The Stata Journal (2022)
22, Number 1, pp. 195–223 DOI: 10.1177/1536867X221083927

Fitting mixture models for feeling and uncertainty for rating data analysis

Giovanni Cerulli
IRCrES-CNR
Rome, Italy
giovanni.cerulli@ircres.cnr.it
https://orcid.org/0000-0002-5892-3372

Rosaria Simone
Department of Political Sciences
University of Naples Federico II
Naples, Italy
rosaria.simone@unina.it
https://orcid.org/0000-0002-6844-6418

Francesca Di Iorio
Department of Political Sciences
University of Naples Federico II
Naples, Italy
fdiiorio@unina.it
https://orcid.org/0000-0002-9586-3380

Domenico Piccolo
Department of Political Sciences
University of Naples Federico II
Naples, Italy
domenico.piccolo@unina.it
https://orcid.org/0000-0002-5198-1078

Christopher F. Baum
Department of Economics
Boston College
Chestnut Hill, MA
baum@bc.edu
https://orcid.org/0000-0003-4766-3699

Abstract. In this article, we present the command `cub`, which fits ordinal rating data using combination of uniform and binomial (CUB) models, a class of finite mixture distributions accounting for both feeling and uncertainty of the response process. CUB identifies the components that define the mixture in the baseline model specification. We apply maximum likelihood methods to estimate feeling and uncertainty parameters, which are possibly explained in terms of covariates.

An extension to inflated CUB models is discussed. We also present a subcommand, `scattercub`, for visualization of results. We then illustrate the use of `cub` using a case study on students' satisfaction for the orientation services provided by the University of Naples Federico II in Italy.

Keywords: st0669, cub, scattercub, CUB, mixture models, rating data, maximum likelihood estimation

1 Motivation

Several estimation commands, such as `ologit`, `oprobit`, or `oglm`, are available to Stata users to analyze ordinal data based on classical modeling approaches (Tutz 2012). The repository at http://users.stat.ufl.edu/~aa/ordinal/ord.html reports related examples of the benchmark reference (Agresti 2010). A command to compare distributions of ordinal data has been recently introduced in Stata (Jenkins 2020).

In this article, we add to this literature by presenting a command implementing the class of combination of uniform and binomial (CUB) models for ordinal data (Piccolo and Simone 2019a,b), uniform and binomial being the two distributions used to jointly model feeling and uncertainty of the response process via a mixture specification. Beyond this baseline definition, this new paradigm for ordinal data modeling (Piccolo 2003; D'Elia and Piccolo 2005) includes a richer class of models, which has shown to be of interest to a broad audience of applied scholars because of a versatile and multifaceted range of applications (Balirano and Corduas 2008; Arboretti Giancristofaro, Bordignon, and Carrozzo 2014; Capecchi and Piccolo 2016; Fin et al. 2017; Capecchi, Simone, and Ghiselli 2019) and the flexibility to perform more complex analysis (Cappelli, Simone, and Di Iorio 2019; Simone, Cappelli, and Di Iorio 2019; Simone, Tutz, and Iannario 2020; Manisera and Zuccolotto 2014; Bonnini et al. 2012; D'Elia 2008). From the methodological point of view, see Piccolo, Simone, and Iannario (2019) for a comparative analysis with cumulative models.

The innovative aspect of the combination of uniform and binomial (CUB) paradigm is the modeling of uncertainty arising from the ensemble of individuals or framing effects surrounding the evaluation on rating scales. This component is meant to convey indecision, fuzziness, and the heterogeneity of responses (Di Nardo and Simone 2019), yielding a twofold interpretation of response patterns. Uncertainty blurs the assessment of respondents' sentiments toward the trait being investigated (preference, satisfaction, and so on). Thus, the CUB paradigm involves a mixture between the least informative uniform distribution over the discrete support and an adequate model for feeling to analyze both heterogeneity and location of the responses, respectively. Linking estimable uncertainty and feeling parameters to subjects' covariates adds further value. This feature allows the derivation of interpretable response profiles useful for understanding and prediction of response behaviors.

This approach to the analysis of the rating process can be extended to account for other response phenomena, such as overdispersion and an inflated frequency in a given category, by modifying the baseline distributions in the model specification. Frequency

inflation occurs when one category is a refuge or shelter option for the response choice because of its peculiar wording, because of response styles, or to avoid the cognitive burden of a more precise choice. Hence, we refer to this as a shelter effect. We present an example in the illustrative case study in section 4.

CUB models have proven to be parsimonious yet valuable in research and applications in social and behavioral studies, particularly in terms of their effective visualization features. In addition to providing a response distribution for each covariate profile, estimated feeling and uncertainty measures can be represented as points in the parameter space. Hereafter, we call this representation `scattercub`. In this way, effective comparative analysis can be pursued when several rating variables or groups of respondents are investigated jointly. Sections 3 and 4 describe this feature in detail.

Currently, the CUB (Iannario, Piccolo, and Simone 2020) and FASTCUB (Simone 2020) libraries are available for the R environment and for the GRETL community (Simone, Di Iorio, and Lucchetti 2019) as open-source software. They both include the implementation of the expectation-maximization algorithm (McLachlan and Krishnan 2008) for maximum likelihood inference (Piccolo 2006). For Stata users, no related tool is available. To fill this gap, we present the commands `cub` and `scattercub` to provide Stata users with new tools for ordinal data analysis.

This article is intended to provide a concise yet comprehensive introduction of CUB models, illustrating their applications and interpretation. Section 2 briefly reviews the methodological background. See Piccolo and Simone (2019a,b) and the references for an overview and a comprehensive description of the state of the art on methodology and applications. Section 3 sets out the syntax of the `cub` and `scattercub` commands. Section 4 provides an illustrative application from a survey carried out in 2002 to evaluate students' satisfaction for the orientation services at the University of Naples Federico II, Italy. Customer satisfaction, in its broadest sense, is one of the most emblematic applications for the class of CUB models. This dataset is downloadable with the package. Section 5 concludes the article by highlighting possible applications and identifying future software developments that could make a more comprehensive use of CUB models attractive for researchers.

2 CUB model specification

For a sample of size n, let R_i be the ordinal rating response provided to a given item (of a questionnaire) by the ith subject, $i = 1, \ldots, n$. Assume that the response is collected on a Likert-type scale with m ordered categories, with $m > 3$ required for identifiability (Iannario 2010). For convenience, categories will be coded as the first m integers to convey their position along the scale. CUB models' paradigm prescribes that the rating process arise from the combination of two main components: feeling, addressing the perception of the item being investigated (attraction, satisfaction, agreement, and so on); and uncertainty, conveying the fuzzy elements of the response. The CUB model for the rating response mechanism is then specified as a two-component mixture distribution of these components. In the baseline definition, uncertainty is modeled by a uniform

discrete distribution over the first m integers to contribute to model parsimony, while feeling is modeled by a shifted Binomial distribution of the parameter $\xi_i \in (0,1)$:

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

Thus, if $\boldsymbol{w}_i, \boldsymbol{y}_i$ are the row vectors of selected covariates for the ith subject, for $i = 1, \ldots, n$, a CUB model for the response R_i is specified by the two-component mixture on the discrete support:

$$\Pr(R_i = r \mid \boldsymbol{y}_i, \boldsymbol{w}_i, \boldsymbol{\theta}) = \pi_i \, b_r(\xi_i) + (1-\pi_i) \frac{1}{m}, \quad r = 1, \ldots, m \tag{1}$$

Here the uncertainty parameter $\pi_i = \pi(\boldsymbol{y}_i, \boldsymbol{\beta}) \in (0, 1)$ is introduced to weight the two components. Specifically, an increase of π_i implies a reduced impact of the uncertainty component on the response distribution. In (1), the logit transforms[1] of the feeling and uncertainty parameters ξ_i and π_i link these model features to subject characteristics:

$$\text{logit}(1-\pi_i) = -\boldsymbol{y}_i \boldsymbol{\beta}; \qquad \text{logit}(1-\xi_i) = -\boldsymbol{w}_i \boldsymbol{\gamma}; \qquad i = 1, \ldots, n \tag{2}$$

Then $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\gamma}')$ is the estimable parameter vector characterizing the distribution of $(R_1, \ldots, R_n)$, with $\boldsymbol{\beta}, \boldsymbol{\gamma}$ denoting the column vectors of regression coefficients for the uncertainty and feeling parameters, respectively.

