The Stata module CUB for fitting mixture models for ordinal data

Christopher F. BAUM[1], Giovanni CERULLI[2], Francesca DI IORIO[3], Domenico PICCOLO[3], Rosaria SIMONE[3]

[1] Boston College
[2] IRCrES-CNR, Roma
[3] Università degli Studi di Napoli Federico II

November 15th, 2018

XV Convegno Italiano degli Utenti di STATA
Bologna

Outline

1. The statistical framework

2. Empirical evidence

3. The Stata module for CUB

Ordinal data

Human and relational variables such as *happiness, job satisfaction, quality of life, consumers' preferences*, etc. are considered as the main responses in official sample surveys

Ordinal variables:

Associate positive *integers* to discrete choices

Ranking:

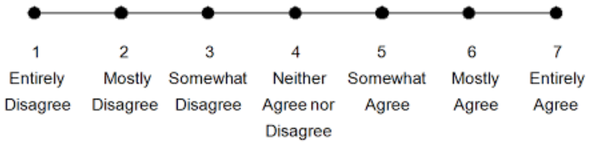
Numbers convey the `location/preference` of the "object" in a given ordered list

items, products, sports, applicants, sentences, teams, songs, . . .

Rating:

Numbers (*ordinal scores*) convey the `level/evaluation` of a "perception"

perception, opinion, taste, fear, worry, agreement . . .



Cumulative models

Analyses of these data are generally performed in the context of Generalized Linear Models (McCullagh 1980; McCullagh and Nelder, 1989):

- Let R_i be the ordinal score marked by the i-th respondent to an item of a questionnaire for $i = 1, \dots, n$:

1	**2**	**. . .**	**r**	**. . .**	**m**

- The discrete response is obtained by grouping the (continuous) latent variable $R_i^\star$ in classes by means of cut-points $(-\infty = \alpha_0 < \alpha_1 < \cdots < \alpha_m = +\infty)$:

$$\alpha_{r-1} < R_i^\star \le \alpha_r \iff R_i = r, \qquad r = 1, \dots, m$$

- A systematic relationship is set between the cumulative function and selected subjects' variables $\mathbf{t}_i$ (*covariates*) via regression coefficients $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)$:

$$R_i^\star = \boldsymbol{\beta}^T \mathbf{t}_i + \epsilon_i, \qquad \Leftrightarrow \qquad Pr(R_i \le r | \boldsymbol{\theta}, \mathbf{t}_i) = F_\theta\left(\alpha_r - \boldsymbol{\beta}^T \mathbf{t}_i\right)$$

(Proportional odds model -POM)

$$logit\left(Pr(R_i \le r | \mathbf{t}_i)\right) = \alpha_r - \boldsymbol{\beta}^T \mathbf{t}_i, \qquad \left(logit(p) = log\left(\frac{p}{1-p}\right), \quad p \in (0,1)\right)$$

In Stata: `ologit`, `oprobit`, `oglm`, ...

Rationale: CUB models paradigm

Psychologists assess that two main aspects are activated when people have to express their evaluation (agreement, worry, etc.) towards an item by selecting a category out of a list of m ordered alternatives (Tourangeau *et al.* (2000)):

Perceptual aspects: *the rater's perception of the item content*

Decisional aspects: *the rater's use of the available scale*

CUB models (Piccolo, 2003) assume that the data generating process is structured as the combination of:

Feeling: *generated by the sound perception of the respondent*

Uncertainty: *generated by the intrinsic fuzziness of the final choice*

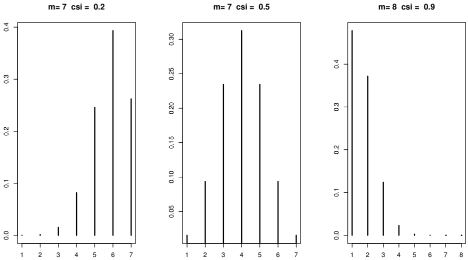
Modelling Feeling

Feeling is the result of a continuous (latent) variable that is discretized: depending on the framework, it is a direct measure of worry, satisfaction, preference, involvement, happiness, ...

