Quaderni di Statistica
Vol. 11, 2009

Generalized residuals in CUB models

Francesca Di Iorio Domenico Piccolo
Dipartimento di Scienze Statistiche, Università di Napoli Federico II
E-mail: fdiiorio@unina.it, domenico.piccolo@unina.it

Summary: A relevant issue in regression studies is the model diagnostics performed by using estimated residuals and their transformations. Such tools cannot be exploited in Generalized Linear Models framework where a decomposition of response into fitted and error components is no more available; thus, some kind of generalized residuals are derived by first-order conditions of maximum likelihood equations. In this paper, we will introduce generalized residuals for CUB models, discuss their main characteristics and test their usefulness on two real data sets. This approach brings out as a noticeable feature the possibility to detect a differential effect of covariates on the probability of choice for ordinal categories.

Keywords: Generalized residuals, CUB models, Model diagnostics

1. Introduction

In regression models, residuals analysis is a key tool for different purposes: validating the structure of the model, checking the nature and persistence of the postulated dependence, detecting outliers and/or influential data, assessing the validity of classical linear hypotheses (homoscedastic and uncorrelated errors) and, finally, when necessary, verifying distributional assumptions (Gaussianity, skewness, heavy tails, and so on). A relevant issue is the study of residuals pattern in order to find omitted covariates in the models. Moreover, by exploiting their variability and distributional properties, standard R^2 measures and F tests are derived together with several generalizations (Magee, 1990).

These objectives are strongly interrelated and graphical tools, trans-

formations and formal tests have been proposed in current literature for assessing one or few of them (Atkinson, 1985; Belsey *et al.*, 1980; Cook and Weisberg, 1994; Fox, 1991; 1997, 267-366; Mosteller and Tukey, 1968; Seber, 1977).

However, when the response variable is not continuous and standard regression paradigm is converted to the Generalized Linear Models (GLM) framework, the definition of residuals is not so evident. In fact, it is not possible to rely on a simple decomposition as:

$$\textit{response=expectation} + \textit{error} \Longleftrightarrow \textit{observation=fitted} + \textit{residual}\,,$$

which is regularly assumed in classical linear models. In this vein, several proposals have been advanced in order to mimic the main assumptions of regression models leading to Pearson, Anscombe and deviance residuals, respectively (McCullagh and Nelder, 1989, 37-40; 396-415), among others.

This kind of problems becomes more awkward when studying models with qualitative data, specifically, ordinal data since there is no *natural* definition of residuals for these models. Indeed, the very concept of "outlier" is not so evident for ordinal data as the admissible range of observation is finite and discrete. Thus, statisticians should look for more general definitions that hopefully preserve some of the fundamental requirements of standard residuals.

The paper is organized as follows: in section 2, we briefly review a likelihood-based approach to generalized residuals and in section 3 we will formally derive these residuals for a peculiar class of ordinal models, called CUB . Then, in section 4 we will apply this new definition to empirical data in order to highlight limits and usefulness of the proposal; specifically, we will enhance a noticeable feature of this approach in detecting differential effects of covariates on the probabilities of choice. Some final remarks end the paper.

2. Likelihood-based definition of generalized residuals

In a standard regression model: $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, the Ordinary Least Squares (OLS) solutions for the $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$ parameters vector

imply that the estimated residuals: $e_i = y_i - \boldsymbol{x}_i\hat{\boldsymbol{\beta}}$, for $i = 1, 2, \ldots, n$, should obey the requirements:

$$\sum_{i=1}^{n} \boldsymbol{x}_i' \, (y_i - \boldsymbol{x}_i\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{n} \boldsymbol{x}_i' \, e_i = \mathbf{0} \, . \tag{1}$$

or, explicitly,

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} x_{i1} \, e_i = \cdots = \sum_{i=1}^{n} x_{ip} \, e_i = 0 \, .$$

where $\boldsymbol{x}_i = (x_{i0}, x_{i1}, \ldots, x_{ip})$ and $x_{i0} = 1$, for $i = 1, 2, \ldots, n$.