Model selection for the best covariate specification can be performed by fitting several models and choosing the one that attains the lowest values of the Akaike information criterion (Akaike 1974) or Bayesian information criterion (Schwarz 1978). As with all mixture models, variable-selection procedures should be based on a crossed search for the best covariate specification for both feeling parameter ξ_i and uncertainty parameter π_i. This could be pursued via best-subset search algorithms as described in Simone (2021).

A focal point of this class of models is that covariate specification is not compulsory. A CUB model could describe a given rating distribution in terms of a global measure of feeling ξ and uncertainty π, in which case one refers to the model as CUB$(0,0)$. In this case, $(1-\pi)$ is a (normalized) measure of the uncertainty implied by the model in terms of the overall heterogeneity of the distribution (Capecchi and Piccolo 2017). Thus, CUB models allow characterization of different rating responses in terms of only two parameters (π, ξ), ranging in $(0,1] \times [0,1]$ and leading to effective visualization tools. Indeed, for different items or response profiles, obtained by conditioning (1) on selected values of covariates, estimated uncertainty and feeling parameters identify a point in the parameter space, yielding a scatterplot that gives a unified picture of the data at hand. This can be visualized with the `scattercub` command. For instance, one can identify which items or response profiles are associated with a higher feeling or more

1. The choice of the logit link is mainly motivated by the ease of interpretation of the covariates' effect. In fact, any admissible link may be used; for instance, the probit link is considered as an option in Hernández Barajas, Usuga Manco, and García Muñoz (2018), where the estimation steps are carried out without resorting to expectation-maximization procedures.

homogeneous patterns and so on, possibly determining clusters of models (Corduas 2011). We discuss an illustration of this specific feature for the case in which the model includes covariates in section 4. Along these lines, model-based composite indicators for multivariate rating data have been proposed (Capecchi and Simone 2019).

A further remark on interpretation is worthwhile. According to common motivations for mixture models (McLachlan and Peel 2000), CUB models should imply two clusters of respondents such that, in the more uncertain group, people randomly select an ordinal score. Although this is a possible interpretation, in the case when no covariates are specified, the CUB parameterization of the response variable directly on its support is intended to be a synthesis of the overall distribution in terms of location and heterogeneity. Otherwise, the CUB parameterization provides a method to assess individuals' level of uncertainty, which can be interpreted as subjective indecision and feeling in terms of subjects' characteristics.

2.1 Inflated CUB models

One of the major advantages of the CUB paradigm is the ease of extending the model to encompass other circumstances that may affect the rating response process. One typical scenario concerns the inflation in frequency for a category that attains a peculiar meaning or role for the respondents. Inflated CUB models include a so-called shelter effect (Corduas, Iannario, and Piccolo 2009; Iannario 2012) located at a known category $s \in \{1, \ldots, m\}$. This category is excessively frequent, beyond that accounted for by the standard CUB mixture. To fit this circumstance, the CUB model is extended with the introduction of a degenerate distribution:[2]

$$\Pr(R = r \mid \boldsymbol{\theta}) = \delta \left(D_r^{(s)} \right) + (1 - \delta) \left\{ \pi^* \, b_r(\xi) + (1 - \pi^*) \, \frac{1}{m} \right\}, \quad r = 1, 2, \ldots, m \quad (3)$$

The estimable parameter vector is then $\boldsymbol{\theta} = (\pi^*, \xi, \delta)'$, quantifying the shelter effect with the parameter δ. As for CUB mixtures, covariates' effects can be tested by specifying a logit link with model parameters. Because the standard CUB model is nested into specification (3), a likelihood-ratio test can be implemented to assess the potential improvement from a shelter specification.

3 The cub and scattercub commands

3.1 Syntax for cub

The model fit by `cub` considers a sample of the ordinal response variable R as the main input (*outcome*); notice that the response values should be coded as integers from 1 to m. Further inputs include a series of covariates (either continuous or categorical),

2. If c denotes the shelter category, the degenerate distribution $D_r^{(c)}$ is such that $D_r^{(c)} = 1$ if $r = c$ and $D_r^{(c)} = 0$ if $r \neq c$, for $r = 1, \ldots, m$.

possibly explaining the uncertainty (*varlist_pi*) and the feeling (*varlist_xi*) parameters, as well as an optional shelter effect at a given category r.

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

`fweights` and `pweights` are allowed; see [U] **11.1.6 weight**.

3.2 Options for cub

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

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

`shelter(`#`)` specifies the shelter, that is, 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 *outcome* has zero observed frequency. By default, the procedure will set `m()` at the maximum observed response value.

`prob(`*newvar*`)` generates a new variable containing the model-fitted probabilities.

`graph` generates a graph displaying a plot of the actual and predicted probabilities.

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

`save_graph(`*filename*`)` saves the graph generated by the `graph` option.

3.3 Syntax for scattercub

`scattercub` fits the CUB model without covariates for each element of a list of rating variables (*varlist*) (possibly collected on scales with different numbers of response options) and generates the scatterplot of the corresponding feeling versus uncertainty measures ($1-\xi$ and $1-\pi$, respectively) in the unit square $(0,1]\times[0,1]$.

`scattercub` *varlist* [*if*] [*in*] [*weight*] [`,` `m(`*numlist*`)` `save_data(`*filename*`)` `save_graph(`*graph_name* `)`]

`fweights` and `pweights` are allowed; see [U] **11.1.6 weight**.

3.4 Options for scattercub

`m(`*numlist*`)`, for each ordinal dependent variable in *varlist*, optionally specifies the total number of categories. This total number comprises both observed and unobserved categories. By default, only observed categories are used to determine the number of response options `m()` needed for CUB specification and estimation.

`save_data(`*filename*`)` saves in *filename* the dataset containing the uncertainty and feeling measures.

`save_graph(`*graph_name*`)` saves in *filename* the feeling versus uncertainty scatterplot.

4 CUB models at work

This section is meant to help Stata users become familiar with the `cub` command by illustrating its usage within a cross-sectional data analysis. Specifically, the goal of this section is to show how to exploit the methodology of CUB models in the context of an illustrative case study on customer satisfaction.

4.1 Application

`universtata.dta` arises from a sample survey on students' evaluation of the orientation services that has been administered to students across the Faculties of the University of Naples Federico II, in Italy, in five time waves. Participants were asked to express their ratings on a 7-point Likert-type scale (1 = very unsatisfied, 7 = extremely satisfied) on the following aspects:

- `informat`: Level of satisfaction about the acquired information
- `willingn`: Level of satisfaction about the willingness of the staff
- `officeho`: Level of satisfaction about the opening hours
- `competen`: Level of satisfaction about the competence of the staff
- `global`: Level of global satisfaction

Hereafter, we consider the data collected in 2002, consisting of 2,179 observations. The remaining 7 variables correspond with subjects' covariates (for instance, the dichotomous variable `gender`, equal to 0 for men and to 1 for women, and the continuous measurement `age`).

