo, Padova, Vicenza, Ferrara, Palermo, Sassari, Turin, Bari, Rome, Florence, Benevento, Milan, Pavia, Potenza, Catania, . . .), **Germany**, **Switzerland**, **France**, **Netherlands**, **Israel**, **USA**, **Argentina**, **Malaysia**, **South-Korea**, **China**, **Brasil**.

A list of references .... continuously increasing

- Piccolo D. (2003). On the moments of a mixture of uniform and shifted binomial random variables. *Quaderni di Statistica*, **5**, 85–104.
- D'Elia A., Piccolo D. (2005). A mixture model for preference data analysis. *Computational Statistics & Data Analysis*, **49**, 917–934.
- .... *about 200 publications* ....
- Piccolo D., Simone R. (2019). The class of CUB models: statistical foundations, inferential issues and empirical evidence. *Statistical Methods & Applications*, **28**, 389–435; with discussion (pp.437–475) and rejoinder (pp.477–493). (**(with hundreds of references)**
- .... *recent publications* ....
- Simone R. (2023). Uncertainty Diagnostics of Binomial Regression Trees for Ordered Rating Data. *Journal of Classification*, 40: 79–103, Springer.
- Venson A.H., Jacinto P.A. and Sbicca, A. (2023). Cognitive Dissonance in the Self-assessed Health in Brazil: A CUB Model Analysis Using 2013 National Health Survey Data. *Integrative Psychological and Behavioral Science*, `https://doi.org/10.1007/s12124-023-09768-x`. **19 May 2023**

Sometimes. . . I cannot read a paper on CUB models



Journal of the Korean Data &
Information Science Society
2015, **26**(2), 313–322

http://dx.doi.org/10.7465/jkdi.2015.26.2.313
한국데이터정보과학회지

이항-퇴화 혼합분포의 최우추정법[†]

황선영[1] · 손승혜[2] · 오창혁[3]

[123]영남대학교 통계학과

접수 2015년 1월 17일, 수정 2015년 2월 9일, 게재확정 2015년 3월 18일

요 약

본 연구에서는 하나의 균일분포 또는 퇴화분포와 두 개의 이항분포의 혼합분포 모형에 대하여 최우추정법을 소개하며, 제시된 모형에 대하여 시뮬레이션을 통해 최우추정량의 성질을 밝히며, 실험을 통해 얻은 강의 평가 자료에 대하여 퇴화분포를 가지는 혼합분포에 대하여 적용하여 보았다. 특히 퇴화분포는 한국의 문화 특성상 가운데 값을 선호하는 현상을 모형화하는데 유용하게 사용될 수 있음을 보였다.

주요용어: 우도함수, 이산균일분포, 최우추정법, 퇴화분포, 혼합분포.

1. 서론

혼합분포는 분포의 이질성을 나타내는 유용한 방법이며 자료가 얻어지는 모집단이 두 개 이상의 이질적 집단으로 구성되어 있는 경우에 여러 분야에서 폭넓게 사용되고 있다 (McLachlan과 Peel, 2001). 혼합분포는 몇 개의 성분분포로 이루어지며, 성분분포는 연속형 또는 이산형이 될 수 있다. 성분분포가 이산형인 경우는 이항분포, 포아송분포, 이산균일분포 등이 흔히 사용된다. 그중에서 이항분포를 성분으로 가지는 혼합분포의 이론과 적용에 대한 많은 연구가 이루어져 왔다 (Blischke, 1964; Johnson 등, 2005; Liu 등, 2006). 한편, Oh (2014)는 이동 이항분포의 혼합분포의 최우추정치를 찾는 방법을 제안하였고, Bonnini 등 (2012)은 이항분포와 이산균일분포의 혼합분포에서, Domenico (2003)는 이산균일분포와 이동 이항분포의 혼합분포에서, Lee와 Oh (2006)와 Oh (2006)는 이동 포아송분포의 혼합분포에서 모수의 추정과 적용문제를 다루었다.

The very source of uncertainty

➤ Uncertainty ***is not*** the stochastic component related to the sampling experiment (so that different people generates different ratings).

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

- *Limited set of information, Knowledge/Ignorance* of properties and/or characteristics of the object/item to be evaluated.
- *Personal interest/Engagement* in activities related to the specific or related field of interest.
- *Amount of time* devoted to the response.
- *Operational mode* for responding: face-to-face, questionnaire form, telephone, mobile, PC, mail, Email, etc.
- *Nature of the scale* in terms of range and wording.
- *Tiredness or fatigue* for a correct comprehension of the wording.
- *Willingness to joke and fake*.
- *Lack of self-confidence* of the respondent.
- *Laziness/Apathy/Boredom* in the selection mechanism.
- ....