If we adopt independence and Gaussianity for the random vector $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)'$, the same orthogonality constraint (1) is derived by assuming that score functions are identically 0, that is:

$$\frac{\partial \, \log L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \frac{\partial \, \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \mathbf{0} \, ,$$

where we denote by $L(\boldsymbol{\beta})$ and $\ell(\boldsymbol{\beta})$ the likelihood and log-likelihood functions, respectively. Sometimes, these are defined as first-order conditions of Maximum Likelihood (ML) estimation method.

Adopting a similar structure, we can assume that "residuals" are the numerical estimated quantities such that orthogonality with covariates is preserved, when the first derivatives of the log-likelihood function are performed. From an operational point of view, this new conception considers residuals as quantities possessing properties shared by classical regression models in contexts where it is not possible to deduce them as difference among observed and fitted values.

This approach, starting from Pregibon (1981), leads to *generalized residuals* and it has been successfully pursued with dichotomous and ordered polytomous (probit and logit) analysis. Their introduction has been mainly suggested for checking the presence of outliers and the validation of the model (as in Franses and Paap (2001), 62;123;172-173, for instance) or for suggesting further R^2-type measures (as in Hübler, 1997).

3. Generalized residuals of CUB models

In the framework of models for ordinal data, CUB models have been proposed by Piccolo (2003), D'Elia and Piccolo (2005), Iannario and Piccolo (2009a), with increasing levels of generalizations.

Briefly, they are characterized by a discrete mixture where two parameters (related to *feeling* and *uncertainty* of the respondent, respectively) are able to generate a flexible range of distributions with different location, heterogeneity and shape. This class of models has been applied in several fields (as reported in the last reference) and immediate and interesting interpretations are derived when significant subject's covariates for feeling and uncertainty parameters are explicitly included.

Formally, given a Likert-type m-point ordered scale, for any $m > 3$, a sample of ratings $\boldsymbol{r} = (r_1, r_2, \ldots, r_n)'$ is collected on n respondents together with a set of discrete and/or continuous covariates for each subject. Thus, the sample data consist of: $(r_i,\, \boldsymbol{y}_i,\, \boldsymbol{w}_i)'$, for $i = 1, 2, \ldots, n$ where we are denoting by $\boldsymbol{y}_i$ and $\boldsymbol{w}_i$ the covariates related to uncertainty and feeling parameters, respectively.

Then, we assume that ordered responses r_i, $i = 1, 2, \ldots, n$ are the realizations of a random variable R whose distribution is defined by:

$$p_i(\boldsymbol{\theta}) = Pr\,(R = r_i \mid \boldsymbol{y}_i; \boldsymbol{\beta},\, \boldsymbol{\gamma}) = \pi_i \binom{m-1}{r-1} (1-\xi_i)^{r-1} \xi_i^{m-r} + (1-\pi_i)\,\frac{1}{m}\,,$$

and links functions are:

$$\pi_i = \pi_i(\boldsymbol{\beta}) = \frac{1}{1+e^{-\boldsymbol{y}_i\,\boldsymbol{\beta}}}\,; \quad \xi_i = \xi_i(\boldsymbol{\gamma}) = \frac{1}{1+e^{-\boldsymbol{w}_i\,\boldsymbol{\gamma}}}\,; \quad i = 1, 2, \ldots, n\,.$$

where $\boldsymbol{\beta} = (\beta_0,\, \beta_1,\, \ldots,\, \beta_p)'$ and $\boldsymbol{\gamma} = (\gamma_0,\, \gamma_1,\, \ldots,\, \gamma_p)'$. Notice that the peculiar structure of the model allows that some and/or all covariates $\boldsymbol{y}_i$ and $\boldsymbol{w}_i$ may overlap without losing identifiability.

In the quoted literature, it is common to denote such structure as CUB (p, q) models, where p and q are the number of significant covariates useful for explaining the respondents' behavior with reference to uncertainty and feeling, respectively. Then, a CUB model without covariates is denoted as CUB $(0, 0)$ and it is parameterized by $\boldsymbol{\theta} = (\pi, \xi)'$.