As a first step, we show how to simultaneously visualize the ordinal variables included in `universtata.dta` with the command `scattercub` by fitting a CUB$(0, 0)$ model to every ordinal variable. Then each fitted model is represented as a point in the parameter space corresponding with the maximum likelihood estimates of the uncertainty $1 - \widehat{\pi}$ and feeling $1 - \widehat{\xi}$ parameters. This exploratory graphical analysis reveals how response heterogeneity and feeling vary among the different aspects of satisfaction in a comparative perspective. To enhance visualization, figure 1 displays only a subspace of the parametric space.

```
. sysuse universtata
. quietly scattercub informat willingn officeho compete global, m(7 7 7 7 7)
> save_graph(mygraph1.png)
```

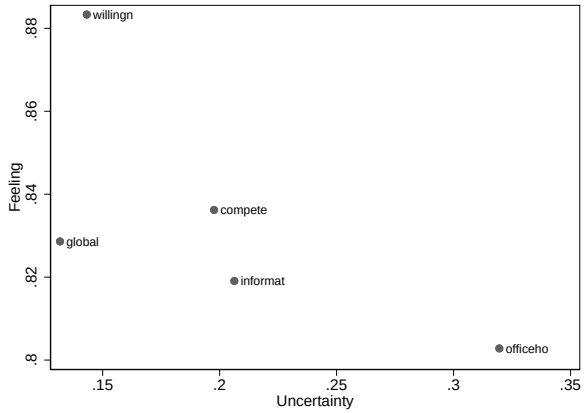

Figure 1. CUB models without covariates for satisfaction items in `universtata.dta` ($m = 7$)

From figure 1, we observe that the highest feeling has been expressed for the willingness of the staff, whereas the lowest corresponds with the scheduled office hours. Because the latter item is affected by the highest uncertainty, it deserves further investigation. Thus, we focus on the item `officeho`; there are no categories of the original measurement scale (with $m = 7$ response options) with zero frequency, so it is redundant to provide the option `m()`. We include it for illustrative purposes only.

We first show how to estimate the parameters of a CUB$(0, 0)$ model for the `officeho` item; this is the simplest model specification possible because parameters do not depend on covariates. This goal is attained by simply typing

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

officeho	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
pi_beta						
_cons	.7557921	.0889482	8.50	0.000	.5814568	.9301274
xi_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 'pi', and 'xi' ********************************
******************************************************************************
          pi: 1/(1+exp(-_b[pi_beta:_cons]))
------------------------------------------------------------------------------
    officeho | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
          pi |   .6804395    .019341    35.18   0.000     .6425317    .7183472
------------------------------------------------------------------------------
          xi: 1/(1+exp(-_b[xi_gamma:_cons]))
------------------------------------------------------------------------------
    officeho | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
          xi |   .1971891   .0058808    33.53   0.000     .1856629    .2087152
------------------------------------------------------------------------------
******************************************************************************
```

To provide a unique output format for fitted CUB models possibly with covariate effects, we also report the logit transformation of feeling and uncertainty parameters (2) when no covariate is specified. In this circumstance, the "constants" denoted as `pi_beta:_cons` and `xi_gamma:_cons` are related to π and ξ, respectively, by

$$\pi = \frac{1}{1+e^{-\beta_0}}; \qquad \xi = \frac{1}{1+e^{-\gamma_0}}$$

By default, the output tables report the $\widehat{\pi}$ and $\widehat{\xi}$ estimates (second panel) as well as the inverse logit transformations (first panel), that is, $\widehat{\beta}_0$ and $\widehat{\gamma}_0$, respectively. Notice that, given the orientation of the response scale, the actual satisfaction sentiment toward an item is the complement to one of the feeling parameters ξ; thus, hereafter $1-\xi$ will be considered as a feeling indicator, increasing with latent satisfaction.

The `prob()` and `graph` options within the `cub` command return the plot displayed in figure 2, comparing the observed and fitted probabilities for variable `officeho`, as well as a table setting out these probabilities:

```
. cub officeho, prob(_PROB) graph
  (output omitted)
Actual vs. fitted probabilities
    officeho | fitted_~b  actual_~b
-------------+----------------------
           1 |  .0456915   .0399266
           2 |  .0466287   .0330427
           3 |  .0555973   .0702157
           4 |  .0996408   .1032584
           5 |  .2105055   .2464433
           6 |  .3141179   .2308398
           7 |  .2278183   .2762735
```

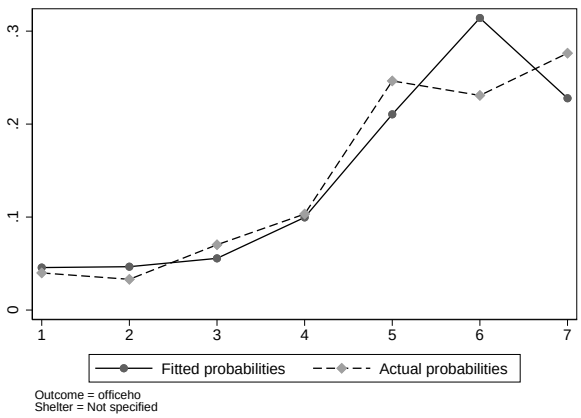

Figure 2. Plot of the observed versus fitted probabilities for variable `officeho` under a CUB$(0, 0)$ model

It follows that the model does not sufficiently fit responses observed for categories 5, 6, and 7. Specifically, a moderate inflation in frequency for the fifth category seems unaccounted for by the model. Thus, we test for a possible shelter effect at category $s = 5$ by calling

```
. cub officeho, shelter(5) prob(_PROB) graph save_graph(mygraph2)
> outname("OFFICEHO")
                                                  Number of obs = 2,179
                                                  Wald chi2(0)  =     .
Log likelihood = -3741.6643                       Prob > chi2   =     .
```

officeho	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
pi_beta						
_cons	.3800763	.1057342	3.59	0.000	.1728411	.5873114
xi_gamma						
_cons	-1.722511	.0860041	-20.03	0.000	-1.891076	-1.553946
lambda						
_cons	-2.213185	.1787122	-12.38	0.000	-2.563454	-1.862915

```
The number of categories of variable officeho is M = 7
```

```
*******************************************************************************
*************** Estimates of 'pi', and 'xi' ***********************************
*******************************************************************************
         pi: 1/(1+exp(-_b[pi_beta:_cons]))
------------------------------------------------------------------------------
    officeho | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
          pi |   .5938915   .0255014    23.29   0.000     .5439096    .6438734
------------------------------------------------------------------------------

         xi: 1/(1+exp(-_b[xi_gamma:_cons]))
------------------------------------------------------------------------------
    officeho | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
          xi |    .151548   .0110585    13.70   0.000     .1298737    .1732223
------------------------------------------------------------------------------
*******************************************************************************

*******************************************************************************
******** Estimation of the shelter parameters 'delta'  ************************
*******************************************************************************
       delta: 1/(1+exp(-_b[lambda:_cons]))
------------------------------------------------------------------------------
    officeho | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
       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
------------------------------------------------------------------------------
file mygraph2.gph saved
```

As indicated by both an improvement in the log likelihood and the significance of parameter δ, it can be inferred that category 5 is perceived as a shelter for the assessment of satisfaction on office hours. Notice that the first panel reports the logit transform also for δ [specifically, $\text{logit}(\delta) = \lambda$] to give a unified presentation of results.

The fit improvement entailed by the specification of the shelter effect can be additionally inspected with a graphical comparison between observed frequencies and estimated probabilities (see figure 3).

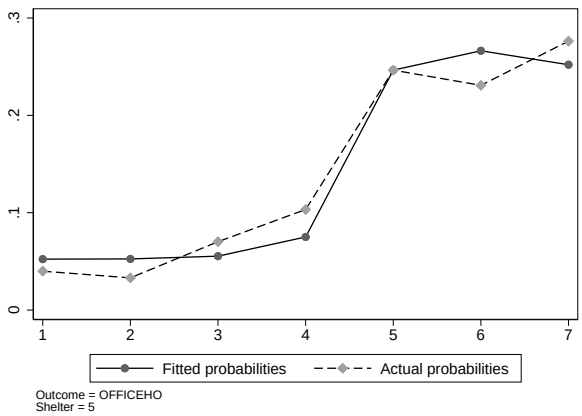

Figure 3. Plot of the observed versus fitted probabilities for variable `officeho` under a CUB model without covariates with shelter at category 5

Next, to enrich the interpretation of results, we introduce some covariates in the model to identify the main determinants of satisfaction for office hours in terms of students' characteristics. As a first example, we test if the model components can be explained by the dichotomous covariate `freqserv`, indicating the users' frequency of the service, with levels 0 and 1 for nonregular and regular users, respectively.

