An Inflated Model to Account for Large Heterogeneity in Ordinal Data

Stefania Capecchi, Rosaria Simone, and Domenico Piccolo

Abstract In sample surveys where people are asked to express their opinions, a high level of indecision among respondents may generate sub-optimal statistical analyses caused by large heterogeneity in the responses. We discuss a model belonging to the class of generalized CUB models that is suitable for this kind of surveys. Then, we examine a real case study where the observed heterogeneity as well as respondents' indecision can be analyzed within the theoretical framework of the proposed model leading to convincing interpretations. A comparison with current literature and some concluding remarks end the paper.

1 Introduction

Ordinal data are common in many fields involving Social Sciences and Humanities, Medicine and Marketing, among others. As a matter of fact, this typology of responses is collected in sample surveys concerning personal beliefs, habits, opinions, preferences, tastes, political orientation, well-being, work related issues, job satisfaction, etc. In such cases, although quite often the answer modalities are designed to be expressed by natural numbers to simplify coding and synthesis, ordinal data convey categorical information; thus, adequate statistical analysis should be performed to avoid oversimplified interpretations [1, 27].

The *response style* is a prominent issue in psychological and marketing studies when people respond to questionnaires according to a subjective disposition that could hide the "true" score. The issue is widely investigated, in different disciplines and fields. Generally, it has been argued [2, 22] that the response styles are tendencies to respond systematically to questionnaire items on the basis other than what the items were specifically designed to measure; that is, very briefly, a tendency to provide responses to the questionnaire items regardless of their specific content. Thus, a significant proportion of respondents tends to use only a smaller number of the rating scale options [9].

S. Capecchi (✉) • R. Simone • D. Piccolo
Department of Political Sciences, University of Naples Federico II, Via L. Rodinò 22, I-80138 Naples, Italy
e-mail: stefania.capecchi@unina.it; rosaria.simone@unina.it; domenico.piccolo@unina.it

F. Palumbo et al. (eds.), *Data Science*, Studies in Classification, Data Analysis, and Knowledge Organization, DOI 10.1007/978-3-319-55723-6_16

More specifically, we discuss a distinctive typology of responses characterized by a sharp preference towards the item (expressed by a proportion of the respondents) and by diffuse indecision manifested by a large group of the other interviewees. For a number of different motivations, a quota of interviewees selects a specific category which can be considered a sort of “refuge” (*shelter option*). According to different circumstances, this behavior may be caused by indecision, desire of privacy, or a real approval/disapproval of the item (for extreme categories) and it can be referred to one of the most common response styles.

If the item concerns evaluation/judgement which is of interest for a limited portion of the population, responses are concentrated in just one or very few categories. As an instance, consider a questionnaire designed to analyze relational goods and leisure habits. In this case, it has been asked “How important for you is to walk the dog” with an ordinal rating scale from 1 to 10; then, it is common to observe a very high frequency of respondents selecting 1 (since they don’t have a dog) and low frequencies of responses spread over the other categories. A sample of this evidence from real sample surveys for different topics is presented in Fig. 1.