As a consequence, after some algebra, the log-likelihood function for the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\gamma}')'$ of a general CUB (p, q) model turns out to be:

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \left[\frac{1}{1 + e^{-\boldsymbol{y}_i \boldsymbol{\beta}}} \left\{ \binom{m-1}{r_i - 1} \frac{e^{(-\boldsymbol{w}_i \boldsymbol{\gamma})(r_i - 1)}}{(1 + e^{-\boldsymbol{w}_i \boldsymbol{\gamma}})^{m-1}} - \frac{1}{m} \right\} + \frac{1}{m} \right].$$

In this regard, an effective EM algorithm for deriving ML estimates and performing asymptotic inference is available (Piccolo, 2006).

Thus, it may be useful to introduce for such models some sort of residuals in order to achieve diagnostics information. A peculiar difficulty of CUB model stems from the circumstance that covariates (and related estimable coefficients) may differ for both uncertainty and feeling parameters. Then, there is no unique definition of residuals and we have to introduce both π-*generalized residuals* (related to uncertainty parameter) and ξ-*generalized residuals* (related to feeling parameter), respectively. This conceptual situation stems from the explicit definition of the links relating covariates to parameters. As a consequence, we will only consider CUB models with just a single set of covariates affecting either uncertainty or feeling parameters.

By means of the link functions, we may express both parameters π and ξ in a CUB $(0, 0)$ model as:

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

thus, a similar approach can be applied for defining generalized residuals in models without and with covariates, respectively.

3.1. Models without covariates

In CUB models without covariates, it is well known that information in sample data are equivalent to that contained in the vector of relative frequencies $(f_1, f_2, \ldots, f_m)'$, for any given $m > 3$. Thus, all expressions of this subsection relate to these quantities since they exhaust sample information for both parameters.

If we equate to 0 the derivative of log-likelihood function with respect to π, we get the equation:

$$\sum_{r=1}^{m} \left(\frac{f_r}{p_r(\boldsymbol{\theta})} - 1 \right) = 0 \,.$$

Then, we define the *generalized π-residuals* of a CUB model without covariates as

$$e_r^{(\pi)} = \frac{f_r}{p_r(\boldsymbol{\theta})} - 1 \,, \quad r = 1, 2, \ldots, m \,. \tag{2}$$

In a similar way, if we equate to 0 the derivative of log-likelihood function with respect to ξ, we get the equation:

$$\sum_{r=1}^{m} f_r \left(1 - \frac{1-\pi}{m\, p_r(\boldsymbol{\theta})} \right) \frac{m - r - \xi(m-1)}{\xi\,(1-\xi)} = 0 \,.$$

Then, we define the *generalized ξ-residuals* of a CUB model without covariates as

$$e_r^{(\xi)} = f_r \left(1 - \frac{1-\pi}{m\, p_r(\boldsymbol{\theta})} \right) \left[m - r - \xi(m-1) \right] , r = 1, 2, \ldots, m \,. \tag{3}$$

It is evident that *π-residuals* are immediately related to a comparison of observed and predicted probability estimated by ML methods, and thus they convey useful information for fitting measures. As a matter of fact, following and independent line of reasoning, Iannario (2009) introduced a measure of fitting for CUB model based on the average of the squared $e_r^{(\pi)}$. Finally, it is interesting to notice that in a different context (nonparametric estimation of mixtures), Lindsay and Roeder (1992) defined a *residual function* strictly equivalent to (2).

Instead, it seems difficult to derive an immediate interpretation from the cumbersome expression of the *ξ-residuals* and we will try to find conditions leading to small residuals (near 0). From (3), it is evident that such residuals tends to 0 when either $f_r \to 0$ or $p_r(\boldsymbol{\theta}) \to (1-\pi)/m$. The first condition implies that the r-th category is absent while the second condition requires that the r-th Binomial component of the mixture should tend to 0. In both case, it seems that small *ξ-residuals* are expected at categories selected by very few respondents.