```
. cub officeho, pi(freqserv) xi(freqserv) prob(_PROB)
> graph save_graph(mygraph2) outname("OFFICEHO")
                                                  Number of obs =  2,179
                                                  Wald chi2(1)  =   0.14
Log likelihood = -3704.2854                       Prob > chi2   = 0.7057
```

officeho	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
pi_beta						
freqserv	-.0688113	.1822338	-0.38	0.706	-.4259829	.2883604
_cons	.8144389	.1146983	7.10	0.000	.5896343	1.039243
xi_gamma						
freqserv	-.8253573	.0944552	-8.74	0.000	-1.010486	-.6402285
_cons	-1.149044	.040662	-28.26	0.000	-1.228739	-1.069348

```
The number of categories of variable officeho is M = 7
```

```
Actual vs. fitted probabilities
    officeho | fitted_~b  actual_~b
-------------+-----------------------
           1 |  .0446193   .0399266
           2 |  .0454518   .0330427
           3 |  .0537141   .0702157
           4 |    .09586   .1032584
           5 |  .2065812   .2464433
           6 |  .3175473   .2308398
           7 |  .2362263   .2762735
-------------------------------------
file mygraph2.gph saved
```

Because the regression coefficient for $\text{logit}(\xi_i)$ is negative, it follows that regular users have a higher feeling $1-\xi_i$ than occasional users, whereas there is no statistically significant difference in terms of heterogeneity between these groups. Notice that covariate specification in CUB models can be similar or different for uncertainty and feeling parameters. In this case, one could fit a CUB model by including the `freqserv` covariate only to explain feeling and considering the uncertainty parameter π not depending on any covariate by calling

```
. cub officeho, xi(freqserv)
```
(*output omitted*)

If covariates are specified in the model, then the `cub` command returns—in addition to the estimation results—a table comparing observed relative frequencies with the average of estimated probabilities given the covariates for each category. Indeed, the estimated probability $\Pr(R_i = r|\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}, \boldsymbol{y}_i, \boldsymbol{w}_i)$ for each observed individual response can be computed once subject-specific π_i and ξ_i are obtained via (2) from estimated parameters $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\gamma}}$ and covariates values $\boldsymbol{y}_i, \boldsymbol{w}_i$ for the ith subject. Then the estimated probabilities are grouped by response value, and the average of fitted probabilities for each category is returned.

It can be insightful to compare the estimated probability distribution for the two groups of respondents (regular and nonregular users in the case under examination). This goal can be obtained via the following commands, returning a matrix and a plot comparing observed relative frequencies and fitted probabilities for the two groups (see figure 4).

```
. * CUB for "freqserv==1"
. quietly cub officeho if freqserv==1, prob(predicted_probs) save_graph(gr_m)
> graph
. matrix P_m=e(M)
. * CUB for "freqserv==0"
. quietly cub officeho if freqserv==0, prob(predicted_probs) save_graph(gr_f)
> graph
. matrix P_f=e(M)
. * Generate the adjoint matrix P
. matrix P=P_m,P_f
```

```
. matrix list P
```
(*output omitted*)
```
. * Generate the four probabilities' variables (actual and fitted by freqserv)
. preserve
. svmat2 P, rnames(categories)
. destring categories, replace
categories: all characters numeric; replaced as byte
(2172 missing values generated)
. drop if categories==.
(2,172 observations deleted)
. keep categories P*
. * Run the graph
. sort categories
. tw (connected P1 categories) (connected P2 categories)
> (connected P3 categories) (connected P4 categories),
> legend(order(1 "freqserv_1 fitted" 2 "freqserv_1 actual"
> 3 "freqserv_0 fitted" 4 "freqserv_0 actual"))
. restore
. matrix list P
P[7,4]
   fitted_prob  actual_prob  fitted_prob  actual_prob
1    .04597005    .03800786    .04398401    .04096045
2    .04606406    .02228047    .04639857    .03884181
3     .0477008    .05635649    .06395741    .07768362
4    .06260974    .08125819     .1284419    .11511299
5    .13586293    .19003932    .24402751    .27683616
6    .30494848    .20183486    .29648811    .24646893
7    .35684395     .4102228    .17670254    .20409605
```

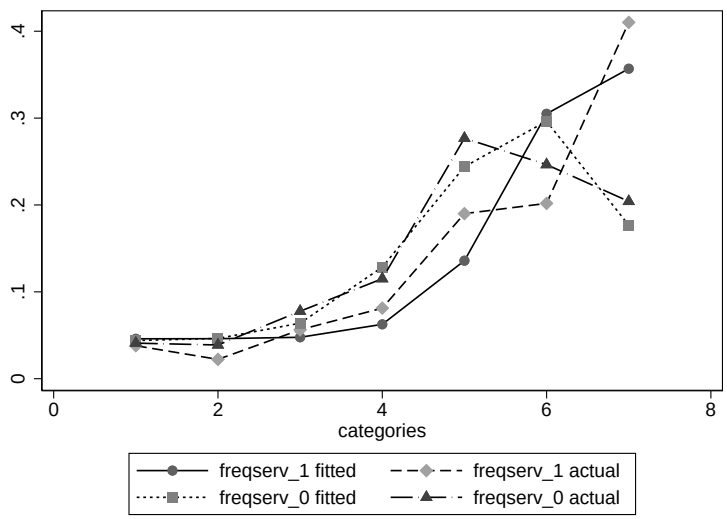

Figure 4. Plot of the observed frequencies and fitted probabilities for variable `officeho` under a CUB$(0, 0)$, conditional to `freqserv`

According to the fitted models, it follows that relevant differences between the two profiles appear only in categories 4, 5, and 7; specifically, nonregular users are more likely to score lower categories 4 and 5 than regular ones. Conversely, regular users are more likely to score the highest grade $R = 7$ than the nonregular ones. In particular, figure 4 indicates that inflation in category 5 is mainly due to regular users, whereas inflation in the last category should be accounted for instead for the ratings assigned by nonregular users to further improve the fit. This circumstance could be assessed by fitting the following models and comparing classical goodness-of-fit statistics; results are displayed in figure 5.

