The Stata module for CUB models for rating data analysis

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

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

September 10th, 2021

2021 UK Stata Conference
Virtual

- Human and relational variables such as *satisfaction, well-being, consumers' preferences and opinions*, etc. are considered as the main responses in official sample surveys
- Beyond cumulative models [1], a different approach foresees to model ordinal response variables for preference data directly on the discrete support ($\{c_1 \prec c_2 \prec \cdots \prec c_m\}$) rather than on the continuous latent scale
- In this case, for the observed sample $(r_1, \ldots, r_n)$ with relative frequency distribution $(f_1, \ldots, f_m)$, the fitting result is directly a probability model $(p_1(\boldsymbol{\theta}), \ldots, p_m(\boldsymbol{\theta}))$, where possibly $\boldsymbol{\theta} \equiv \boldsymbol{\theta}_i$ depending on subjects' covariates
- Binomial distribution

Allik J (2014) A mixed-binomial model for Likert-type personality measure. *Frontiers in Psychology* 5:1–13

Pinto da Costa JF, Alonso H, Cardoso JS (2008) The unimodal model for the classification of ordinal data. *Neural Networks*, **21**, 78–91. *Corrigendum* in: (2014). *Neural Networks*, **59**, 73–75

Zhou H, Lange K (2009) Rating movies and rating the raters who rate them. *The Amer Stat*, 63:297–307

[1] In Stata: `ologit, oprobit, oglm,...`

Jenkins S.P. (2020). Comparing distributions of ordinal data. *The Stata Journal*, 20(3), 505–531.

CUB models[a] Rationale

[a]CUB: Combination of Uniform and Binomial

Psychologists (Tourangeau *et al.* (2000)) assess that the ordinal choice is the results of the combination of:

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

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

CUB mixture models 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*


Piccolo D., D'Elia A. (2005). A mixture model for preference data analysis. *Computational Statistics & Data Analysis*, **49**(3), 917–934.


Piccolo D., Simone, R. (2019). The class of CUB models: statistical foundations, inferential issues and empirical evidence. *Statistical Method and Applications*, **28**, 389–435 (with discussions and rejoinder)

The class of CUB mixture models for ordinal variables $(R_1, \ldots, R_n)$ is grounded on the specification of an *uncertainty* and a feeling component:

$$Pr\left(R_i = r \mid \boldsymbol{x}_i, \boldsymbol{w}_i, \boldsymbol{\theta}\right) = \pi_i \, b_r(\xi_i \mid \boldsymbol{w}_i) + (1 - \pi_i) \frac{1}{m}, \qquad r = 1, \ldots, m, \; i = 1, \ldots, n$$

Shifted Binomial:

$$b_r(\xi_i) = \binom{m-1}{r-1} \xi_i^{m-r} (1 - \xi_i)^{r-1}$$

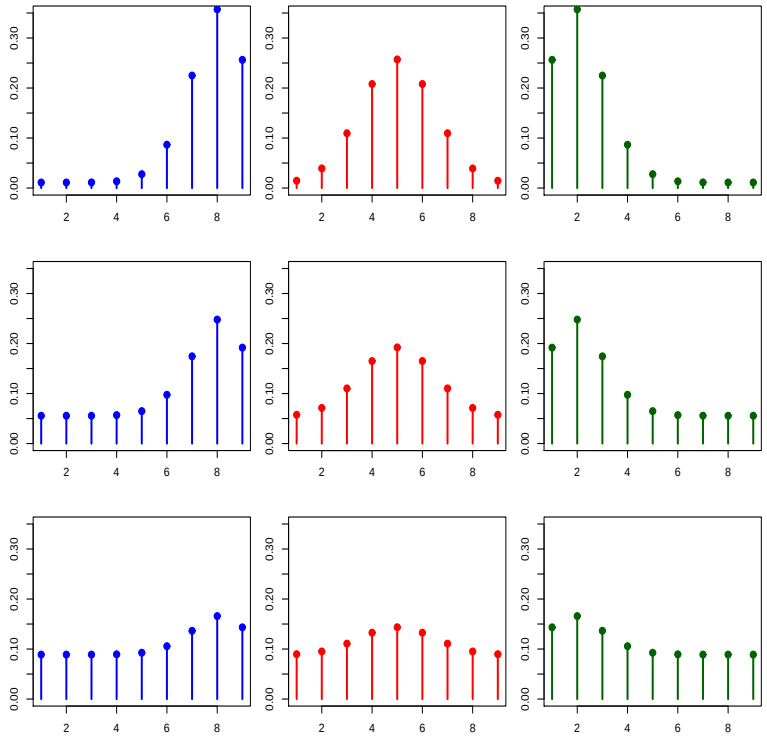
Systematic components:

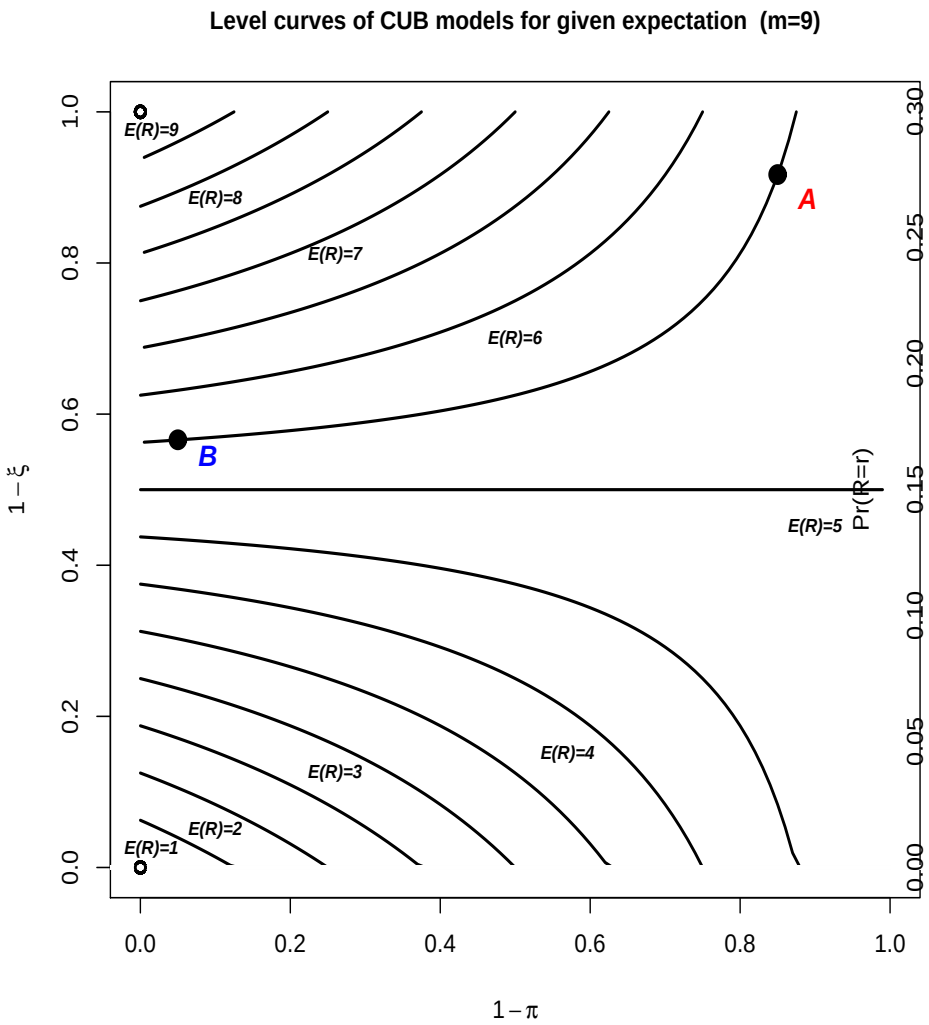
$$\text{logit}\left(\pi_i\right) = \boldsymbol{x}_i \boldsymbol{\beta}$$