3.2. Models with covariates

In a CUB $(p, 0)$ model, the p covariates $\boldsymbol{y}_i$ only affect the π parameter. Then, after some algebra, for such model, the first-order conditions on the log-likelihood function are:

$$\sum_{i=1}^{n} y_{is} \left[(1 - \pi_i(\boldsymbol{\beta})) \left(1 - \frac{1}{m\, p_i(\boldsymbol{\theta})} \right) \right] = 0\,,\ s = 0, 1, 2, \ldots, p.$$

Then, by emulating (1), we define as *generalized π-residuals* for a CUB $(p, 0)$ model the quantities:

$$e_i^{(\pi)} = (1 - \pi_i(\hat{\boldsymbol{\beta}})) \left(1 - \frac{1}{m\, p_i(\hat{\boldsymbol{\theta}})} \right)\,, \quad i = 1, 2, \ldots, n\,. \tag{4}$$

Similarly, for a CUB $(0, q)$ model, where the q covariates $\boldsymbol{w}_i$ only affect the ξ parameter, we get the equations:

$$\sum_{i=1}^{n} w_{it} \left[\left(1 - \frac{1 - \hat{\pi}}{m\, p_i(\hat{\boldsymbol{\theta}})} \right) \left[m - r_i - \xi_i(\hat{\boldsymbol{\gamma}})(m - 1) \right] \right] = 0\,,\ t = 0, 1, 2, \ldots, q.$$

Then, as before, the *generalized ξ-residuals* are defined as:

$$e_i^{(\xi)} = \left(1 - \frac{1 - \hat{\pi}}{m\, p_i(\hat{\boldsymbol{\theta}})} \right) \left[m - r_i - \xi_i(\hat{\boldsymbol{\gamma}})(m - 1) \right]\,, i = 1, 2, \ldots, n\,. \tag{5}$$

An important property of these residuals derives from the compulsory presence of a unit covariate in any CUB model. This circumstance implies that:

$$\sum_{i=1}^{n} e_i^{(\pi)} = \sum_{i=1}^{n} e_i^{(\xi)} = 0\,,$$

a result that increases their similarity with traditional residuals interpretations.

Now, by using empirical data, we will look for possible usefulness of these generalized residuals when applied to standard use, that is for outliers detection and the study of omitted/relevant covariates.

4. An artificial experiment

We examine an experimental data set of $n = 2000$ observations obtained by simulating a CUB $(1, 0)$ model with $m = 7$ and by assuming $\xi = 0.3$. The uncertainty covariate is a dummy one: $D_i = 0, 1$, and thus:

$$\pi_i = \frac{1}{1 + e^{-\beta_0 - \beta_1 D_i}}; \quad i = 0, 1.$$

We chose parameters (β_0, β_1) such that: $\pi_0 = 0.2$, when $D_i = 0$, and $\pi_1 = 0.9$, otherwise.

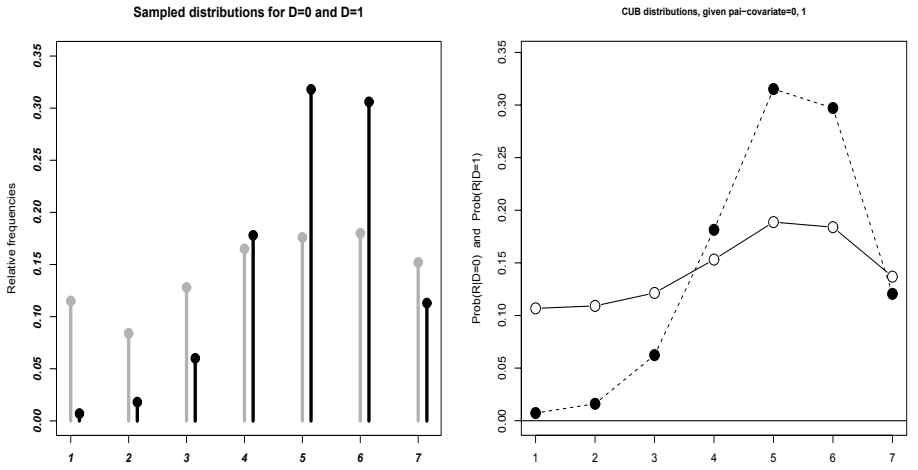

Figure 1. Observed and probability distributions for faked data

The peculiarity of this data set, fully discussed in Iannario (2009), derives from an high confounding effect that induces to suspect a unique population whereas two distinct subgroups are in fact present with regard to uncertainty behaviour. Instead, two clusters are easily detected when a CUB $(1, 0)$ model with a dummy covariate is proposed.