```
. cub officeho if freqserv==0, shelter(5) prob(_Prob0) graph
> save_graph(mygraph2)
```

```
                                                Number of obs = 1,416
                                                Wald chi2(0)  =     .
Log likelihood = -2458.7964                     Prob > chi2   =     .
```

officeho	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
pi_beta						
_cons	.5772791	.134363	4.30	0.000	.3139324	.8406258
xi_gamma						
_cons	-1.260895	.0616	-20.47	0.000	-1.381629	-1.140162
lambda						
_cons	-2.616979	.3055337	-8.57	0.000	-3.215814	-2.018144

```
The number of categories of variable officeho is M = 7
```

```
*******************************************************************************
*************** Estimates of 'pi', and 'xi' ***********************************
*******************************************************************************
         pi: 1/(1+exp(-_b[pi_beta:_cons]))
```

officeho	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
pi	.6404411	.0309406	20.70	0.000	.5797986	.7010836

```
         xi: 1/(1+exp(-_b[xi_gamma:_cons]))
```

officeho	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
xi	.2208198	.0105988	20.83	0.000	.2000465	.241593

```
*******************************************************************************
```

```
******************************************************************************
******** Estimation of the shelter parameters 'delta'  **********************
******************************************************************************
       delta: 1/(1+exp(-_b[lambda:_cons]))
```

officeho	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
delta	.0680537	.0193777	3.51	0.000	.0300741	.1060332

```
(763 missing values generated)
(763 missing values generated)
(763 missing values generated)
(763 missing values generated)
```

Actual vs. fitted probabilities

officeho	fitted_~b	actual_~b
1	.0479391	.0409605
2	.049335	.0388418
3	.0607937	.0776836
4	.1086733	.115113
5	.2768358	.2768362
6	.2749865	.2464689
7	.1814365	.204096

```
file mygraph2.gph saved
. cub officeho if freqserv==1, shelter(7) prob(_Prob1) graph
> save_graph(mygraph2)

                                                      Number of obs = 763
                                                      Wald chi2(0)  =   .
Log likelihood = -1208.9212                           Prob > chi2   =   .
```

officeho	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
pi_beta						
_cons	.8162676	.1832598	4.45	0.000	.4570851	1.17545
xi_gamma						
_cons	-1.110301	.0911449	-12.18	0.000	-1.288942	-.9316604
lambda						
_cons	-.8956222	.1250323	-7.16	0.000	-1.140681	-.6505634

```
The number of categories of variable officeho is M = 7
```

```
*******************************************************************************
*************** Estimates of 'pi', and 'xi' ***********************************
*******************************************************************************
          pi: 1/(1+exp(-_b[pi_beta:_cons]))
------------------------------------------------------------------------------
    officeho | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
          pi |   .6934435   .0389573    17.80   0.000     .6170886    .7697984
------------------------------------------------------------------------------
          xi: 1/(1+exp(-_b[xi_gamma:_cons]))
------------------------------------------------------------------------------
    officeho | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
          xi |   .2478147   .0169897    14.59   0.000     .2145156    .2811138
------------------------------------------------------------------------------
*******************************************************************************

*******************************************************************************
******** Estimation of the shelter parameters 'delta'  ************************
*******************************************************************************
       delta: 1/(1+exp(-_b[lambda:_cons]))
------------------------------------------------------------------------------
    officeho | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
       delta |    .289951   .0257416    11.26   0.000     .2394984    .3404035
------------------------------------------------------------------------------
(1,416 missing values generated)
(1,416 missing values generated)
(1,416 missing values generated)
(1,416 missing values generated)

Actual vs. fitted probabilities
    officeho | fitted_~b  actual_~b
-------------+----------------------
           1 |  .0312098   .0380079
           2 |  .0331726   .0222805
           3 |  .0468555   .0563565
           4 |  .0948758   .0812582
           5 |  .1762882   .1900393
           6 |  .2073753   .2018349
           7 |  .4102228   .4102228

file mygraph2.gph saved
```

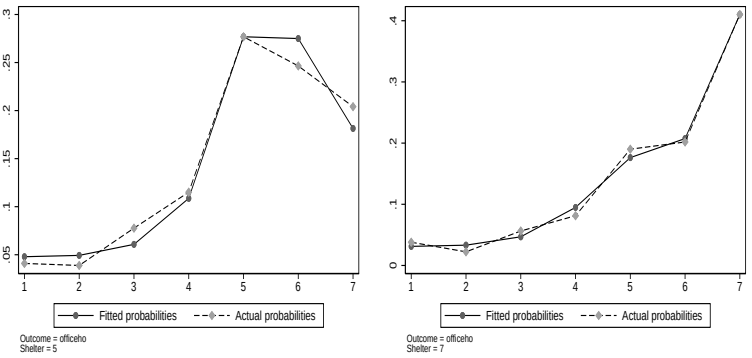

Figure 5. Separate fit of CUB models with shelter for ratings on `officeho`, given `freqserv` (left: shelter at $s = 5$ for nonregular users; right: shelter at $s = 7$ for regular users)

As an example of a more complex covariate specification, we show how to check for possible age effects. We consider the deviation from the mean of the logarithmic transform of age (covariate `slnage`) together with `gender` to explain uncertainty. When covariates are included for both parameters as in the previous example, the estimates will be the corresponding coefficients of the logistic link for the uncertainty and the feeling parameters. Then the resulting CUB model is specified by

$$\begin{aligned} \text{logit}(1-\pi_i) &= -\beta_0 - \beta_1\, \texttt{slnage}_i - \beta_2\, \texttt{gender}_i \\ \text{logit}(1-\xi_i) &= -\gamma_0 - \gamma_1\, \texttt{slnage}_i - \gamma_2\, \texttt{freqserv}_i \end{aligned} \tag{4}$$

After the `slnage` variable is generated, the command to implement this model is

```
. generate lage=ln(age)
. egen mlage=mean(lage)
. generate slnage=lage-mlage
. cub officeho, pi(slnage gender) xi(slnage freqserv)
```

The output of the estimation procedure is given below and indicates that regular users have a higher feeling than occasional users and that younger users have lower feeling than older ones and higher uncertainty.[3] In addition, responses provided by women are more heterogeneous than those provided by men.

3. Here we consider as young those individuals whose age (in logarithmic scale) is lower than the average.

```
                                                  Number of obs =  2,179
                                                  Wald chi2(2)  =  11.71
Log likelihood = -3693.8876                       Prob > chi2   = 0.0029
```

officeho	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
pi_beta						
slnage	1.237235	.6134223	2.02	0.044	.0349497	2.439521
gender	.4956044	.1687882	2.94	0.003	.1647857	.8264232
_cons	.564898	.1177285	4.80	0.000	.3341543	.7956417
xi_gamma						
slnage	-.5905128	.2409135	-2.45	0.014	-1.062695	-.1183311
freqserv	-.8228392	.0849996	-9.68	0.000	-.9894353	-.6562431
_cons	-1.14668	.0403469	-28.42	0.000	-1.225758	-1.067601

The number of categories of variable officeho is M = 7

Because no plot is directly provided as output for complex covariate specifications, the results of the CUB model estimation with significant covariates on feeling and uncertainty parameters may be represented, for instance, as in figure 6, obtained with the following commands:

```
. * Run CUB with covariates
. cub officeho, pi(slnage gender) xi(slnage freqserv)
  (output omitted)
. * Produce linear predictions
. predict pred_csi, equation(xi_gamma) xb
. predict pred_pai, equation(pi_beta) xb
. * Form the four groups
. generate group=.
(2,179 missing values generated)
. replace group=1 if gender==1 & freqserv==1
(385 real changes made)
. replace group=2 if gender==0 & freqserv==1
(378 real changes made)
. replace group=3 if gender==1 & freqserv==0
(664 real changes made)
. replace group=4 if gender==0 & freqserv==0
(752 real changes made)
. * Generate the "feeling" and the "uncertainty" variables
. generate feeling=invlogit(1-pred_csi)
. generate uncertainty=invlogit(1-pred_pai)
```