1. *How likely is it that you would recommend this brand to a friend or colleague?* (Net Promoter Score)
2. *Does your family easily make ends meet?*
3. ...

CUB paradigm prescribes a **shifted Binomial random variable** for feeling:

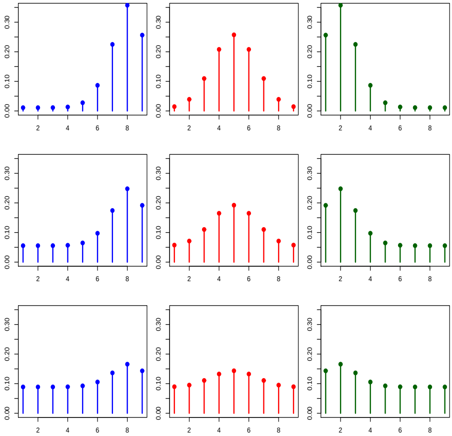
$$b_r(\xi) = \binom{m-1}{r-1} \xi^{m-r} (1-\xi)^{r-1}, \quad r = 1, \dots, m$$



- *Pragmatic view.* The shifted Binomial distribution allows for modal values to be located everywhere on the support $\{1, 2, \dots, m\}$ and on the basis of a single parameter (ξ), related in a simple way to both mode and expectation
- *Statistical view.* When a respondent selects a single value in a list of ordered categories, he/she is comparing each score with all the others. The Binomial distribution "counts" the number of successes (number of times that the selected category is outclassed by the previous ones).

Modelling uncertainty

1. *Limited set of information, Knowledge/Ignorance* about the item
2. *Personal interest/Engagement* in activities related to item
3. *Amount of time* devoted to the response
4. *Range and wording* of the scale
5. *Tiredness or fatigue* for a correct comprehension of the wording
6. *Willingness to joke and fake*
7. *Laziness/Apathy/Boredom*
8. ...



The discrete Uniform random variable U maximizes the entropy among all the discrete distributions with finite support

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

CUB MODELS SPECIFICATION

Let $R_i \in \{1, 2, \ldots, m\}$ the ordinal response given by the i-th subject characterized by variables $\mathbf{t}_i \in \mathbf{T}$. If $\mathcal{C}_i = (r, \mathbf{t}_i)$ denotes the information collected on the i-th subject, then the CUB mixture is defined by:

1. A **stochastic component**:

$$Pr(R_i = r \mid \mathcal{C}_i, \boldsymbol{\theta}) = \pi_i \underbrace{[b_r(\xi_i)]}_{\textbf{feeling}} + (1-\pi_i) \underbrace{\left[\frac{1}{m}\right]}_{\textbf{uncertainty}}, \quad r = 1, 2, \ldots, m; \; i = 1, 2, \ldots, n.$$

   where $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T)^T$

2. Two **systematic components**: if subject's covariates $\boldsymbol{x}_i$ and $\boldsymbol{w}_i$ are chosen to explain π_i and ξ_i, respectively:

$$logit(\pi_i) = \mathbf{x}_i \boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

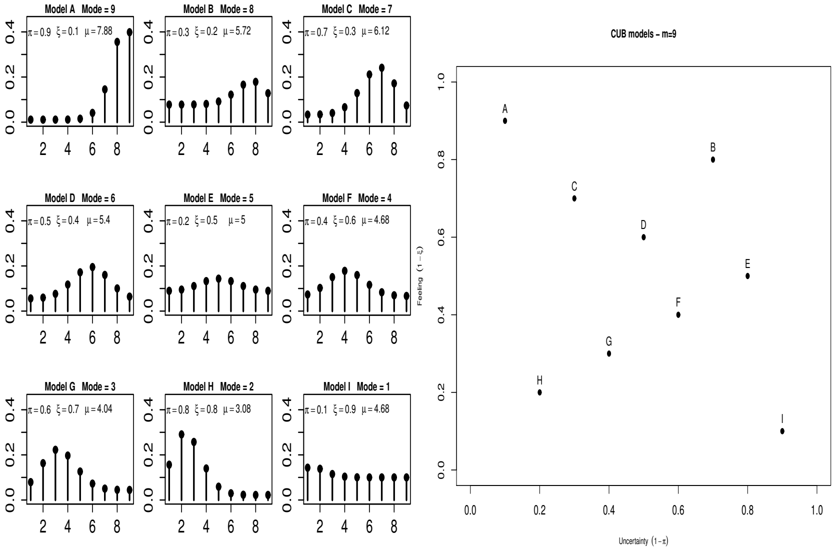
$$logit(\xi_i) = \mathbf{w}_i \boldsymbol{\gamma} = \gamma_0 + \gamma_1 w_{i1} + \cdots + \gamma_q w_{iq}$$