What is *really* uncertainty?

- In essence, uncertainty includes at least three points of view:

1. ***subjective indecision***: when we examine $1-\pi_i$, it is possible to consider it as a measure of personal indecision of the i-th respondent as a function of selected covariates.
2. ***heterogeneity***: when we analyse a global CUB model for the given item, it is possible to consider $1-\pi$ as a measure of heterogeneity of the responses.
3. ***predictability***: it is possible to consider π as a direct measure of predictability of the model with respect to two extremes:
   - *minimum* $\rightarrow$ responses follow a (pure) discrete Uniform distribution ($\pi \rightarrow 0$)
   - *maximum* $\rightarrow$ responses follow a (pure) Binomial distribution ($\pi = 1$)

CUB models are indeed with **varying** uncertainty

$$Pr(R_i = r | \boldsymbol{x}_i,\, \boldsymbol{w}_i) = \pi_i \underbrace{\left[\binom{m-1}{r-1}(\boldsymbol{1-\xi_i})^{r-1}\xi_i^{m-r}\right]}_{\textit{feeling distribution}} + (\boldsymbol{1-\pi_i}) \underbrace{\left[\frac{1}{m}\right]}_{\textit{uncertainty distribution}}$$

for $r = 1, 2, \ldots, m$, where $\pi_i \in (0, 1]$ and $\xi_i \in [0, 1]$, for $i = 1, 2, \ldots, n$.

➤ CUB models estimates the *weight of uncertainty* $1 - \pi_i$, **not** the parameters of the probability distribution assumed for the uncertainty.

➤ As a consequence, *if covariates are significant*, any CUB model assumes a **non-constant uncertainty** $1 - \pi_i$ ***which modifies with subjects*** $(i = 1, 2, \ldots, n)$ **not with categories** $(r = 1, 2, \ldots, m)$.

An application of CUB models with varying uncertainty............(1)

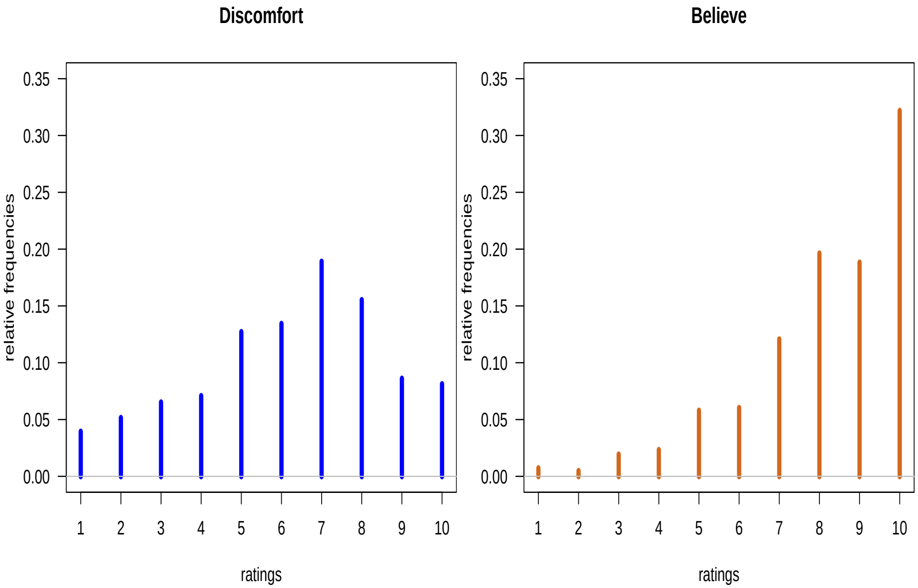
➤ A recent survey (8-22 May 2023) about the distress/discomfort (*disagio*) of University students has been planned at University of Basilicata.