For this model, we compute the generalized π-residuals defined in the previous section by (4). For a check, we notice that the average and variance of these generalized π-residuals are -4.691×10^{-6} and 0.017, respectively. Then, in Figure 2 we plot the same residuals with respect to the dummy covariate in order to obtain a more stylized shape: this plot enhances that ordinal data with m categories and a dummy explanatory covariate (with 2 categories) can only generate $2 \times m$ unique residuals. We label them in correspondence with the admissible values of $r = 1, 2, \ldots, m$; thus, it is immediate to observe that, for this data set, the order of residuals is strictly related to the probability $p_i(\hat{\boldsymbol{\theta}})$ for either conditioning of D_i. As a consequence, these generalized π-residuals relate to the probabilities and not to the values of observed ratings.

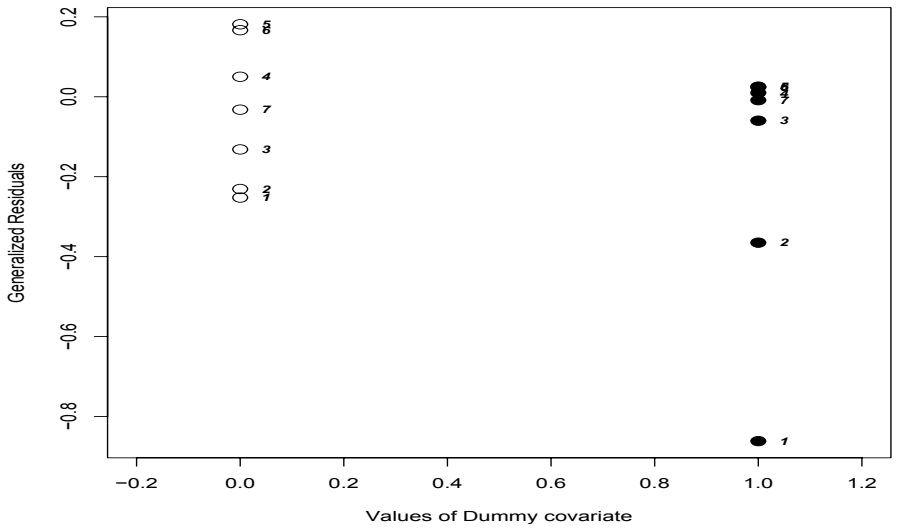

Figure 2. Generalized π-residuals with respect to dummy covariate

As a further consideration, Figure 2 shows that π-residuals pertaining to the $D_i = 0$ subgroup vary in a limited range as probabilities

are more similar each other. Finally, we found that only 7 are related to the bottom right value in Figure 2. These few extreme residuals are related to observations when $D_i = 1$ and $r_i = 1$, an event with estimated probability as low as 0.0086068; so, they appear as comparatively extreme. However, their expected occurrence in the observed sample is $0.0086068 \times 2000 = 17.214$ and we can not consider such residuals as outliers since the model predicts them with a low probability.

5. A real data set

As a further case study, we consider the expressed perceptions of subjective survival probabilities to 75 years collected by a large sample survey $(n = 20184)$ conducted in Italy by ISFOL during 2006; they have been analyzed by means of classical ordinal models by Peracchi and Perotti (2009). Instead, Iannario and Piccolo (2009b) categorized them and fitted a CUB models with covariates where $m = 7$.