If no covariate is included: $\pi_i = \pi \in (0, 1]$ and $\xi_i = \xi \in [0, 1]$:

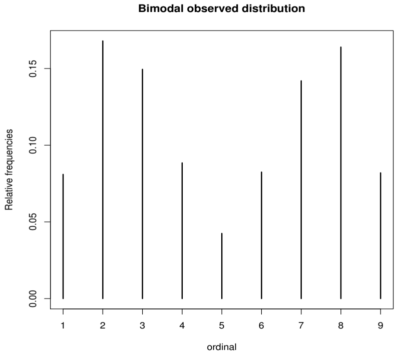
$$logit(\pi) = \beta_0 \iff \pi = \frac{1}{1 + \exp(-\beta_0)}$$

$$logit(\xi) = \gamma_0 \iff \xi = \frac{1}{1 + \exp(-\gamma_0)}$$

CUB MODELS VISUALIZATION



CUB models and bimodality



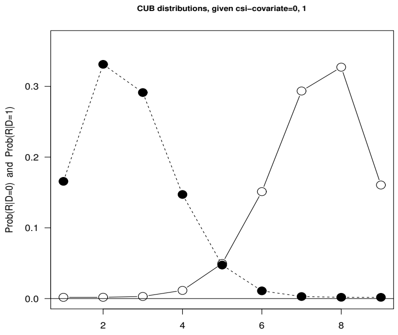


Figure shows the simulated and estimated distributions (conditional to $D_i = 0, 1$, respectively) of the shifted Binomial model ($m = 9$):

$$\begin{cases} Pr\,(R_i = j) = \binom{8}{j-1}\,\xi_i^{8-j}\,(1-\xi_i)^{j-1}\,; \\ logit(\xi_i) = -1.362 + 2.744\,D_i\,; \end{cases} \qquad j = 1, 2, \ldots, 9; \;\; i = 1, 2, \ldots, n.$$

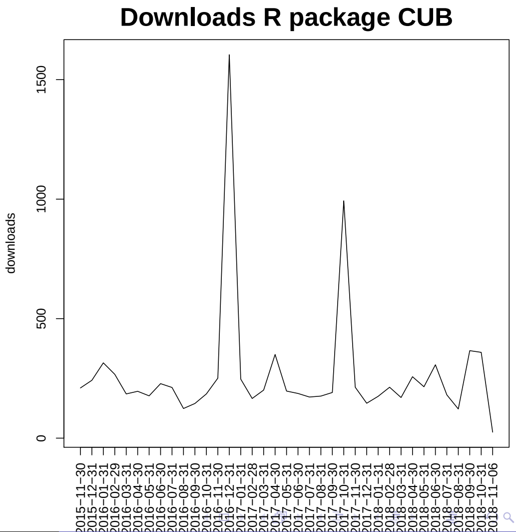
Inferential Issues

Estimation relies on Maximum Likelihood methods:

- Maximum likelihood (ML) estimates of parameters can be obtained by means of the E-M algorithm
- Standard ML asymptotic results apply by using observed information matrix. (Piccolo, 2006).
- For models with covariates, to test significance of each parameter estimate $\hat{\beta}_i$ (or $\hat{\gamma}_j, \hat{\alpha}_i$), Wald test (and Likelihood Ratio test (LRT) in case of nested models) are exploited
- The library 'CUB' is available for the `R` environment on CRAN (previously, Gauss program and R script shared among interested researchers...)