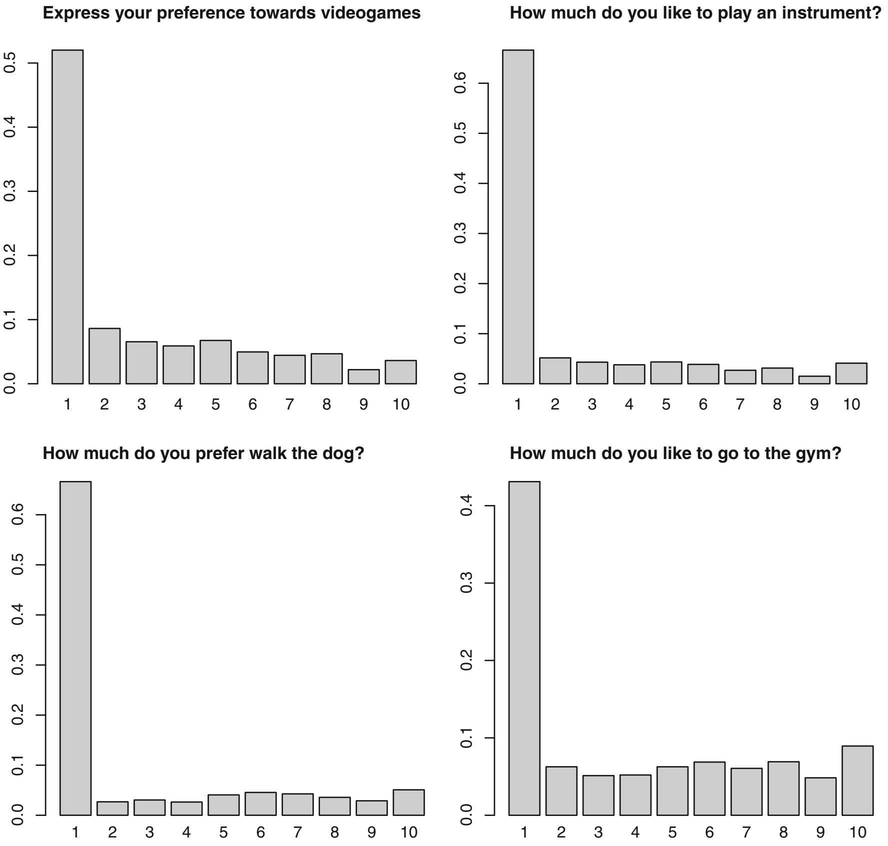

Fig. 1 Observed distributions with large heterogeneity

Similarly, in working conditions surveys, if some of the items concern the possibility of workers to select their own partners for a specific job/duty, most of the people would select the category "Never" (a large proportion of respondents indeed does not have the chance to choose their job partners or co-workers) whereas all the others would pick different categories. It is evident that without a specific approach, the analysis of these responses may cause bias in the interpretation.

Thus, it seems useful to take these information into specific account and search for adequate statistical modelling of this biased behavior and to relate it to subjects' characteristics when significant. In this regard a focal point is the relationship between individual indecision and observed heterogeneity of the responses. In fact, if the reaction of the interviewees is substantially homogeneous with respect to the item and their responses are mostly concentrated in few categories, then we observe a very low heterogeneity; on the contrary, interviewees express quite different responses when their indecision is high and thus we notice high heterogeneity in the observed distribution.

The paper is organized as follows: in the next section, we motivate the introduction of a parsimonious parametrization in modelling ordinal data in presence of large heterogeneity; then, in Sect. 3 we discuss the inferential issues related to estimates and test of the parameters that characterize the model. In Sect. 4 a real case study is presented and the role of significant covariates is discussed. Some concluding remarks end the work.

2 A Model for Large Heterogeneity

Motivated by psychological aspects of the decisional process, an alternative framework for modelling preferences and evaluations expressed as ordinal data has been proposed [5, 23, 24]. The main features of the approach are: parsimony in the number of parameters, high flexibility of shapes in the derived distributions, sharp visualization of the estimated models (an open source software is available as R package: [14]).

More specifically, the generating process leading to ordinal data is interpreted as a mixture where the indecision of the choice and the attractiveness/repulsion towards the item are explicitly modelled according to discrete distributions [13]. This class of models includes several generalizations and extensions to cope with real situations [10, 12, 19, 20, 25].

A model may be specified as the mixture of two distributions allowing for both *shelter effect* and extreme uncertainty, respectively, in case data originating from sample surveys suggest a large indecision among respondents (except for the ones who do not select a response category to be interpreted as a *shelter* one: [11]). This model is a *C*ombination of a discrete *U*niform distribution with a *SH*elter effect and will be called CUSH model.

If R is the ordinal random variable defined on the support $\{1, 2, \ldots, m\}$, for a given m and $c \in \{1, 2, \ldots, m\}$ is the known location of the *shelter effect*, a CUSH model is defined by:

$$Pr(R = r \mid \boldsymbol{x}_i) = \delta_i D_r^{(c)} + (1 - \delta_i)\,\frac{1}{m}\,, \quad r = 1, 2, \ldots, m; \quad logit(\delta_i) = \boldsymbol{x}_i \boldsymbol{\omega}\,; \quad i = 1, \ldots, n\,. \tag{1}$$

where $D_j^{(c)} = I(r = c)$ and $I(A)$ is the indicator function such that is 1 when A is true and is 0 elsewhere. The row vector $\boldsymbol{x}_i = (x_{i0}, x_{i1}, \ldots, x_{is})$ includes the explanatory covariates of the *shelter effect* for the ith subject, with the convention: $x_{i0} = 1$, $i = 1, \ldots, n$.