```
. * Plot the graph
. sort(uncertainty)
. twoway
> (line feeling uncertainty if group==1 & age<=22, lwidth(medthick)
>    lpattern(solid))
> (line feeling uncertainty if group==2 & age<=22, lwidth(medthick)
>    lpattern(dash))
> (line feeling uncertainty if group==3 & age>22, lwidth(medthick)
>    lpattern(longdash_dot))
> (line feeling uncertainty if group==4 & age>22, lwidth(medthick)
>    lpattern(dash_dot)),
> legend(label(1 "Young man user") label(2 "Young woman user")
> label(3 "Old man not-user") label(4 "Old woman not-user"))
> scheme(s1mono) xtitle("Uncertainty (1-pi)") ytitle("Feeling (1-xi)")
> saving(mygraph3, replace)
file mygraph3.gph saved
```

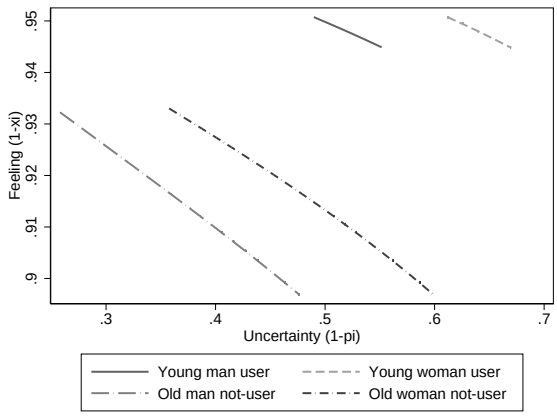

Figure 6. Plot of CUB model with covariates (4): age and gender effect for uncertainty, age and frequency effect for feeling

Figure 6 is meant to display how feeling and uncertainty vary together with the continuous variable `slnage` when conditioning to the selected dummy variables. For illustrative purposes, for each of the possible four groups identified by gender and frequency of service, we have plotted only a restricted set of values to identify four profiles: young male and female regular users and older male and female nonregular users. In light of the comments discussed above on estimated parameters, the bottom point of each curve segment corresponds with younger ages within each group.

As discussed in section 2, the class of CUB mixture models includes a specific extension to fit the so-called shelter effect, arising in the presence of an inflated category. For illustrative purposes, we show how to perform the analysis of a possible shelter effect in the previous model using category 5 as the shelter choice for `officeho` if covariate effects are also investigated for feeling and uncertainty measures. The code is as follows:

```
. cub officeho, shelter(5) pi(slnage gender) xi(slnage freqserv)
                                                  Number of obs =  2,179
                                                  Wald chi2(2)  =  13.35
Log likelihood = -3667.9995                       Prob > chi2   = 0.0013
```

officeho	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
pi_beta						
slnage	1.318132	.5887743	2.24	0.025	.1641556	2.472108
gender	.4970248	.160193	3.10	0.002	.1830523	.8109973
_cons	.270122	.117399	2.30	0.021	.0400241	.5002199
xi_gamma						
slnage	-.685046	.3182336	-2.15	0.031	-1.308772	-.0613196
freqserv	-1.243455	.1306411	-9.52	0.000	-1.499507	-.9874033
_cons	-1.301998	.0566919	-22.97	0.000	-1.413112	-1.190884
lambda						
_cons	-2.267073	.1447959	-15.66	0.000	-2.550868	-1.983278

```
The number of categories of variable officeho is M = 7
```

```
******************************************************************************
******** Estimation of the shelter parameters 'delta'  **********************
******************************************************************************
       delta: 1/(1+exp(-_b[lambda:_cons]))
```

officeho	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
delta	.0938869	.0123181	7.62	0.000	.0697439	.11803

The sign of the regression coefficients for the selected covariates for both uncertainty and feeling parameters confirms, overall, the interpretations derived from inspection of figure 6. In addition, we observe that the significance test for parameter δ suggests that accounting for inflation in category 5 improves the fit even after controlling for the selected covariates.

5 Categories with zero frequencies

The `cub` command also allows consideration of settings where the dataset presents zero-frequency categories, that is, categories that are part of the response measurement support but that no respondents have chosen. For illustrative purposes, we consider two artificial scenarios assuming that the rating variable `officeho` was collected on a scale with eight and nine categories and for which nonzero frequencies were observed for the first seven categories only. We then compare the results graphically with the `scattercub` command, including the original measurement, to see how feeling and uncertainty are adjusted when specifying the complete rating scale length in case of unobserved response options; see figure 7.

```
. generate officeho8 = officeho
. generate officeho9 = officeho
. scattercub officeho officeho8 officeho9, m(7 8 9)
(0 observations deleted)
```
(*output omitted*)

The estimates are updated to account for the presence of the extra categories having zero frequency. The total number of categories considered in the procedure (including those with zero frequency) is reported in the last line of the first panel of the output table (here not reported for brevity).

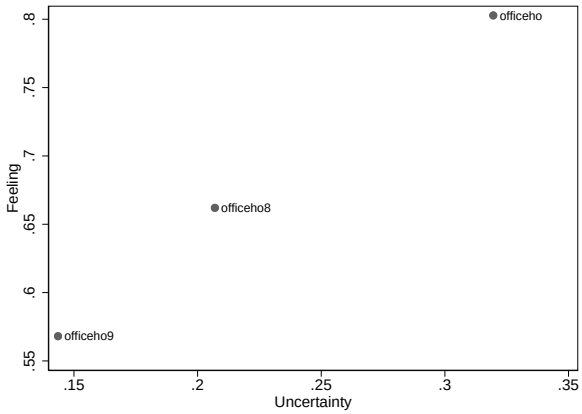

Figure 7. Scatterplot to display the effect of misspecification of the length of the original rating scale in case there are categories with zero observed frequencies

To discuss the case of zero-frequency categories not at the extreme of the scales, we consider for illustrative purposes the ratings on satisfaction for the willingness of the staff of the orientation office, and we artificially set frequencies of the second and third category at 0 by shifting those responses to 1. Figure 8 shows the graphical output of the code below; also, in this case the shelter effect is tested at $s = 7$.

```
. generate w1=willingn
. replace w1=1 if willingn==2 | willingn==3
(86 real changes made)
. cub w1, prob(_PROB) graph save_graph(miss_cat.jpg) outname("w1")
```
(*output omitted*)
```
. cub w1, prob(_PROB) graph shelter(7) save_graph(miss_cat.jpg) outname("w1")
```
(*output omitted*)

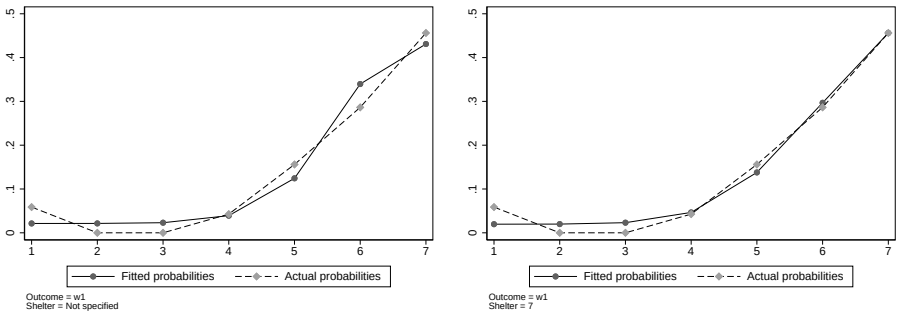

Figure 8. Observed and estimated distributions restricted to observed categories of satisfaction ratings for willingness of the staff (without and with shelter at $s = 7$)

Accordingly, the output will report the estimated uncertainty and feeling estimation results (given below only for the fitted CUB model with shelter at $s = 7$).