The R package 'CUB'

start	end	downloads
2015-11-01	2015-11-30	210
2015-12-01	2015-12-31	242
2016-01-01	2016-01-31	315
2016-02-01	2016-02-29	267
....	....	....
2016-06-01	2016-06-30	228
2016-07-01	2016-07-31	212
....	....	....
2016-11-01	2016-11-30	251
2016-12-01	2016-12-31	1604
2017-01-01	2017-01-31	247
....	....	....
2017-04-01	2017-04-30	350
2017-05-01	2017-05-31	197
....	....	....
2017-10-01	2017-10-31	993
2017-11-01	2017-11-30	213
....	....	....
2018-02-01	2018-02-28	213
2018-04-01	2018-04-30	257
2018-06-01	2018-06-30	307
2018-09-01	2018-09-30	366



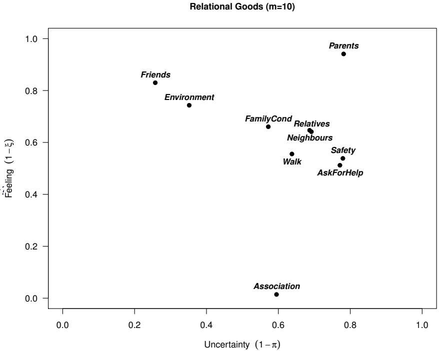
Empirical evidence

- **Preferences**
  - Cities where to live
  - Sensometric analysis and consumers' behaviors
  - Italian newspapers
- **Evaluations**
  - Quality of counseling services for students provided by Universities
  - Services for E-bay users
  - Political affairs: Left/Right self-placement
  - Customers' satisfaction of European consumers towards salmon
- **Perception**
  - Urban audit surveys about city emergencies
  - Chronic pain threshold in TMD (*temporomandibular disorders*)
  - Synonymy and semantic space of words
  - European Union objectives and policies
  - Perception of financial security and job satisfaction in SHIW
  - Subjective survival probability to 75 and 90 years
  - Measure of Happiness

Survey on Relational Goods

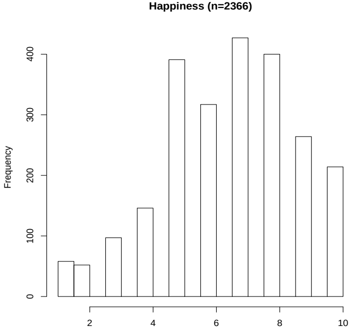
In 2014, $n = 2366$ respondents filled a questionnaire about relational goods and some related issues: items were rated on a scale from 1 to $m = 10$ (1 meaning *"Never, Not at all"* and 10 standing for *"Always, Totally, Absolutely Yes"*).

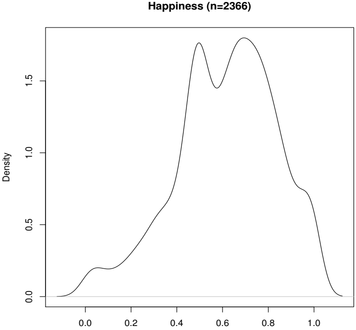
`Walk`	*How often do you walk?*
`Parents`	*How often do you speak with at least one of your parents?*
`Relatives`	*How often do you meet other relatives?*
`Associations`	*How often are you involved in associations?*
`Friends`	*How good are your relationships with friends?*
`Neighbours`	*How good are your relationships with neighbours?*
`Ask-for-help`	*Is it easy for you to ask for help?*
`Environment`	*How good are the relationships with the surrounding environment?*
`Safety`	*Do you feel safe in the place where you live?*
`FamilyCond`	*Does your family easily make ends meet?*



Empirical evidence: Relational Goods

Perceived Happiness: Respondents were asked to self-evaluate their level of happiness by marking a sign along a horizontal line of 110 mm, the left-most extremity standing for "extremely unhappy", and the right-most extremity corresponding to the status "extremely happy".