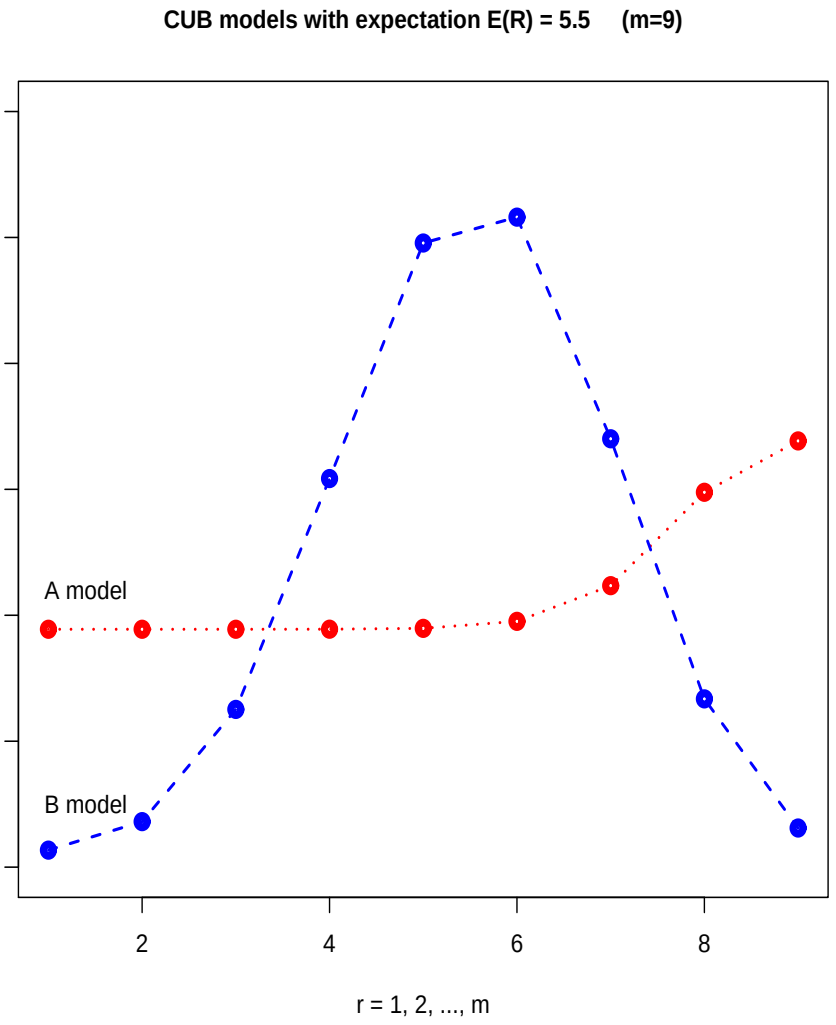
$$\text{logit}\left(\xi_i\right) = \boldsymbol{w}_i \boldsymbol{\gamma}$$

No covariate:

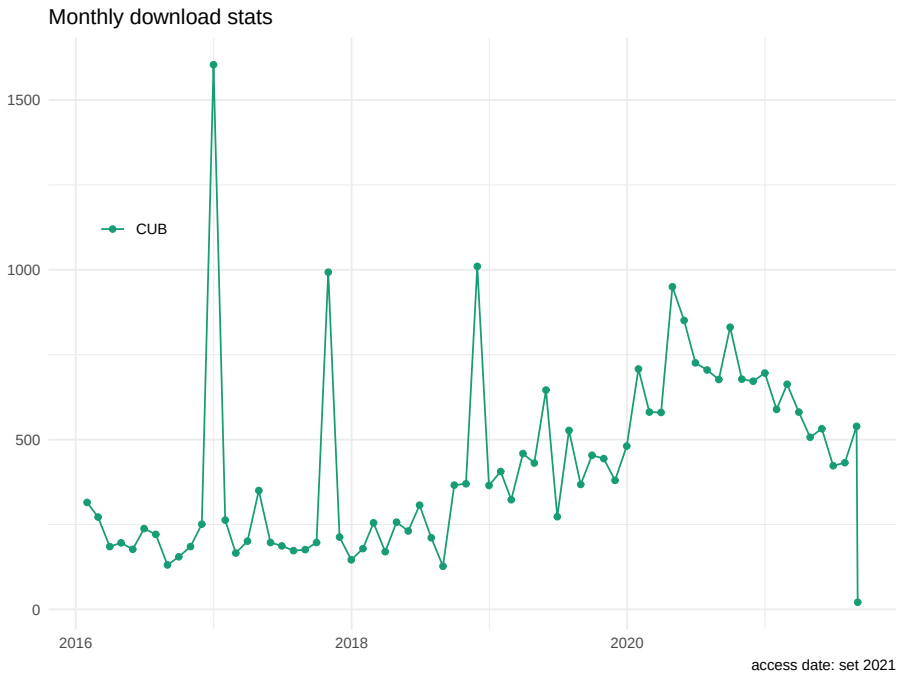
$$\pi_i = \pi \in (0, 1], \qquad \xi_i = \xi \in [0, 1]$$