```
. cub w1, prob(_PROB) graph shelter(7)

                                                  Number of obs = 2,179
                                                  Wald chi2(0)  =     .
Log likelihood = -3000.0991                       Prob > chi2   =     .
```

w1	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
pi_beta						
_cons	1.621426	.1010487	16.05	0.000	1.423374	1.819478
xi_gamma						
_cons	-1.768895	.057773	-30.62	0.000	-1.882128	-1.655662
lambda						
_cons	-1.616902	.19687	-8.21	0.000	-2.00276	-1.231044

```
The number of categories of variable w1 is M = 7
```

```
********************************************************************************
*************** Estimates of 'pi', and 'xi' ************************************
********************************************************************************
         pi: 1/(1+exp(-_b[pi_beta:_cons]))
```

w1	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
pi	.8349917	.0139225	59.97	0.000	.807704	.8622794

```
         xi: 1/(1+exp(-_b[xi_gamma:_cons]))
```

w1	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
xi	.1456798	.0071903	20.26	0.000	.1315871	.1597724

```
********************************************************************************
```

```
*******************************************************************************
******** Estimation of the shelter parameters 'delta'  **********************
*******************************************************************************
       delta: 1/(1+exp(-_b[lambda:_cons]))
------------------------------------------------------------------------------
          w1 | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
       delta |   .1656326   .0272071     6.09   0.000     .1123076    .2189575
------------------------------------------------------------------------------

Actual vs. fitted probabilities
          w1 | fitted_~b  actual_~b
-------------+----------------------
           1 |  .0196749   .0587425
           2 |  .0199025          0
           3 |  .0231036          0
           4 |  .0465297   .0426801
           5 |  .1378123   .1560349
           6 |  .2968046   .2863699
           7 |  .4561725   .4561726
------------------------------------------------------------------------------
```

6 Conclusions

Compared with more consolidated approaches mainly derived from cumulative models (McCullagh 1980), which is the leading pathway to analyze ordinal data, the CUB paradigm offers wider possibilities from both the interpretative and graphical points of view. In addition, an important consequence is the circumstance that CUB models are not constrained to include covariates as explanatory tools to fit consistent models for data fitting, prediction, and classification. This opportunity allows the introduction of more flexible methods to manage and compare rating responses.

In this framework, the `cub` command for ordered rating data is presented to provide a new analytical tool for Stata users interested in ordinal data modeling, thus broadening the extent of application of the class of CUB mixture distributions. This methodology provides measures of both latent uncertainty and feeling of the response process, which could be possibly linked to subjects' covariates. The command also allows for the estimation of CUB models with shelter to account for an inflated frequency. The visualization features for CUB models have been emphasized because this is one of the most obvious advantages of this modeling approach.

Improvements of `cub` could include extra functions to fit model extensions, as CUB model extensions to account for overdispersion (Piccolo 2015).

Beyond extended methodologies, CUB modeling is also under active development for applications; in this respect, we quote original marketing research in the field of food preferences and sensory analysis: Piccolo and D'Elia (2008), Iannario et al. (2012), Corduas, Cinquanta, and Ievoli (2013), Capecchi et al. (2016), Mauracher, Procidano, and Sacchi (2016), and Contini et al. (2016). We also quote recent new perspectives

and applications (Hwang, Sohn, and Oh 2015; Low 2017; Finch and Hernández Finch 2020; Hu, Zhou, and Sharma 2020; and Xu and Zhang 2021), providing evidence of an increasing international interest toward the CUB paradigm.

7 Programs and supplemental materials

To install a snapshot of the corresponding software files as they existed at the time of publication of this article, type