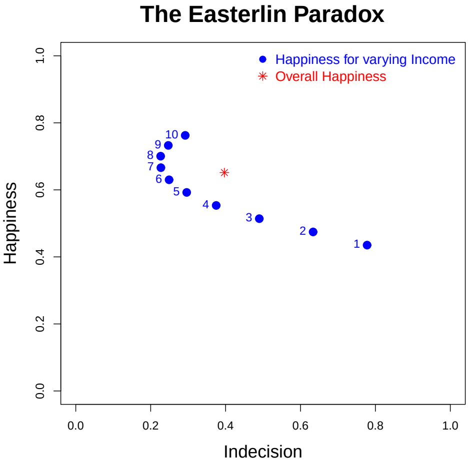




The Easterlin Paradox?

(loglik $= -5099.33$, $BIC = 10237.63$)

$$\begin{cases} logit(1-\pi_i) = \underset{(0.344)}{1.253} - \underset{(0.145)}{0.763}\ FamilyCond_i + \underset{(0.014)}{0.058}\ FamilyCond_i^2 \\ logit(1-\xi_i) = \underset{(0.072)}{-0.261} + \underset{(0.012)}{0.159}\ FamilyCond_i \end{cases}$$



Evaluation of the Orientation Services 2002

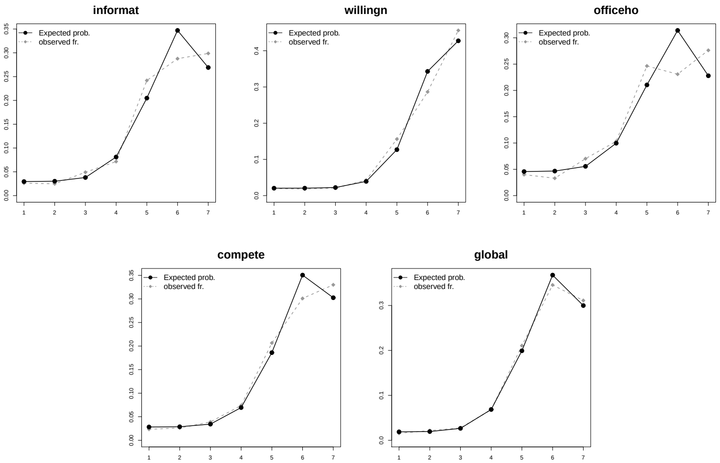
A sample survey on students evaluation of the Orientation services was conducted across the 13 Faculties of University of Naples Federico II in five waves: participants were asked to express their ratings on a 7 point scale (1 = "very unsatisfied", 7 = "extremely satisfied").

Rating variables

- `informat`: Level of satisfaction about the collected information
- `willingn`: Level of satisfaction about the willingness of the staff
- `officeho`: Judgement about the Office hours
- `competen`: Judgement about the competence of the staff
- `global`: Global satisfaction

Subjects' covariates

- `freqserv`: a dummy with levels: 0 = for not regular users, 1 = for regular users
- `age`: a variable indicating the age of the respondent in years
- `gender`: a dummy with levels: 0 = man, 1 = woman
- ....


informat
Expected prob.
observed fr.
willingn
Expected prob.
observed fr.
officeho
Expected prob.
observed fr.
compete
Expected prob.
observed fr.
global
Expected prob.
observed fr.

Shelter effect

If c denotes the *shelter* category, let

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

$R \sim \text{CUB}_{\,she}(\pi^\star, \xi, \delta)$, with shelter at c, if:

$$Pr(R = r|\boldsymbol{\theta}^\star) = (1-\delta)\left(\pi^\star b_r(\xi) + (1-\pi^\star)\frac{1}{m}\right) + \delta\, D_r^{(c)}$$

Possibly, with subjects covariates $\mathbf{v}_i$:

$$logit(\delta_i) = \boldsymbol{v}_i\,\boldsymbol{\omega}$$

Overview

```
*******************************************************************
* "cub" // estimates the cub model (the MAIN one)
*******************************************************************
* "prob_pred" // estimates model predicted probability
*******************************************************************
* "scattercub" // produces the scatterplot of "Uncertainty" and
"Feeling" for cub00
*******************************************************************
* "graph_prob" // produces the graph comparing the actual and the
expected (or model) probabilities for cub00
*******************************************************************
```

Help

help cub

Title

cub —Ordinal outcome model estimated by a mixture of a Uniform and a shifted Binomial

Syntax

cub *outcome* [*if*] [*in*] [*weight*], **pi(***varlist_pi***)** **xi(***varlist_xi***)**

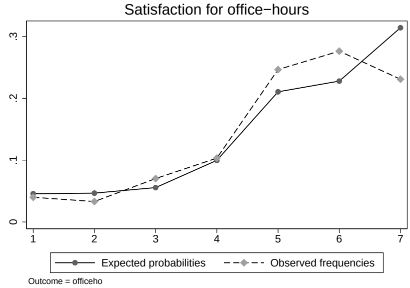
fweights, **pweight**s, **iweight**s are allowed; see weight.