Level curves of CUB models for given expectation (m=9)
E(R)=9
E(R)=8
E(R)=7
E(R)=6
E(R)=5
E(R)=4
E(R)=3
E(R)=2
E(R)=1
A
B
1 − ξ
1 − π
0.0
0.2
0.4
0.6
0.8
1.0


CUB models with expectation E(R) = 5.5 (m=9)
A model
B model
Pr(R=r)
r = 1, 2, ..., m
0.00
0.05
0.10
0.15
0.20
0.25
0.30
2
4
6
8

R package 'CUB'



- Maximum likelihood (ML) estimates of parameters can be obtained via E-M algorithm for mixtures or direct optimization;
- Standard ML asymptotic results apply by using observed information matrix or Louis'identity within EM.
- For models with covariates, to test significance of each parameter estimate $\hat{\beta}_i$ (or $\hat{\gamma}_j, \hat{\alpha}_i$), Wald test (and LRT for nested models) can be exploited

Iannario M., Piccolo D. Simone R (2018). CUB: A Class of Mixture Models for Ordinal Data. R package version 1.1.3. https://CRAN.R-project.org/package=CUB

Simone R. (2020). FastCUB: Fast EM and Best-Subset Selection for CUB Models for Rating Data. R package version 0.0.2. https://CRAN.R-project.org/package=FastCUB

Simone R., Di Iorio F., Lucchetti R. (2019). CUB for GRETL. In: Gretl 2019: Pro- ceedings of the International Conference on the GNU Regression, Econometrics and Time Series Library, Eds. Di Iorio - Lucchetti, feDOA University Press, ISBN: 978-88-6887-057-7, pp. 147–166, http://ricardo.ecn.wfu.edu/gretl/cgi-bin/gretldata_cgi?opt=SHOW_FUNCS

`cub` *depvar* [*if*] [*in*] [*weight*] [, `xi(`*varlist_xi*`)` `pi(`*varlist_pi*`)` `shelter(#)` `m(#)` `prob(`*newvarname*`)` `graph` `outname(`*name*`)` `save_graph(`*filename*`)`]

Options

`xi(`*varlist_xi*`)` specifies the covariates explaining the "feeling" parameter.

`pi(`*varlist_pi*`)` specifies the covariates explaining the "uncertainty" parameter.

`shelter(#)` specifies the "shelter", i.e. the category associated with an inflated frequency.

`m(#)` specifies the total number of categories of the dependent variable. **It is important to provide this input if any category in *depvar* has zero observed frequency. If this option is not specified, the procedure will set *m* at the maximum observed response value.**

`prob(`*newvarname*`)` allows the user to generate a new variable containing the model fitted probabilities.

`graph` allows the user to generate a graph displaying a plot of the actual and predicted probabilities.

`outname(`*name*`)` allows the user to specify a convenient name for the outcome variable to appear in the graph, when the `graph` option is invoked.

`save_graph(`*filename*`)` allows the user to save the graph generated by the `graph` option.

Cerulli G. (2020). "CUB: Stata module to estimate ordinal outcome model estimated by a mixture of a uniform and a shifted binomial," Statistical Software Components S458727, Boston College Department of Economics, revised 22 Jun 2021.

Cerulli G., Simone R., Di Iorio F., Piccolo D., Baum C.F. (2021) The CUB Stata module: mixture models for feeling and uncertainty of rating data, *The STATA Journal* (under review)

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
- .....

```
clear all
use universtata.dta , clear
. cub officeho

                                                  Number of obs     =      2,179
                                                  Wald chi2(0)      =          .
Log likelihood = -3759.9171                       Prob > chi2       =          .
------------------------------------------------------------------------------
*************** Estimates of 'pi' and 'xi' *********************************
------------------------------------------------------------------------------
officeho |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
---
-------------------------------------------------------
pi |   .6804395    .019341    35.18   0.000     .6425317    .7183472
------------------------------------------------------------------------------
officeho |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 xi |   .1971891   .0058808    33.53   0.000     .1856629    .2087152
------------------------------------------------------------------------------
******************************************************************************
------------------------------------------------------------------------------
```

```
scattercub informat willingn officeho compete global , m(7 7 7 7 7),
save_graph(mygraph1)
```