➤ A large sample of students ($n = 1243$, where the population size is $N \simeq 6000$) answered on a 10 point Likert scale to several items. Two of them were:

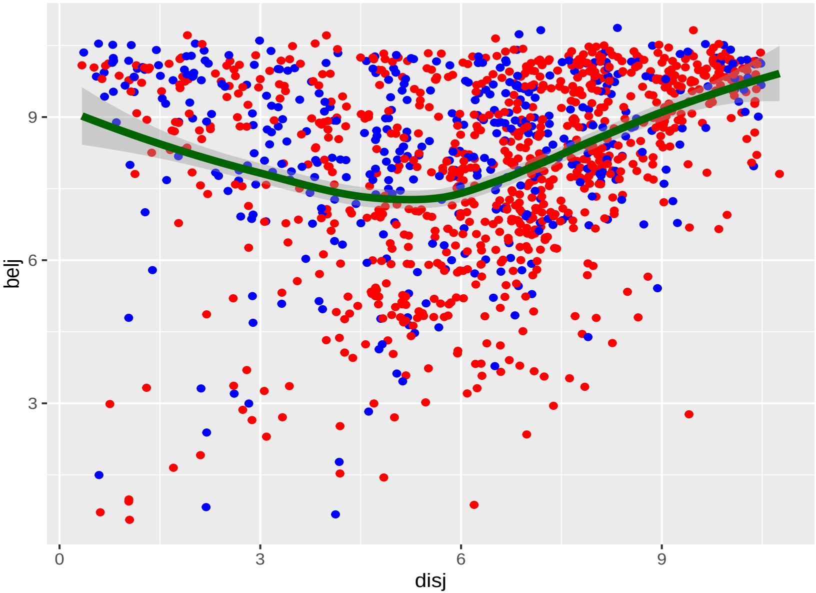
1. ***After carefully thinking, how strong is the personal discomfort/distress you are experiencing?*** ($Dist_i$)
2. ***How strongly do you believe in the answer you provided to the previous question regarding your level of discomfort/distress?*** (Bel_i)

➤ The opportunity to ask this pair of questions has been strongly supported by several colleagues (Carpita, Bartolucci, Pennoni, ...) in order to investigate *the role of uncertainty in* CUB *models*.

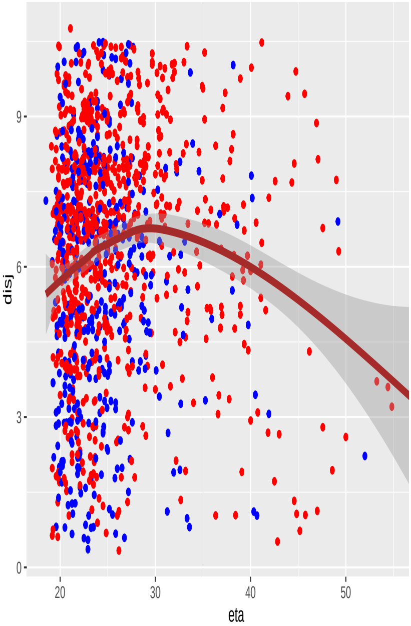
An application of CUB models with varying uncertainty . . . . (2)

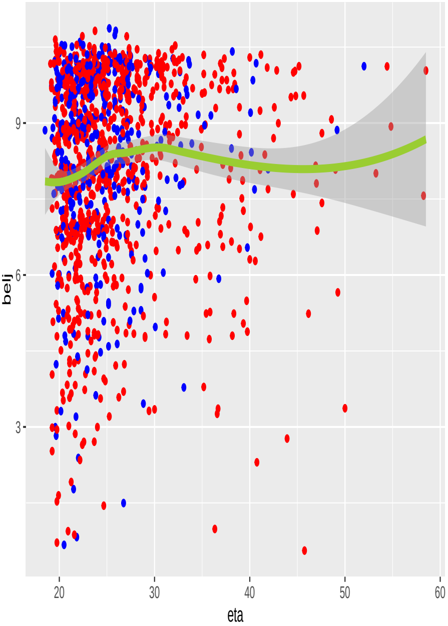


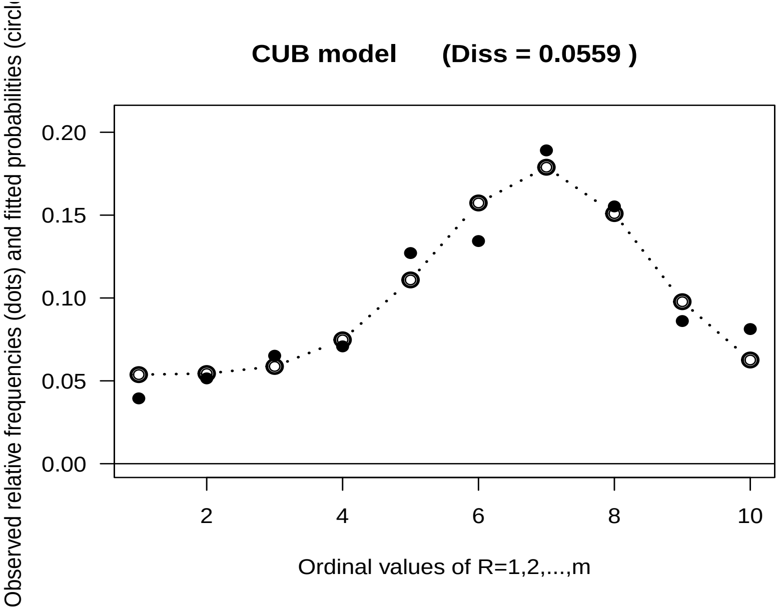
An application of CUB models with varying uncertainty . . . . (3)