Description

cub estimates a probability model for an ordinal *outcome* variable, where the probability to observe a specific ordinal value (a preference for a given commodity, for instance) is modeled as a mixture of a Uniform and a shifted Binomial distribution. The Uniform distribution models individual *uncertainty* in setting a preference, whereas the shifted Binomial distribution is the law of probability governing individual *feeling* on the item. The user can specify the covariates expected to drive individual uncertainty, as well as those possibly affecting individual feeling. The estimation is performed by maximum likelihood.

```
. cub officeho , pai() csi() vce(oim)

                                                Number of obs     =      2,179
                                                Wald chi2(0)      =          .
Log likelihood = -3759.9171                     Prob > chi2       =          .

------------------------------------------------------------------------------
    officeho |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
pai_beta     |
       _cons |   .7557921   .0889482     8.50   0.000     .5814568    .9301274
-------------+----------------------------------------------------------------
csi_gamma    |
       _cons |  -1.403956   .0371485   -37.79   0.000    -1.476766   -1.331147
------------------------------------------------------------------------------
The number of categories of variable officeho is M = 7
------------------------------------------------------------------------------

-------------------------------------------------------------------------------
******************************************************************************
*********************** Estimates of 'pai' and 'csi' ************************
******************************************************************************

         pai:  1/(1+exp(-_b[pai_beta:_cons]))

------------------------------------------------------------------------------
    officeho |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         pai |   .6804395    .019341    35.18   0.000     .6425317    .7183472
------------------------------------------------------------------------------

         csi:  1/(1+exp(-_b[csi_gamma:_cons]))

------------------------------------------------------------------------------
    officeho |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         csi |   .1971891   .0058808    33.53   0.000     .1856629    .2087152
------------------------------------------------------------------------------
******************************************************************************
```

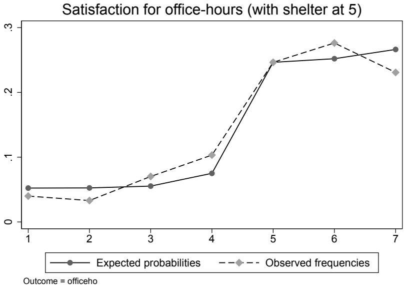
graph_prob officeho



_PROB	Percent
.0456915	3.99
.0466287	3.30
.0555973	7.02
.0996408	10.33
.2105055	24.64
.2278183	27.63
.3141179	23.08