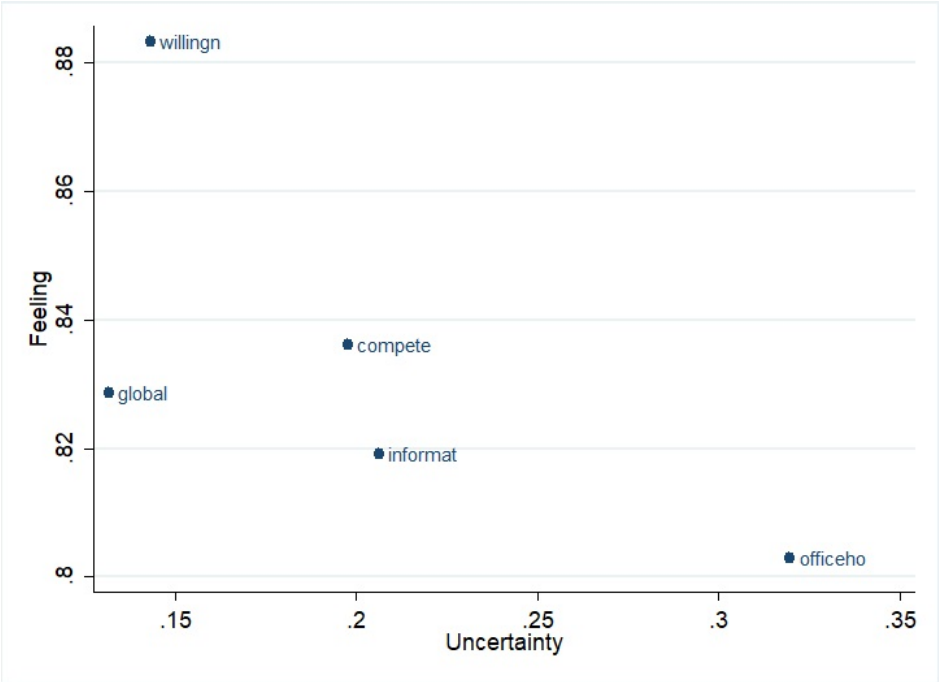
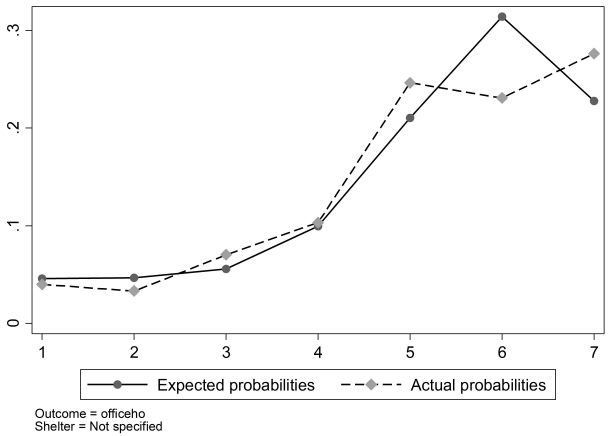


FIGURE: CUB models without covariates for Satisfaction Items in `universtata.dta` ($m = 7$).



If $c \in \{1, \ldots, m\}$ 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:

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

```
cub officeho, shelter(5) prob(_PROB) graph save_graph(mygraph2)
------------------------------------------------------------------------------
Log likelihood = -3741.6643
------------------------------------------------------------------------------
    officeho |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          pi |   .5938915   .0255014    23.29   0.000     .5439096    .6438734
          xi |    .151548   .0110585    13.70   0.000     .1298737    .1732223
       delta |   .0985727   .0158797     6.21   0.000     .0674491    .1296963
```

Actual vs. fitted probabilities

officeho	fitted_~b	actual_~b
1	.0523032	.0399266
2	.0525146	.0330427
3	.0553459	.0702157
4	.0750582	.1032584
5	.2464432	.2464433
6	.2663273	.2308398
7	.2520075	.2762735