If the previous model is conditioned on a specific pattern of the subjects' covariates (so that $\delta_i = \delta$), a CUSH model without covariates is:

$$Pr\,(R = r \mid \delta) = \delta\, D_r^{(c)} + (1 - \delta)\,\frac{1}{m}\,, \qquad \delta \in [0, 1]\,. \tag{2}$$

As a matter of fact, δ measures the differential effect of a preferred category with respect to all the others. A CUSH model may be considered as a *c-inflated* model with respect to the discrete Uniform distribution in the same line of reasoning leading to the well-known zero-inflated models for Poisson [16] and Negative Binomial distributions [8], for instance. In Fig. 2 CUSH models of different shapes are obtained by varying c and δ.

For given m and c, a CUSH model is fully characterized by the parameter δ; thus, expectation and variance are given by:

$$\mathbb{E}\,(R) = \delta\, c + (1 - \delta)\,\frac{m + 1}{2}; \qquad Var(R) = (1 - \delta)\left[\delta\left(c - \frac{m + 1}{2}\right)^2 + \frac{m^2 - 1}{12}\right].$$

The mean value is a convex combination of the mean of the discrete Uniform distribution and a degenerate random variable at the shelter category, whereas the variance is a parabolic function whose maximum depends on m and c. The variance is 0 when $\delta = 1$ since the CUSH model collapses to a degenerate distribution at $R = c$.

More noticeably, the (normalized) heterogeneity [7] index turns out to be:

$$G = \frac{m}{m - 1}\left(1 - \sum_{r=1}^{m} [Pr(R = r)]^2\right) \;=\; 1 - \delta^2\,;$$

thus, it is independent of the location c of the shelter and also of the number m of categories. As a consequence, δ may be interpreted as an inverse measure of heterogeneity: for increasing δ the probability mass distribution becomes more and more concentrated on the shelter category.

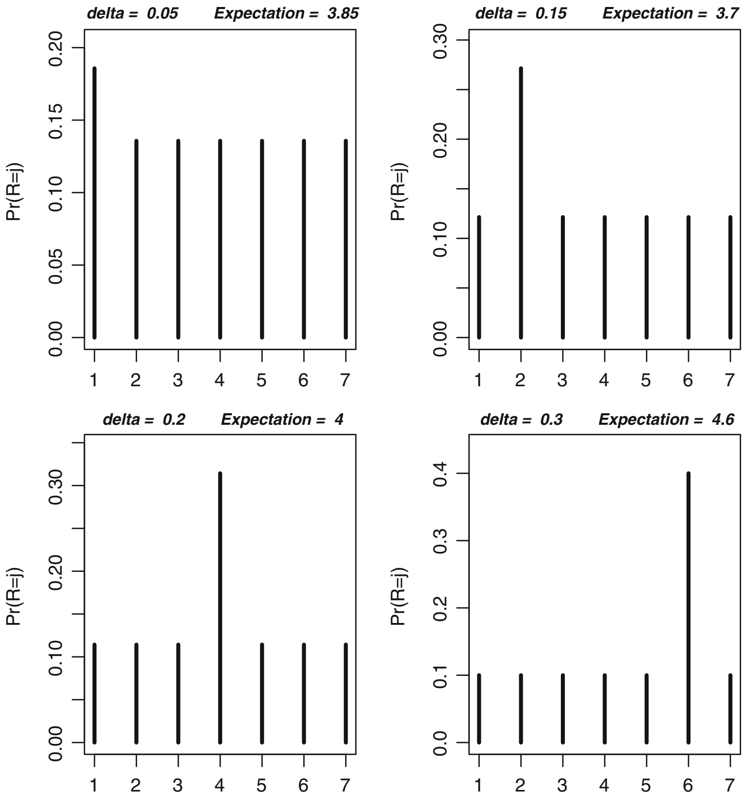

Fig. 2 A sample of typologies for CUSH distributions

However, Gini index is not so selective (it is invariably high in most of the circumstances) and this invariance with respect to m may be a limitation. Thus, the normalized [15] index may be preferred. In general, and for CUSH models, this index turns out to be:

$$\mathscr{H} = \frac{1}{m-1}\left[\left(\sum_{i=1}^{m}[Pr(R=r)]^2\right)^{-1} - 1\right] = \frac{1-\delta^2}{1+(m-1)\,\delta^2}\,.$$

This measure is monotonically related to the Gini index, since $\mathscr{H} = G/[m-G(m-1)]$, but it is more selective and depends on m.