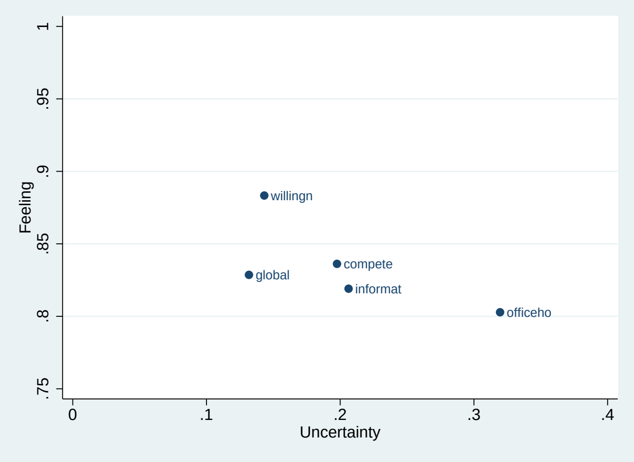
```
. global shelter=5

.
. cub officeho, pai() csi() vce(oim)

                                                Number of obs     =      2,179
                                                Wald chi2(0)      =          .
Log likelihood = -3741.6643                     Prob > chi2       =          .

------------------------------------------------------------------------------
    officeho |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
pai_beta     |
       _cons |   .3800759   .1057342     3.59   0.000     .1728408    .5873111
-------------+----------------------------------------------------------------
csi_gamma    |
       _cons |  -1.722511   .0860042   -20.03   0.000    -1.891076   -1.553946
-------------+----------------------------------------------------------------
delta        |
       _cons |   .0985729   .0158797     6.21   0.000     .0674493    .1296965
------------------------------------------------------------------------------
The number of categories of variable officeho is M = 7
------------------------------------------------------------------------------

--------------------------------------------------------------------------------
********************************************************************************
*********************** Estimates of 'pai' and 'csi' ***********************
********************************************************************************

         pai:  1/(1+exp(-_b[pai_beta:_cons]))

------------------------------------------------------------------------------
    officeho |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         pai |   .5938914   .0255014    23.29   0.000     .5439095    .6438733
------------------------------------------------------------------------------

         csi:  1/(1+exp(-_b[csi_gamma:_cons]))

------------------------------------------------------------------------------
    officeho |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         csi |    .151548   .0110585    13.70   0.000     .1298737    .1732223
------------------------------------------------------------------------------
********************************************************************************
```

graph_prob officeho

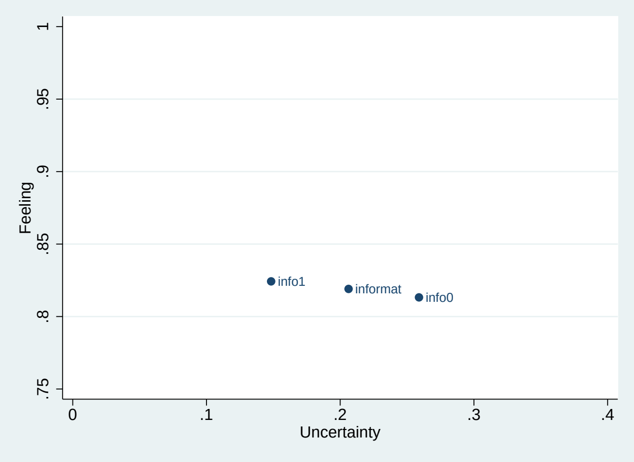


_PROB	Percent
.0523032	3.99
.0525146	3.30
.0553459	7.02
.0750582	10.33
.2464433	24.64
.2520075	27.63
.2663272	23.08

```
scattercub informat willingn officeho compete global
```



```
gen info1=informat if gender==1
gen info0=informat if gender==0
scattercub info0 info1 informat
```



```
. cub global, pai(freqserv) csi(freqserv) vce(oim)

                                                Number of obs     =      2,179
                                                Wald chi2(1)      =      15.99
Log likelihood = -3182.2549                     Prob > chi2       =     0.0001

------------------------------------------------------------------------------
      global |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
pai_beta     |
    freqserv |   1.195018    .2988201     4.00   0.000     .6093414    1.780695
       _cons |   1.636063    .1276279    12.82   0.000     1.385917    1.886209
-------------+----------------------------------------------------------------
csi_gamma    |
    freqserv |  -.5308179    .0616086    -8.62   0.000    -.6515685   -.4100672
       _cons |  -1.385247    .0349435   -39.64   0.000    -1.453735   -1.316759
------------------------------------------------------------------------------
The number of categories of variable global is M = 7
------------------------------------------------------------------------------

. estimate store mod0
```

```
. cub global, pai(freqserv gender) csi(freqserv age) vce(oim)

                                                Number of obs     =      2,179
                                                Wald chi2(2)      =      27.37
Log likelihood = -3168.0857                     Prob > chi2       =     0.0000

------------------------------------------------------------------------------
      global |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
pai_beta     |
    freqserv |   1.201542   .3020578     3.98   0.000     .6095193    1.793564
      gender |   .8916488   .2460514     3.62   0.000      .409397    1.373901
       _cons |   1.292251   .1524861     8.47   0.000     .9933838    1.591118
-------------+----------------------------------------------------------------
csi_gamma    |
    freqserv |  -.5444547   .0614369    -8.86   0.000    -.6648688   -.4240407
         age |  -.0301607   .0080406    -3.75   0.000      -.04592   -.0144015
       _cons |  -.7001192   .1849068    -3.79   0.000     -1.06253   -.3377085
------------------------------------------------------------------------------
The number of categories of variable global is M = 7
------------------------------------------------------------------------------


. estimate store mod1

. lrtest mod0 mod1

Likelihood-ratio test                                 LR chi2(2)  =     28.34
(Assumption: mod0 nested in mod1)                     Prob > chi2 =    0.0000
```

