STATISTICA, anno LXXII, n. 2, 2012

RESIDUAL DIAGNOSTICS FOR INTERPRETING CUB MODELS[1]

F. Di Iorio, M. Iannario

1. INTRODUCTION

In ordinary linear regression, graphical diagnostics can be very useful for detecting anomalous features in a fitted model (Atkinson, 1985; Belsey *et al.*, 1980; Cook and Weisberg, 1994; Daniel and Wood, 1971; Fox, 1991, 1997; Weisberg, 1980). This validation step is pursued by residual analysis and it is worth considering for several purposes: to check the structure of the model, verifying the nature and persistence of the postulated dependence, to detect outliers and/or influential data, to assess the validity of classical linear hypotheses (homoscedastic and uncorrelated errors), to test distributional assumptions (Gaussianity, for instance) or shape behaviours (skewness, heavy tails), and so on.

However, when the response variable is not continuous (and specifically of categorical nature), it is difficult to interpret the usual definition of residuals and the related graphical devices. In fact, the results are invariably discrete and standard techniques are not very effective. The problem has been mainly raised in the framework of Generalized Linear Models (GLM), as proposed by McCullagh (1980) and McCullagh and Nelder (1989). In this context, the definition of *generalized residuals* have been derived by first-order conditions of the maximum likelihood equations (Lanweher, Pregibon and Shoemaker, 1984).

The specification and extensions of residual diagnostics to categorical and ordinal data have been successfully pursued with ordered polytomous (probit and logit) analysis, as in Pregibon (1981). In this regard, Agresti (2010) synthesizes several approaches, and we mention the *standardized residuals* given by the ratio of the difference between the observed and the fitted values and the standard error of difference under the hypothesis that the model is correct. In addition, it can be informative to quote the residuals obtained by cumulative totals and referred to as *Pearson-type residuals*. In relation to this frame, Liu *et al.* (2009) proposed graphical diagnostics based on cumulative sums of residuals to check the misspecification of the proportional odd models. Moreover, Pruscha (1994) suggested partial

[1] This work has been partly supported by a MIUR grant for PRIN2008 project of the Research Unit of University of Naples Federico II (PUC E61J10000020001) and FARO project sponsored by Polo SUS. ISFOL survey data has been used under the agreement ISFOL/PLUS 2006/430.

residuals whereas Bender and Benner (2002) introduced a smoothed partial residual plot.

Other issues related to the analysis of residuals from a fitted model, have been obtained by using univariate and bivariate marginal distributions: Bartholomew and Tzamourani (1999) and Jöreskog and Moustaki (2001), among others. Finally, Lindsay and Roeder (1992) introduced residual diagnostics for mixture models.

In this paper, we analyze residual diagnostics with reference to ordinal data modelled by means of CUB models (Piccolo, 2003; D'Elia and Piccolo, 2005; Iannario, 2012; Iannario and Piccolo, 2012). More specifically, we define estimated residuals, their transformations and related graphical methods and then extend the concept of *binned residuals* (Gelman and Hill, 2007) to CUB models. Residual plots, based on first-order conditions of maximum likelihood equations, have been previously discussed by Di Iorio and Piccolo (2009).

The paper is organized as follows: the formulation and some basic features of CUB models are considered in section 2. Section 3 proposes a definition of residuals for such mixture models whereas section 4 introduces binned residual plots. Section 5 checks the usefulness of such device by means of empirical data set. Section 6 contains some final remarks.

2. BACKGROUND AND NOTATION

We analyze ordinal responses expressed on the support $\{1,2,...,m\}$ where $m>3$ for identifiability purposes (Iannario, 2010). To be specific, we consider *ratings* data $\boldsymbol{r}=(r_1, r_2,..., r_n)'$ as the collection of scores assigned by a sample of n raters to a prefixed item.

Formally, CUB models[2] are specified by considering the observed sample $\boldsymbol{r}$ of responses as n realizations of a discrete random variable R whose probability distribution is a mixture of Uniform and shifted Binomial random variables, both defined on the support $\{1,2,...,m\}$.