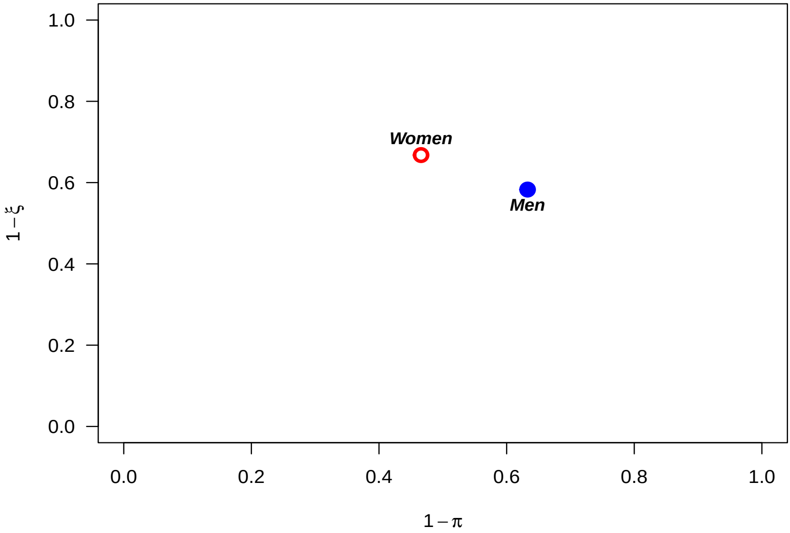
An application of CUB models with varying uncertainty . . . .(4)






CUB model (Diss = 0.0559)
Observed relative frequencies (dots) and fitted probabilities (circl
0.20
0.15
0.10
0.05
0.00
2
4
6
8
10
Ordinal values of R=1,2,...,m

An application of CUB models with varying uncertainty . . . . (6)



An application of CUB models with varying uncertainty . . . . (7)

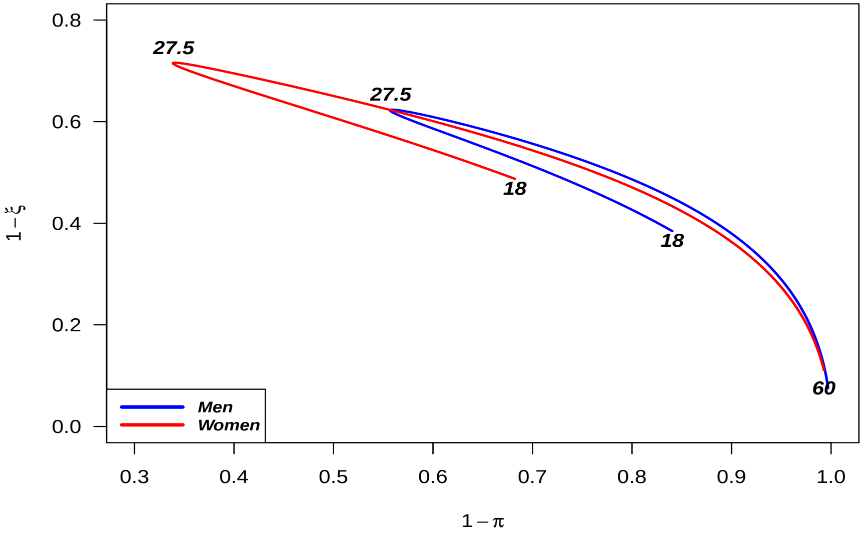
$$\boxed{Pr(\widehat{Dis_i} = r) = \hat{\pi}_i \underbrace{\left[\binom{m-1}{r-1}(1-\hat{\xi}_i)^{r-1}\hat{\xi}_i^{m-r}\right]}_{\textit{feeling distr.}} + (1-\hat{\pi}_i) \underbrace{\left[\frac{1}{m}\right]}_{\textit{uncertainty distr.}} \;,\quad r = 1, 2, \dots, m}$$

$$\begin{cases} logit(1-\hat{\pi}_i) = \underset{(33.209)}{97.271} - \underset{(20.073)}{58.856}\,[\log(Age_i)] + \underset{(3.028)}{8.925}\,[\log(Age_i)]^2 - \underset{(0.281)}{0.922}\,Gender_i \\ logit(1-\hat{\xi}_i) = -\underset{(15.726)}{57.122} + \underset{(9.617)}{34.635}\,[\log(Age_i)] - \underset{(1.469)}{5.203}\,[\log(Age_i)]^2 + \underset{(0.114)}{0.409}\,Gender_i \end{cases}$$