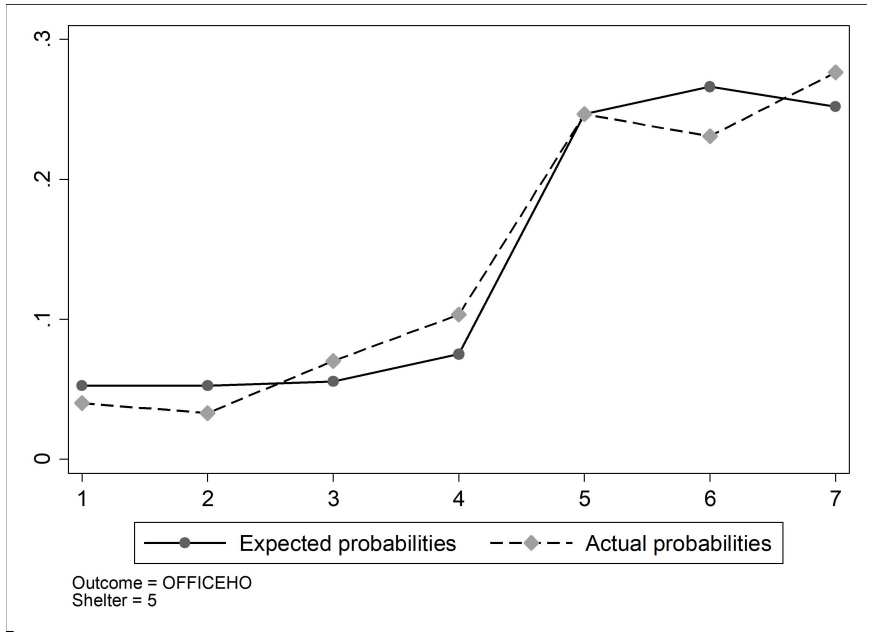


FIGURE: Plot of the observed vs. fitted probabilities for variable officeho under a CUB model without covariates with shelter at category 5

```
cub officeho, pi(freqserv) xi(freqserv)  prob(_PROB)  ///
     graph save_graph(mygraph2)

                                                Number of obs     =      2,179
                                                Wald chi2(1)      =       0.14
Log likelihood = -3704.2854                     Prob > chi2       =     0.7057

------------------------------------------------------------------------------
    officeho |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
pi_beta      |
    freqserv |  -.0688115   .1822338    -0.38   0.706    -.4259831    .2883602
       _cons |   .8144389   .1146983     7.10   0.000     .5896343    1.039244
-------------+----------------------------------------------------------------
xi_gamma     |
    freqserv |  -.8253576   .0944552    -8.74   0.000    -1.010486   -.6402288
       _cons |  -1.149043   .0406619   -28.26   0.000    -1.228739   -1.069347
------------------------------------------------------------------------------
The number of categories of variable officeho is M = 7
------------------------------------------------------------------------------
```