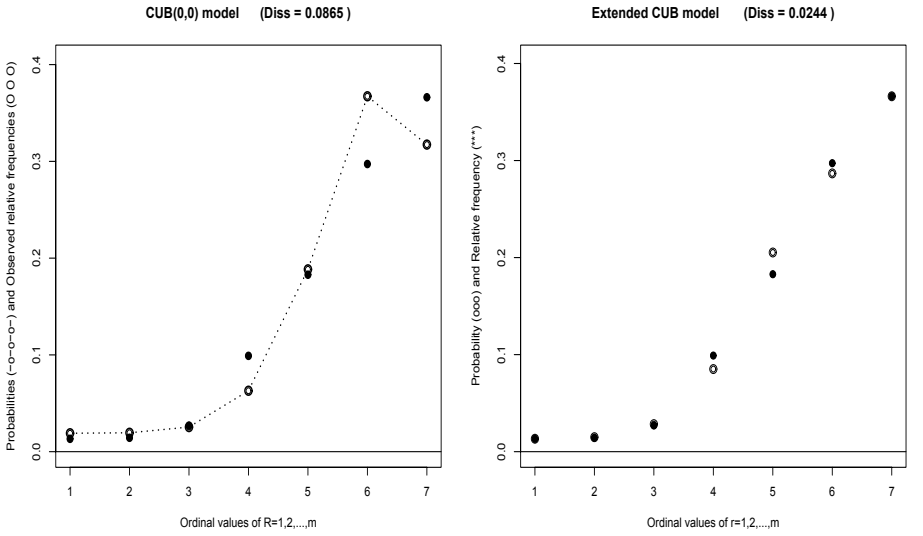

Figure 3. Generalized $\pi-$residuals after CUB *model estimation*

For reference, we show in Figure 3 (left panel) the observed and estimated distributions of responses: a *shelter effect* is evident at $R = 7$ as a consequence of an optimistic view about own survival probability to age 75, and its inclusion produces a very good fitting (right panel).

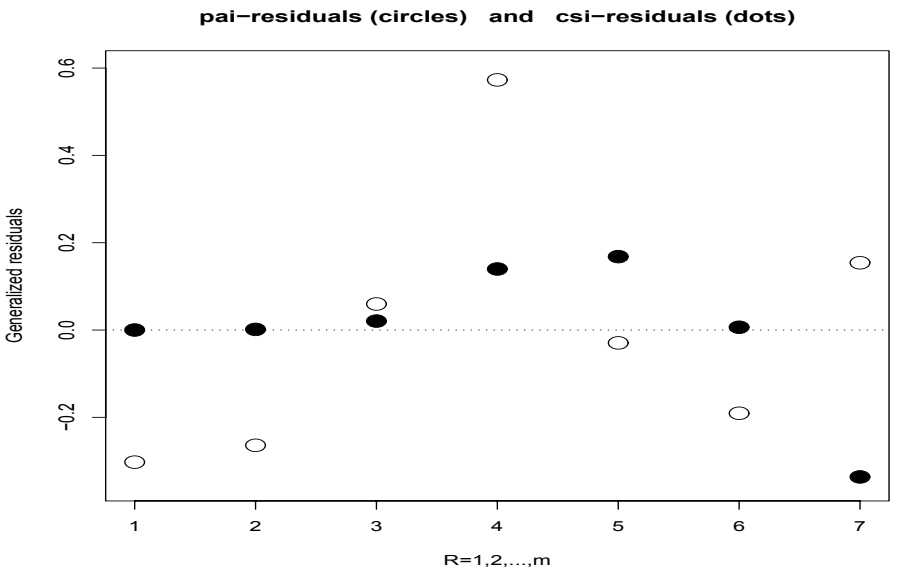

Figure 4. Generalized π- and ξ-residuals of a CUB *(0,0) model*

For this model, we get generalized π- (indicated as circle in Figure 4) and ξ-residuals for a CUB model without covariates by applying formulas (4) and (5), respectively. It is evident that the first kind of residuals expresses the relevance of the obtained fitting by the estimated probability distribution; specifically, it is observed how the weight of discrepancy at $R = 4$ is heavier than those at $R = 6$ and $R = 7$. This confirms that ML estimates does not performs *per se* a fitting objective since discrepancy is

weighted with the inverse of probabilities. On the contrary, the pattern of ξ-residuals does not manifest a clear relationship between observed and fitted distributions since residuals at $R = 6$ is about 0 in presence of a serious discrepancy whereas it is the largest at $R = 7$ (and this is not the maximum deviation).

Then, for the same data set, we consider CUB $(p, 0)$ and CUB $(0, q)$ models, respectively, where the age of respondent has been considered as a relevant covariate after preliminarily logging and transforming for improving the statistical properties of models (this result confirmed a cohort effect).