It is possible to derive preliminary estimators of δ from the sample Gini index, and this measure may be useful also for the selection of covariates in the logit link for the model (1) as fully discussed in [3].

3 Inferential Issues for CUSH Models

If we denote by $(f_1, \dots, f_m)'$ the vector of the relative frequencies obtained from the sample of ordinal data $(r_1, \dots, r_n)'$, the Maximum Likelihood (ML) estimator of δ is obtained by maximizing the log-likelihood function:

$$\ell(\delta) = \sum_{i=1}^{n} \log\left(Pr\left(R = r_i \mid \delta\right)\right) = \sum_{r=1}^{m} n_r \log\left(Pr\left(R = r \mid \delta\right)\right).$$

It is simple to prove that ML estimator exists, it is unique and defined by:

$$\hat{\delta} = \begin{cases} \frac{m f_c - 1}{m-1} = \frac{f_c - 1/m}{1-1/m}, & \text{if } f_c \geq 1/m; \\ 0, & \text{otherwise.} \end{cases}$$

Since for common values of n the probability to observe $f_c < 1/m$ is virtually 0, in the following we consider the first expression as the ML estimator. Notice that $\hat{\delta}$ has a simple interpretation: it compares the relative frequency f_c at the *shelter category c* with the discrete Uniform hypothesis ($f_c = 1/m$), and then normalizes this difference.

It is possible to prove that the ML estimator $\hat{\delta}$ is an unbiased estimator of δ with a variance given by:

$$Var(\hat{\delta}) = \frac{1}{n}\,\frac{1-\delta}{m-1}\,[1 + (m-1)\delta]\,.$$

The standard error of $\hat{\delta}$ is evaluated by plugging the ML estimate into the last expression:

$$es(\hat{\delta}) = \sqrt{\frac{1}{n}\,\frac{1-\hat{\delta}}{m-1}\,[1 + (m-1)\hat{\delta}]} = \frac{f_c}{\sqrt{n}}\,\sqrt{\frac{m}{m-1}}\,.$$

In order to test the presence of a *shelter effect*, both Wald and Likelihood Ratio test may be exploited, and they are asymptotically equivalent. Thus, the Wald test for checking $H_0 : \delta = 0$ is:

$$W = \frac{\hat{\delta}}{es(\hat{\delta})} = \sqrt{n}\,\frac{f_c\, m - 1}{f_c\,\sqrt{m(m-1)}}\,. \qquad (3)$$

After some algebra, the maximized log-likelihood function turns out to be:

$$\ell(\hat{\delta}) = n\left[f_c \log(f_c) + (1 - f_c)\log\left(\frac{1-f_c}{m-1}\right)\right].$$

Since the log-likelihood function under H_0 is $\ell(0) = -n \log(m)$, the Likelihood Ratio test (*LRT*) is:

$$LRT = 2\left(\ell(\hat{\delta}) - \ell(0)\right) = 2\,n\left[f_c \log(f_c) + (1-f_c)\log\left(\frac{1-f_c}{m-1}\right) + \log(m)\right]. \quad (4)$$

As a function of f_c, for a given c, *LRT* is U-shaped with a unique minimum at $f_c = 1/m$. So, for a given α-level, the critical region $LRT > c_{\alpha;n}$ is strictly equivalent to $\mid f_c - 1/m \mid > d_{\alpha;n}$, for a convenient α. In fact, it seems reasonable to reject the hypothesis of a null *shelter effect* when the relative frequency at the category c is significantly greater than the expected totally random proportion $1/m$. Figure 3 depicts this situation in a specific case.

However, when we test a borderline hypothesis (as $H_0 : \delta = 0$) the asymptotic distribution of the *LRT* (4) does not converge to a $\chi^2_{(1)}$ random variable [21, 26, 28] as predicted by the asymptotic standard theory. An acceptable approximate solution is to halve the p-value of a $\chi^2_{(1)}$ distribution and to check this simplification for finite sample sizes by a simulation. For finite sample sizes, this simulation experiment has been performed for varying values of n and m and *LRT* is preferred in regular situations; more evidence of these experiments is reported in [4].