Comparisons between regular and occasional users

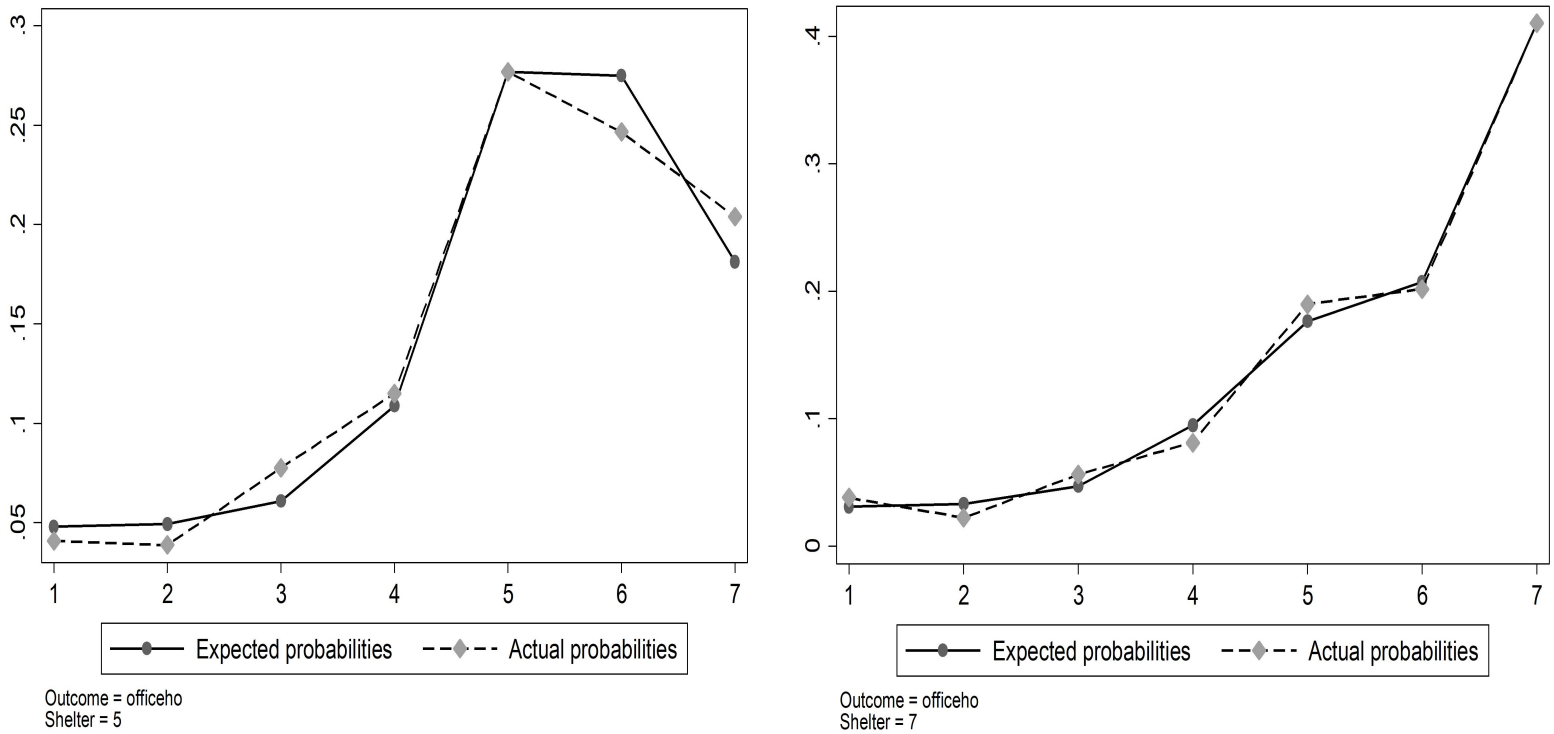


Figure: Separate fit of CUB models with shelter for ratings on `officeho`, given `freqserv` (left: shelter at $c = 5$ for non-regular users; right: shelter at $c = 7$ for regular users)

Thank you for the attention!

Simone (2021) An accelerated EM algorithm for mixture models with uncertainty for rating data. Computational Statistics, **36**:691–714

Simone R., Tutz G., Iannario M. (2020) Subjective Heterogeneity in Response Attitude for Multivariate Ordinal Outcomes. *Econometrics and Statistics*, **14**:145–158.

Cappelli C., Simone R., Di Iorio F. (2019) CUBREMOT: A tool for building model-based trees for ordinal responses. *Expert Systems with Applications*, **124**, 39–49.

Piccolo D. Simone R. (2019) The class of CUB models: statistical foundations, inferential issues and empirical evidence. *Statistical Method and Applications*, **28**: 389–435 (with discussions and rejoinder)

Capecchi S., Simone R., Ghiselli S. (2019) Drivers and uncertainty for job satisfaction the italian graduates. *IJAS - Italian Journal of Applied Statistics,* **31**(2):227-250, 2019.

Capecchi S., Simone R. (2019) A proposal for a model-based composite indicator:experience on perceived discrimination in Europe. *Social Indicators Research*, **141**(1):95-110.

Piccolo D., Simone R., Iannario M. (2019) Cumulative and CUB models for rating data: a comparative analysis. *International Statistical Review*, **87**: 207–236.

Simone R., Iannario M. (2018) Analysing sport data with clusters of opposite preferences. *Statistical Modelling*, **18**(5-6), 1–20.

Simone R., Tutz G. (2018) Modelling uncertainty and response styles in ordinal data. *Statistica Neerlandica*. **72**: 224-245.

D'Elia and Piccolo (2005) A mixture model for preference data analysis. *Computational Statistics & Data Analysis*, **49**(3), 917–934.