Generalizations and work in progress

- CUBE models for overdispersed data
- CUP models: combination of uncertainty and preference model
- CAUB models for response styles
- Random effects and repeated measurements
- Model-based composite indicators
- Zero-inflated and hurdle models
- Acceleration of convergence procedures
- Model-based classification and regression trees
- ......

Thank you for the attention

Benchmark bibliography

Foundations

- D. Piccolo (2003). On the moments of a mixture of uniform and shifted binomial random variables. *Quaderni di Statistica*, **5**, 85–104.
- A. D'Elia, D. Piccolo (2005). A mixture model for preference data analysis. *Computational Statistics & Data Analysis*, **49**(3), 917–934.
- M. Iannario (2012a). Modelling *shelter* choices in a class of mixture models for ordinal responses. *Statistical Methods and Applications*, **21**(1), 1–22.
- D. Piccolo D., R. Simone, M. Iannario (2018). Cumulative and CUB models for rating data: a comparative analysis, *International Statistical Review*, 1–30, doi:10.1111\insr.12282
- M. Iannario, D. Piccolo and R. Simone (2018), *CUB: A Class of Mixture Models for Ordinal Data* (R package version 1.1.2), `http://CRAN.R-project.org/package=CUB`.

CUB models: extensions

- CUB models with **'don't know'** option

  M. Manisera, P. Zuccolotto (2014). Modeling "Don't know" responses in rating scales. *Pattern Recognition Letters*, **45**, 226–234.

- **Non-Linear** CUB

  M. Manisera, P. Zuccolotto (2014), Modeling rating data with Non Linear CUB models, *Computational Statistics & Data Analysis*, **78**, 100–118

- **Latent class** CUB models: mixtures of CUB distribution to account for heterogeneity in clusters

  L. Grilli, M. Iannario, D. Piccolo, C. Rampichini (2014), Latent class CUB models, *Advances in Data Analysis and Classifications*, **8**, 105–119

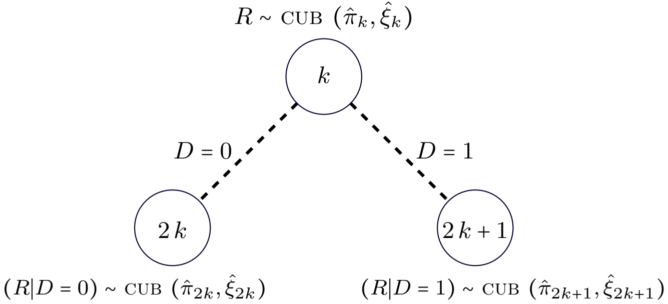
- Logit transform of parameters guarantees robustness

  M. Iannario, A.C. Monti, D. Piccolo, E. Ronchetti (2017), Robust inference for ordinal response models, *Electronic Journal of Statistics* **11**(2), 3407 – 3445.

CUBREMOT (Cappelli, Simone and Di Iorio (2018))

- At node k, corresponding to n_k observations, let $R \sim CUB(\pi_k, \xi_k)$, $m > 3$
- If D is a significant dichotomous covariate D to explain *uncertainty* and/or *feeling*, then:

$$logit(\pi_k) = \beta_0^{(k)} + \beta_1^{(k)}\, D, \quad logit(\xi_k) = \gamma_0^{(k)} + \gamma_1^{(k)}\, D$$



Waiting for Cerulli & Zinilli 's talk: *Calling External Routines in Stata*

Parametric level curves of CUB models