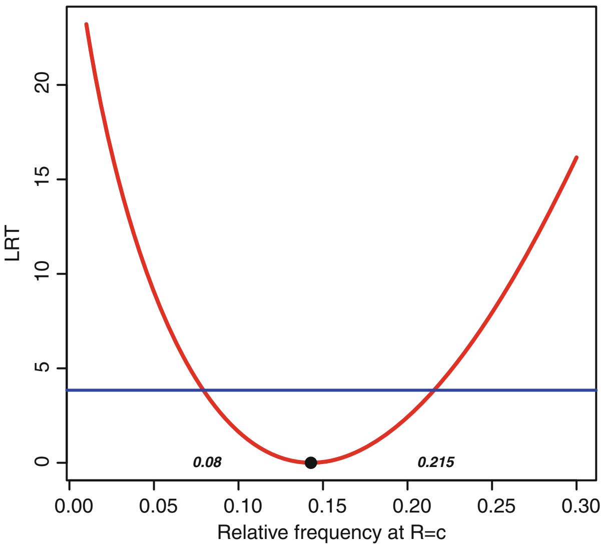

Fig. 3 *LRT* as function of the relative frequency at $R = c$, for a given c ($n = 100$, $m = 7$)

4 A Case Study

Every 5 years, Eurofound carries out the European Working Condition Survey (EWCS) which involves employees and self-employed people with interviews concerning their work and employment conditions across Europe [6].

Data of interest originate from the fifth EWCS and we consider the EU28 Member States only. The sample (a multistage stratified random sampling design) is representative of those aged 15 years and over who are employed and resident in the country being surveyed. The target number of interviews was greater or equal to 1000 in all the countries and global results are based on $n = 31{,}689$ respondents.

The responses to the item: "You have a say in the choice of your working partners" with possible responses: *Never, Rarely, Sometimes, Most-of-the-time, Always* (coded as 1–5) are mostly investigated at an aggregate level. Then, a synthesis of the results for the EU28 Member States is shown. The observed distribution of responses to "*Have a say* ... " is plotted in Fig. 4.

With reference to the selected modelling framework, in Table 1 several alternative and comparable models are presented. First of all, a CUB model without covariates (which is a sort of benchmark for this kind of analysis), then a CUB model

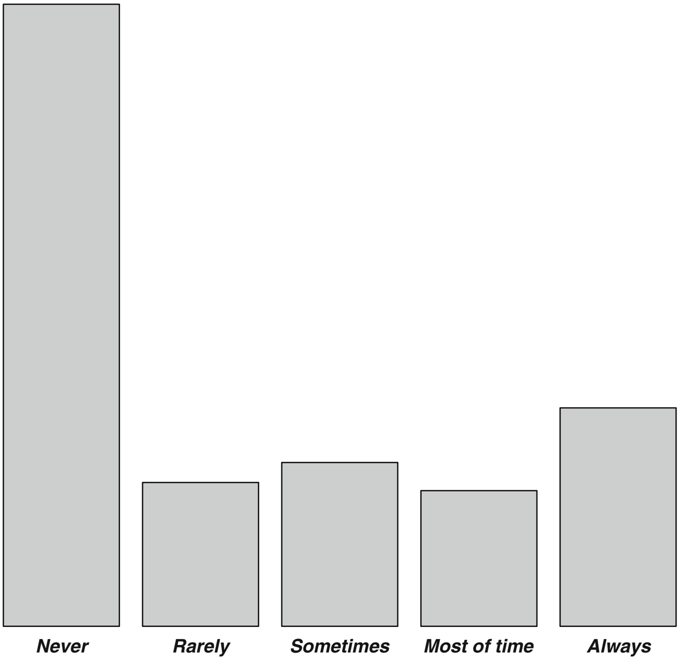

Fig. 4 Frequency distribution of responses to "*Have a say* ... " item

Table 1 Alternative mixture models with uncertainty for responses to *Have a say* . . .