```
. net sj 22-1
. net install st0669     (to install program files, if available)
. net get st0669         (to install ancillary files, if available)
```

8 References

Agresti, A. 2010. *Analysis of Ordinal Categorical Data*. 2nd ed. Hoboken, NJ: Wiley.

Akaike, H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19: 716–723. https://doi.org/10.1109/TAC.1974.1100705.

Arboretti Giancristofaro, R., P. Bordignon, and E. Carrozzo. 2014. Two phase analysis of ski schools customer satisfaction: Multivariate ranking and cub models. *Statistica* 74: 141–154. https://doi.org/10.6092/issn.1973-2201/4994.

Balirano, G., and M. Corduas. 2008. Detecting semiotically-expressed humor in diasporic TV productions. *HUMOR* 21: 227–251. https://doi.org/10.1515/HUMOR.2008.012.

Bonnini, S., D. Piccolo, L. Salmaso, and F. Solmi. 2012. Permutation inference for a class of mixture models. *Communications in Statistics—Theory and Methods* 41: 2879–2895. https://doi.org/10.1080/03610926.2011.590915.

Capecchi, S., I. Endrizzi, F. Gasperi, and D. Piccolo. 2016. A multi-product approach for detecting subjects' and objects' covariates in consumer preferences. *British Food Journal* 118: 515–526. https://doi.org/10.1108/BFJ-10-2015-0343.

Capecchi, S., and D. Piccolo. 2016. Investigating the determinants of job satisfaction of Italian graduates: A model-based approach. *Journal of Applied Statistics* 43: 165–179. https://doi.org/10.1080/02664763.2015.1036844.

———. 2017. Dealing with heterogeneity in ordinal responses. *Quality & Quantity* 51: 2375–2393. https://doi.org/10.1007/s11135-016-0393-3.

Capecchi, S., and R. Simone. 2019. A proposal for a model-based composite indicator: Experience on perceived discrimination in Europe. *Social Indicators Research* 141: 95–110. https://doi.org/10.1007/s11205-018-1848-9.

Capecchi, S., R. Simone, and S. Ghiselli. 2019. Drivers and uncertainty for job satisfaction of the Italian graduates. *Statistica Applicata – Italian Journal of Applied Statistics* 31: 227–250. https://doi.org/10.26398/IJAS.0031-013.

Cappelli, C., R. Simone, and F. Di Iorio. 2019. CUBREMOT: A tool for building model-based trees for ordinal responses. *Expert Systems with Applications* 124: 39–49. https://doi.org/10.1016/j.eswa.2019.01.009.

Contini, C., F. Boncinelli, L. Casini, G. Pagnotta, C. Romano, and G. Scozzafava. 2016. Why do we buy traditional foods? *Journal of Food Products Marketing* 22: 643–657. https://doi.org/10.1080/10454446.2016.1141137.

Corduas, M. 2011. Assessing similarity of rating distributions by Kullback–Leibler divergence. In *Classification and Multivariate Analysis for Complex Data Structures*, ed. B. Fichet, D. Piccolo, R. Verde, and M. Vichi, 221–228. Berlin: Springer. https://doi.org/10.1007/978-3-642-13312-1_22.

Corduas, M., L. Cinquanta, and C. Ievoli. 2013. The importance of wine attributes for purchase decisions: A study of Italian consumers' perception. *Food Quality and Preference* 28: 407–418. https://doi.org/10.1016/j.foodqual.2012.11.007.

Corduas, M., M. Iannario, and D. Piccolo. 2009. A class of statistical models for evaluating services and performances. In *Statistical Methods for the Evaluation of Educational Services and Quality of Products*, ed. P. Monari, M. Bini, D. Piccolo, and L. Salmaso, 99–117. Heidelberg: Physica-Verlag. https://doi.org/10.1007/978-3-7908-2385-1_7.

D'Elia, A. 2008. A statistical modelling approach for the analysis of TMD chronic pain data. *Statistical Methods in Medical Research* 17: 389–403. https://doi.org/10.1177/0962280206071846.

D'Elia, A., and D. Piccolo. 2005. A mixture model for preferences data analysis. *Computational Statistics & Data Analysis* 49: 917–934. https://doi.org/10.1016/j.csda.2004.06.012.

Di Nardo, E., and R. Simone. 2019. A model-based fuzzy analysis of questionnaires. *Statistical Methods & Applications* 28: 187–215. https://doi.org/10.1007/s10260-018-00443-9.

Fin, F., M. Iannario, R. Simone, and D. Piccolo. 2017. The effect of uncertainty on the assessment of individual performance: Empirical evidence from professional soccer. *Electronic Journal of Applied Statistical Analysis* 10: 677–692. https://doi.org/0.1285/i20705948v10n3p677.

Finch, W. H., and M. E. Hernández Finch. 2020. Modeling of self-report behavior data using the generalized covariates in a uniform and shifted binomial mixture model: An empirical example and Monte Carlo simulation. *Psychological Methods* 25: 113–127. https://doi.org/10.1037/met0000225.

Hernández Barajas, F., O. C. Usuga Manco, and S. García Muñoz. 2018. cubm package in R to fit CUB models. *Comunicaciones en Estadística* 11: 219–238. https://doi.org/10.15332/2422474x.3857.

Hu, C., H. Zhou, and A. Sharma. 2020. Application of beta-distribution and combined uniform and binomial methods in longitudinal modeling of bounded outcome score data. *AAPS Journal* 22: 95. https://doi.org/10.1208/s12248-020-00478-5.

Hwang, S., S. H. Sohn, and C. Oh. 2015. Maximum likelihood estimation for a mixture distribution. *Journal of the Korean Data and Information Science Society* 26: 313–322. https://doi.org/10.7465/jkdi.2015.26.2.313.

Iannario, M. 2010. On the identifiability of a mixture model for ordinal data. *Metron* LXVIII: 87–94. https://doi.org/10.1007/BF03263526.

———. 2012. Modelling shelter choices in a class of mixture models for ordinal responses. *Statistical Methods & Applications* 21: 1–22. https://doi.org/10.1007/s10260-011-0176-x.

Iannario, M., M. Manisera, D. Piccolo, and P. Zuccolotto. 2012. Sensory analysis in the food industry as a tool for marketing decisions. *Advances in Data Analysis and Classification* 6: 303–321. https://doi.org/10.1007/s11634-012-0120-4.

Iannario, M., D. Piccolo, and R. Simone. 2020. cub: A class of mixture models for ordinal data. R package version 1.1.4. https://CRAN.R-project.org/package=CUB.

Jenkins, S. P. 2020. Comparing distributions of ordinal data. *Stata Journal* 20: 505–531. https://doi.org/10.1177/1536867X20953565.

Low, Y. C. 2017. Statistical modeling for review ratings data. *International Journal of Knowledge Engineering* 3: 48–51. https://doi.org/10.18178/ijke.2017.3.2.086.

Manisera, M., and P. Zuccolotto. 2014. Modeling "don't know" responses in rating scales. *Pattern Recognition Letters* 45: 226–234. https://doi.org/10.1016/j.patrec.2014.04.012.

Mauracher, C., I. Procidano, and G. Sacchi. 2016. Wine tourism quality perception and customer satisfaction reliability: The Italian Prosecco District. *Journal of Wine Research* 27: 284–299. https://doi.org/10.1080/09571264.2016.1211514.

McCullagh, P. 1980. Regression models for ordinal data. *Journal of the Royal Statistical Society, Series B* 42: 109–142. https://doi.org/10.1111/j.2517-6161.1980.tb01109.x.

McLachlan, G. J., and T. Krishnan. 2008. *The EM Algorithm and Extensions*. 2nd ed. Hoboken, NJ: Wiley.

McLachlan, G. J., and D. Peel. 2000. *Finite Mixture Models*. New York: Wiley.

Piccolo, D. 2003. On the moments of a mixture of uniform and shifted binomial random variables. *Quaderni di Statistica* 5: 85–104.

———. 2006. Observed information matrix for MUB models. *Quaderni di Statistica* 8: 33–78.

———. 2015. Inferential issues for CUBE models with covariates. *Communications in Statistics—Theory and Methods* 44: 771–786. https://doi.org/10.1080/03610926.2013.821487.

Piccolo, D., and A. D'Elia. 2008. A new approach for modelling consumers' preferences. *Food Quality and Preference* 19: 247–259. https://doi.org/10.1016/j.foodqual.2007.07.002.

Piccolo, D., and R. Simone. 2019a. The class of CUB models: Statistical foundations, inferential issues and empirical evidence (with discussions). *Statistical Methods & Applications* 28: 389–475. https://doi.org/10.1007/s10260-019-00461-1.

———. 2019b. Rejoinder to the discussion of "The class of CUB models: Statistical foundations, inferential issues and empirical evidence". *Statistical Methods & Applications* 28: 477–493. https://doi.org/10.1007/s10260-019-00479-5.

Piccolo, D., R. Simone, and M. Iannario. 2019. Cumulative and CUB models for rating data: A comparative analysis. *International Statistical Review* 87: 207–236. https://doi.org/10.1111/insr.12282.

Schwarz, G. 1978. Estimating the dimension of a model. *Annals of Statistics* 6: 461–464. https://doi.org/10.1214/aos/1176344136.

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.

———. 2021. An accelerated EM algorithm for mixture models with uncertainty for rating data. *Computational Statistics* 36: 691–714. https://doi.org/10.1007/s00180-020-01004-z.

Simone, R., C. Cappelli, and F. Di Iorio. 2019. Modelling marginal ranking distributions: The uncertainty tree. *Pattern Recognition Letters* 125: 278–288. https://doi.org/10.1016/j.patrec.2019.04.026.

Simone, R., F. Di Iorio, and R. Lucchetti. 2019. CUB for GRETL. In *Gretl 2019: Proceedings of the International Conference on the GNU Regression, Econometrics and Time Series Library*, ed. F. Di Iorio and R. Lucchetti, 147–166. Napoli, Italy: Federico II Open Access University Press.

Simone, R., G. Tutz, and M. Iannario. 2020. Subjective heterogeneity in response attitude for multivariate ordinal outcomes. *Econometrics and Statistics* 14: 145–158. https://doi.org/10.1016/j.ecosta.2019.04.002.

Tutz, G. 2012. *Regression for Categorical Data*. Cambridge: Cambridge University Press.

Xu, H., and N. Zhang. 2021. From contextualizing to context-theorizing: Assessing context effects in privacy research. https://doi.org/10.2139/ssrn.3624056.

About the authors

Giovanni Cerulli is a senior researcher at the CNR-IRCrES, Research Institute on Sustainable Economic Growth, National Research Council of Italy, Rome. His research interest is in applied econometrics, with a special focus on causal inference and machine learning. He has developed original causal inference models and provided several implementations. He is currently editor in chief of the *International Journal of Computational Economics and Econometrics*.

Rosaria Simone is an assistant professor in statistics at the Department of Political Sciences of the University of Naples Federico II, in Italy. Her main research interests involve computational statistics, statistical modeling, and classification for categorical data, with applications focusing mainly on ratings and rankings.

Francesca Di Iorio is an associate professor in economic statistics at the Department of Political Sciences, University of Naples Federico II. She is an associate editor of the *Journal of Official Statistics*. Her research focuses mainly on bootstrap inference, computational solutions for indirect inference, time-series analysis, and regression trees.

Domenico Piccolo is Professor Emeritus of Statistics at University of Naples Federico II. During a long teaching career, he promoted graduate and postgraduate curricula in statistics. He introduced the autoregressive metric among time series, chaired the main projects of seasonal adjustment in Italy, and is currently the main supporter of the class of CUB models.

Christopher F. Baum is a professor of economics and social work at Boston College, where he chairs the department of economics. He is an associate editor of the *Stata Journal*. Baum founded and manages the Boston College Statistical Software Components (SSC) Archive at RePEc (http://repec.org). He is the author or coauthor of several community-contributed commands and three books on Stata.