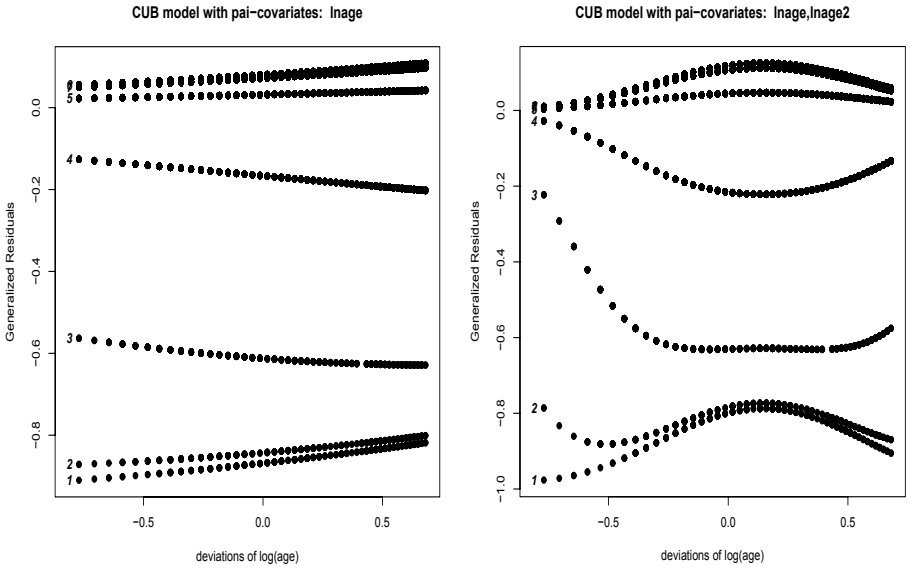

Figure 5. Generalized π-residuals after CUB *model estimation*

In Figure 5 we plot the generalized π-residuals from a CUB model where age and age squared (on left and right panels, respectively) have been used as explanatory for the uncertainty parameter. In all these plots

we denote with $r = 1, 2, \ldots, 7$ the residuals generated from the computation of $Pr\,(R = r)$ given the covariate.

As a first comment we see that age acts in a quite homogeneous manner in determining residuals (and then responses) producing more extreme residuals in correspondence to the rarest events. If we add the squared covariates we see that the effect is quite modest on higher probabilities where substantial modifications are induced on low and intermediate probabilities. Since this parameter is related to uncertainty of responses, we may infer that significance of age (as explained by the reversion effect caused by squaring it) is more sensible on low and medium probability but it is not so important for expressing a high subjective probability to survive.

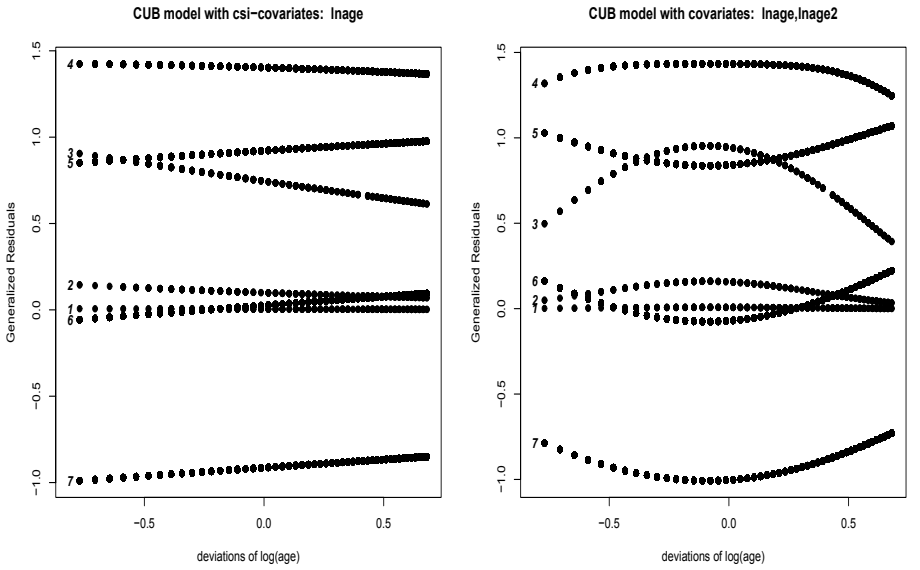

Figure 6. Generalized ξ-residuals after CUB *model estimation*

Similarly, in Figure 6, we consider the generalized ξ-residuals after the introduction of age and age squared (on left and right panels, respectively). In this case, the pattern seems different as the comment. First of

all, as we can see in the left panel, the response at $R = 1, 2, 6$ is not affected by age in a significant manner, whereas $R = 4, 7$ behavior is more extreme as far as perception is concerned. An important consideration arises from analyzing how probabilities at $R = 3, 5, 6, 7$ react in a sensible manner to a squared age covariate (see the right panel); for instance, this added covariate reverses the relative position of $R = 3$ and $R = 5$ probabilities. Finally, the probabilities at $R = 3$ and $R = 7$ manifest in a larger measure the impact of inversion caused by squared covariate, that is a similar behavior of young and elderly respondents.