Models	Characteristics	*Log − lik*	*BIC*
CUB	*no covariates*	−44595.07	89210.87
CUB +*sh*	*shelter at* R=1	−44595.07	89221.24
CUSH	*no covariates*	−44595.07	89200.50
CUSH +*cov*	**Gender, Lage, Univ, Private**	−44046.38	88154.95

with the inclusion of a shelter effect at $R = 1$. Finally, CUSH models without and with covariates are compared.

The first three models are virtually equivalent since all of them capture a dominant frequency of responses located at $R = 1$ (that is, the proportion of interviewees selecting "Never") and CUSH model should be preferred by a parsimony criterion (as confirmed by the *BIC* index). Then, among the several covariates available in the dataset and able to explain the δ values in subgroups, the following ones have been found significant:

$$\texttt{Lage}_i = \text{deviation from the mean of the logged age in years}$$

$$\texttt{Gender}_i = \begin{cases} 0\,, \text{ if the } i\text{th subject is a man;} \\ 1\,, \text{ if the } i\text{th subject is a woman.} \end{cases}$$

$$\texttt{Univ}_i = \begin{cases} 0\,, \text{ if the } i\text{th subject has not a University education;} \\ 1\,, \text{ if the } i\text{th subject has a University education.} \end{cases}$$

$$\texttt{Private}_i = \begin{cases} 0\,, \text{ if the } i\text{th subject does not work in the private sector;} \\ 1\,, \text{ if the } i\text{th subject works in the private sector.} \end{cases}$$

Given the selected covariates $\boldsymbol{x}_i = (Lage_i, Gender_i, Univ_i, Private_i)$ and within the class of CUSH models, the best result is given by the following *stochastic* and *systematic* components, respectively (standard errors in parentheses):

$$Pr(R = r \mid \boldsymbol{x}_i) = \delta_i D_r^{(1)} + (1 - \delta_i)\,\frac{1}{5}\,; \quad r = 1, 2, \ldots, 5; \;\; i = 1, 2, \ldots, n;$$

$$\begin{aligned} logit(\delta_i) = & \underset{(0.038)}{-0.562} + \underset{(0.032)}{0.544}\, Gender_i \underset{(0.039)}{-0.870}\; Univ_i \underset{(0.034)}{-0.295}\; Private_i \\ & \underset{(0.056)}{-0.294}\, Lage_i + \underset{(0.139)}{1.380}\; Lage_i^2\,; \qquad i = 1, \ldots, n\,. \end{aligned}$$

The *shelter effect* shows a negative impact (that is, an increasing probability of a "Never" response) when the respondent is a woman, whereas the impact is positive for people working in the private sector as well as for the ones educated at a University level, mostly. The effect of Lage is parabolic and globally positive since higher ages correspond to a reduced probability of "Never" responses (Fig. 5).

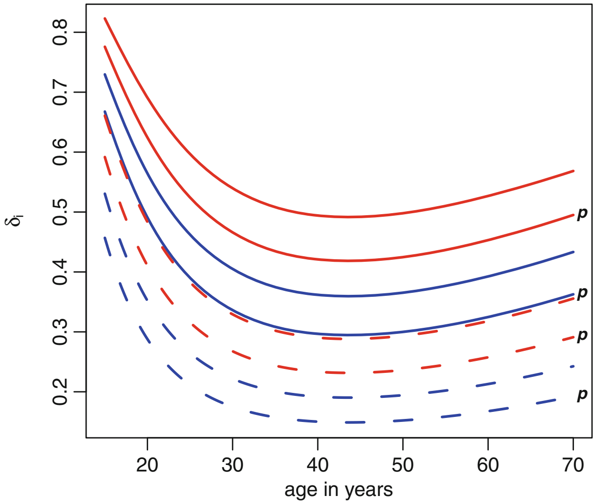

Fig. 5 Estimates of the *shelter effect* for "*Have a say* . . . ". *Broken lines* denote education at University level; women-lines are systematically above the corresponding men-lines; *p-lines* are for people working in private sector

If we compare the previous estimated models with the standard approach—as the proportional odds model (POM) [17, 18]—we get the same significant covariates and a better result in terms of log-likelihood function caused by an increase in the number of parameters to be estimated (9 instead of 6). Indeed, POM models imply a local fit for each frequency whereas CUSH models involve a global fitting and thus they are by far more parsimonious. In addition, the relationship between the *shelter effect* and the covariates may be sharply depicted by CUSH models as shown in Fig. 5.