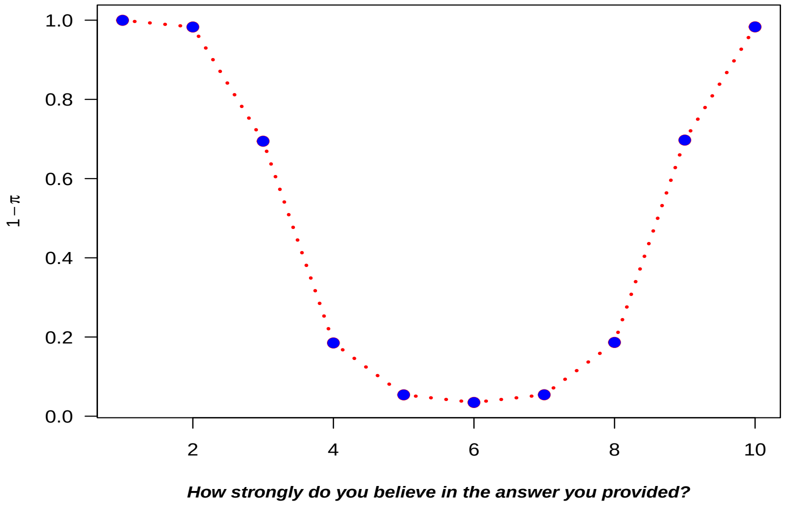
for $i = 1, 2, \dots, n$.

Models	Covariates of (π_i, ξ_i)	*Log-lik*	*BIC*
CUB	=====	-2717.616	5449.465
CUB	*Gender*	-2730.588	5489.677
CUB	*Age, Gender*	-2688.049	5433.028
CUB	*Believe, Gender*	-2577.284	5211.571
CUB	*Believe, Age, Gender*	-2561.832	5209.167

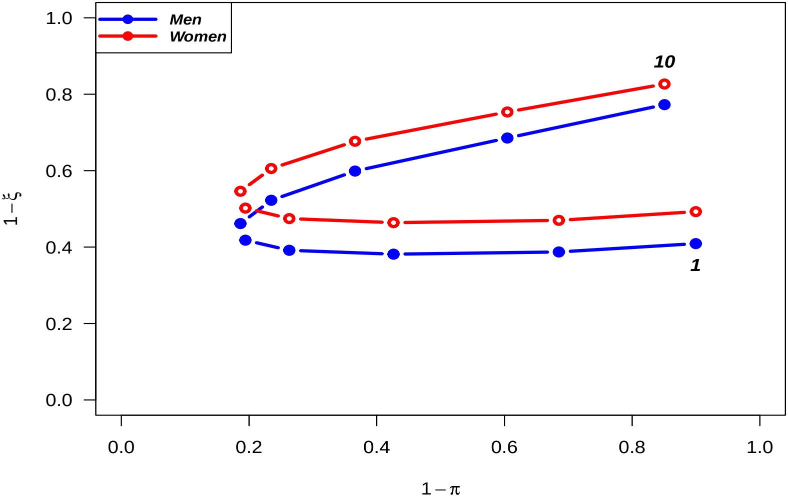
An application of CUB models with varying uncertainty . . . . (8)



An application of CUB models with varying uncertainty . . . . (9)



An application of CUB models with varying uncertainty . . . . (10)



A final consideration

- *We are not working with a single model, a collection of models, a variant of existing models.*

- *Indeed, we are proposing and implementing a whole framework (that is a "paradigm") based on the Generating Data Process of ratings.*

- *This process includes covariates if and only if their effects are significant to explain respondents' behaviour.*

- *The added value of CUB paradigm is a parsimonious model, a visualization feature, a straightforward interpretation of parameters and a direct relationship with subjects' and objects' covariates.*

- *Since statisticians are well aware of the role and importance of uncertainty in human decisions, CUB models may be considered as building blocks of more complex statistical specifications, that is a sort of benchmark to achieve better models . . . which in turn should ever be improved.*