A general consideration applies: the generalized ξ-residuals may help to investigate where and how significant covariates affect the single probabilities and not only the whole distribution, as instead it is enhanced by estimating and validating the model.

6. Concluding remarks

In this paper, we have introduced generalized residuals for a new class of models proposed for the interpretation and fitting of ordinal data even with covariates when they are significant. The formal definition of such residuals has been derived and two empirical examples have been discussed and commented for evaluating limits and implications of such proposals.

This preliminary study shows that standard concepts for detecting outliers and leverage effects should be considered as useless with ordinal data; in fact, extreme values are mostly derived by low probabilities to appear in a sample, and thus deserve homogeneous consideration. Moreover, the effect and the role of covariates on the parameters have to be judiciously considered by stepwise strategies and induced modifications into the generalized residuals. Indeed, it seems that residuals defined by taking uncertainty parameters into consideration are more related to fitting interpretation. On the other hand, residuals defined by taking feeling as a reference do not show a clear pattern in any case, but they help to understate how a covariate induces a differential modification in a single probability or in a subgroup of probabilities.

Acknowledgements: Authors thank referees for constructive comments to a preliminary version of the paper. The research has been partly supported by PRIN-2008: "Modelli per variabili latenti basati su dati ordinali: metodi statistici ed evidenze empiriche'. ISFOL survey data has been used under the agreement ISFOL/PLUS 2006/430. Research structures of CFEPSR, Portici are gratefully acknowledged.

References

Atkinson A.C. (1985), *Plots, transformations, and regressions: an introduction to graphical methods of diagnostic regression analysis*, Clarendon Press, Oxford.

Belsey D.A., Kuh E., Welsch R.E. (1980), *Regression diagnostics: identifying influential data and sources of collinearity*, J. Wiley & Sons, New York.

Cook R.D., Weisberg S. (1994), *An introduction to regression graphics*, J. Wiley & Sons, New York.

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

Fox J. (1991), *Regression diagnostics: an introduction*, Sage, Newbury Park, CA.

Fox J. (1997), *Applied regression analysis, linear models, and related methods*, Sage Publications, Thousand Oaks, CA.

Franses P.H., Paap R. (2001), *Quantitative models in marketing research*, Cambridge University Press, Cambridge.

Hübler O. (1977), Pseudo latent models: goodness of fit measures, residuals, estimation, testing, and simulation, *Statistical Papers*, 38, 271–285.

Iannario M. (2009), Fitting measures for ordinal data models, *Quaderni di Statistica*, 11, 39–72.

Iannario M., Piccolo D. (2009a), A program in R for CUB models inference, Version 2.0, available at *http://www.dipstat.unina.it*

Iannario M., Piccolo D. (2009b), Statistical modelling of subjective survival probabilities, submitted.

Lindsay B.G., Roeder K. (1992), Residual diagnostics for mixture models, *Journal of the American Statistical Association*, 87, 785–792.

Magee L. (1990), R^2 measures based on Wald and likelihood ratio joint significance tests, *The American Statistician*, 44, 250–253.

McCullagh P., Nelder J.A. (1989), *Generalized linear models*, 2^{nd} edition. Chapman and Hall, London.

Mosteller F., Tukey J.W. (1968), *Data analysis and regression*, Addison-Wesley, Reading, MA.

Peracchi F., Perotti V. (2009), Subjective survival probabilities and life tables: an empirical analysis of cohort effects, *Genus*, LXV, 23–57.

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

Piccolo D. (2006), Observed information matrix for MUB models, *Quaderni di Statistica*, 8, 33–78.

Pregibon D. (1981), Logistic regression diagnostics, *The Annals of Statistics*, 9, 705–724.

Seber G.A.F. (1977), *Linear regression analysis*, J. Wiley & Sons, New York.