The estimated CUSH model has been replicated for every country by considering if and where the impact of the selected covariates was homogeneous and significant. This long modelling exercise may be comparatively presented in a series of panels where the impact of each covariate on the *shelter effect* [according to a logit link as in (1)] is shown with the asymptotic level to check for its significance. Figure 6 presents in four panels the effects of the covariates `Gender`, `Lage`, `University`, and `Private`, respectively, on the item "*Have a say*. . . ".

It is evident that `Gender` and `University` have a substantially homogeneous effect which is almost always significant in all countries and in the same direction; less evident are the influences of `Private` and `Lage` which seem to have an important impact only in some specific countries.

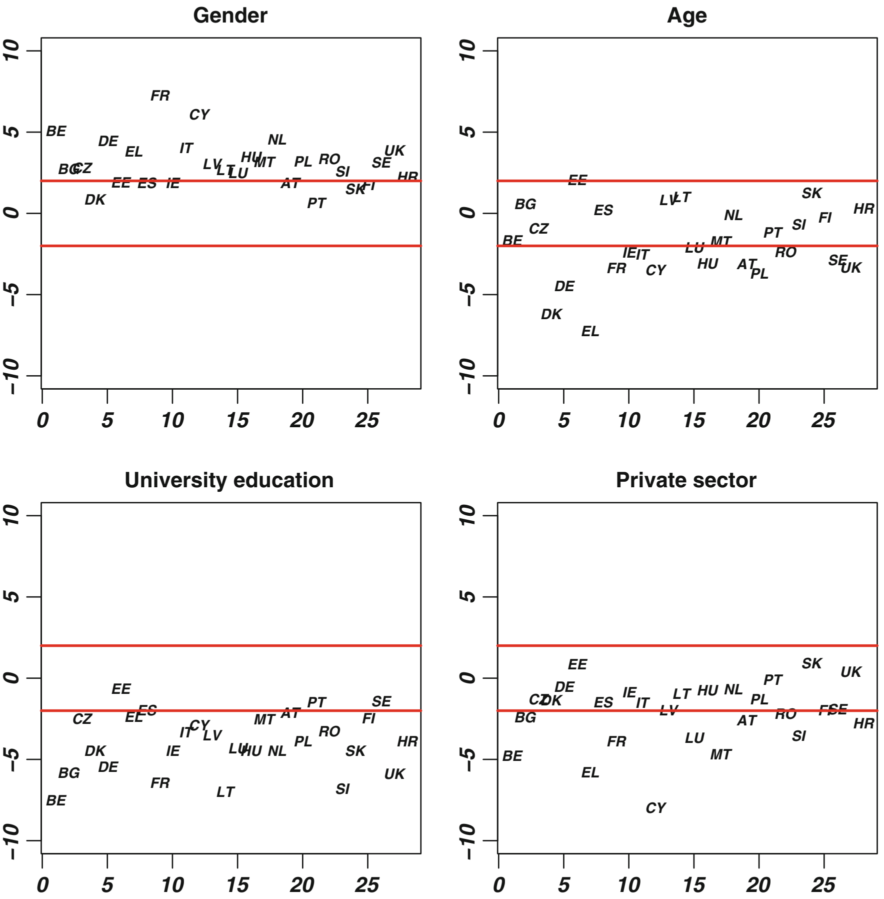

Fig. 6 Differential *shelter effect* on covariates by countries

5 Conclusions

CUSH models are a very simple probability structure introduced to interpret real case studies characterized by large heterogeneity and uncertainty in the responses. Effective methods based on likelihood paradigm can be applied for both estimation and testing. This class of model is particularly parsimonious and offers a sharp graphical representation of the covariates on the inflated category. Some empirical evidence confirms the usefulness of the proposal.

Further open issues worth to be explored include the research of effective tools to select covariates and the investigation of the predictability of these models. Finally, some multivariate suggestions could be very attractive when several items have to be compared and jointly examined.

Acknowledgements This work has been realized within the FIRB2012 project (code RBFR12SHVV) at University of Perugia and the frame of Programme STAR (CUP E68C13000020003) at University of Naples Federico II, financially supported by UniNA and Compagnia di San Paolo.

References

1. Agresti, A.: Analysis of Ordinal Categorical Data, 2nd edn. Wiley, Hoboken (2010)
2. Baumgartner, H., Steenback, J.-B.E.M.: Response styles in marketing research: a cross-national investigation. J. Mark. Res. **38**, 143–156 (2001)
3. Capecchi, S., Iannario, M.: Gini heterogeneity index for detecting uncertainty in ordinal data surveys. METRON (2016). doi:10.1007/s40300-016-0088-5
4. Capecchi, S., Piccolo, D.: Dealing with heterogeneity in ordinal responses. Qual. Quant. (2016). doi:10.1007/s11135-016-0393-3
5. D'Elia, A., Piccolo, D.: A mixture model for preference data analysis. Comput. Stat. Data Anal. **49**, 917–934 (2005)
6. Eurofound: Fifth European Working Conditions Survey. Publications Office of the European Union, Luxembourg (2012)
7. Gini, C.: Variabilità e mutabilità. Studi economico-giuridici, Facoltà di Giurisprudenza, Università di Cagliari, A, III, parte II (1912)
8. Greene, W.H.: Some accounting for excess zeros and sample selection in Poisson and negative binomial regression models. Working Paper EC-94-10, Department of Economics, New York University (1994)
9. Harzing, A.W.K.: Response styles in cross-national survey research: a 26-country study. Int. J. Cross Cult. Manag. **6**, 243–266 (2006)
10. Iannario, M.: Hierarchical CUB models for ordinal variables. Commun. Stat. Theory Methods **41**, 3110–3125 (2012)
11. Iannario, M.: Modelling shelter choices in a class of mixture models for ordinal responses. Stat. Methods Appl. **21**, 1–22 (2012)
12. Iannario, M.: Modelling uncertainty and overdispersion in ordinal data. Commun. Stat. Theory Methods **43**, 771–786 (2014)
13. Iannario, M., Piccolo, D.: CUB models: statistical methods and empirical evidence. In: Kenett, R.S., Salini, S. (eds.) Modern Analysis of Customer Surveys: With Applications Using R, pp. 231–258. Wiley, Chichester (2012)
14. Iannario, M., Piccolo, D., Simone, R.: CUB : a class of mixture models for ordinal data (2016). R package version 0.1. http://CRAN.R-project.org/package=CUB
15. Laakso, M., Taagepera, R.: Effective number of parties: a measure with application to West Europe. Comp. Pol. Stud. **12**, 3–27 (1989)
16. Lambert, D.: Zero-inflated poisson regression, with an application to defects in manufacturing. Technometrics **34**, 1–14 (1992)
17. McCullagh, P.: Regression models for ordinal data (with discussion). J. R. Stat. Soc. Ser. B **42**, 109–142 (1980)
18. McCullagh, P., Nelder, J.A.: Generalized Linear Models, 2nd edn. Chapman & Hall, London (1989)
19. Manisera, M., Zuccolotto, P.: Modeling rating data with Nonlinear CUB models. Comput. Stat. Data Anal. **78**, 100–118 (2014)
20. Manisera, M., Zuccolotto, P.: Modelling "Don't know" responses in rating scales. Pattern Recogn. Lett. **45**, 226–234 (2014)
21. Molenberghs, G., Verbeke, G.: Likelihood ratio, score, and Wald tests in a constrained parameter space. Am. Stat. **61**, 22–27 (2007)

22. Moors, G.: Exploring the effect of a middle response category on response style in attitude measurement. Qual. Quant. **42**, 779–794 (2008)
23. Piccolo, D.: On the moments of a mixture of uniform and shifted binomial random variables. Quaderni di Statistica **5**, 85–104 (2003)
24. Piccolo, D.: Observed information matrix in MUB models. Quaderni di Statistica **8**, 33–78 (2006)
25. Piccolo, D.: Inferential issues on CUBE models with covariates. Commun. Stat. Theory Methods **44**, 5023–5036 (2015)
26. Self, S.G., Liang, K.Y.: Asymptotic properties of maximum likelihood estimators and likelihood ratio test under nonstandard conditions. J. Am. Stat. Assoc. **82**, 605–610 (2003)
27. Tutz, G.: Regression for Categorical Data. Cambridge University Press, Cambridge (2012)
28. Vu, H.T.V., Zhou, S.: Generalization of likelihood ratio tests under nonstandard conditions. Ann. Stat. **25**, 897–916 (1997)