For improving fitting and interpretation, a logistic link among the parameters and some selected subjects' covariates is conveniently assumed. The parameters of a CUB model (denoted by π and ξ, respectively) are inversely related to *uncertainty* and *feeling* features of respondents, respectively. Instead, the observations of subjects' covariates are collected for the two components in the matrices $\boldsymbol{Y}$ and $\boldsymbol{W}$, respectively. Further details have been discussed by Iannario and Piccolo (2012).

Then, information for explaining the rating r_i of the i-th subject, for $i=1,2,...,n$, is:

$$(r_i \mid 1, y_{i1}, y_{i2}, \ldots, y_{ip} \mid 1, w_{i1}, w_{i2}, \ldots, w_{iq}).$$

[2] The acronym CUB derives from the *C*ombination of *U*niform and (shifted) *B*inomial random variables in the mixture which defines the model.

It is related to parameters by the *systematic* links:

$$\pi_i = \pi_i(\beta) = \frac{1}{1+e^{-y_i\beta}}; \qquad \xi_i = \xi_i(\gamma) = \frac{1}{1+e^{-w_i\gamma}}; \qquad i = 1,2,\ldots,n, \tag{1}$$

where $\boldsymbol{y}_i$ and $\boldsymbol{w}_i$ are the covariates of the i-th subject for explaining π_i and ξ_i, respectively, and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$; $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \ldots, \gamma_q)'$, are the vectors of the parameter vectors (supposed fixed across individuals) for $\boldsymbol{y}_i$ and $\boldsymbol{w}_i$, respectively.

As a consequence, for a given $m>3$, the probability distribution of a CUB model with p covariates for explaining uncertainty and q covariates for explaining feeling, hereafter denoted as CUB(p,q) model, turns out to be:

$$Pr(\mathrm{R} = r_i \mid y_i, w_i, \beta, \gamma) = \frac{1}{1+e^{-y_i\beta}}\left[\binom{m-1}{r_i-1}\frac{(e^{-w_i\gamma})^{r_i-1}}{(1+e^{-w_i\gamma})^{m-1}} - \frac{1}{m}\right] + \frac{1}{m}. \tag{2}$$

Given the sample information $I_n = (\boldsymbol{r} \mid \boldsymbol{Y} \mid \boldsymbol{W})$, if we let $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\gamma}')'$ the set of all $p+q+2$ parameters to be estimated, then the log-likelihood function is defined by:

$$\log L(\theta \mid I_n) = \sum_{i=1}^{n} \log[Pr(\mathrm{R} = r_i \mid y_i, w_i, \theta)]. \tag{3}$$

Notice that, for obtaining maximum likelihood (ML) estimates, an effective maximization approach to (3) is required and this is based on EM algorithm (Piccolo, 2006).

In order to assess the closeness between the observed responses and the fitted values, goodness-of-fit statistics have been introduced in the literature. Some of them are based on deviance and divergence measures (Cameron and Windmeijer, 1997; Hastie, 1987; McCullagh and Nelder, 1989) and these criteria are particularly useful in presence of covariates for comparing nested models.

When covariates are absent, the absolute differences between the probability $Pr(\mathrm{R} = r \mid \hat{\theta}) = p_r(\hat{\theta})$, evaluated by the estimated CUB model, and the corresponding observed relative frequency f_r, $r=1,2,\ldots,m$ are often used as inverse measures of goodness of fit. Such quantities may be considered as starting points for building dissimilarity indexes (Leti, 1979; Simonoff, 2003) or direct normalized fitting measures as:

$$F^2 = 1 - \frac{1}{2}\sum_{r=1}^{m} |f_r - p_r(\hat{\theta})|.$$

An alternative normalized measure:

$$L^2 = \left[1 + \frac{1}{m}\sum_{r=1}^{m}\left(\frac{f_r}{p_r(\hat{\theta})} - 1\right)^2\right]^{-1}$$

has been proposed by Iannario (2009). This index has been derived by a first-order property of ML estimates.

As already mentioned, inferential issues for this class of models have been pursued by ML methods, and the related asymptotic inference may be applied by using variance and covariance matrix of estimators (with ML estimates plugged in). Alternatively, nonparametric inference has been introduced in order to check the adequacy of an estimated model or to compare nested CUB models obtained for small sample size (Arboretti *et al.* 2011; Bonnini *et al.* 2011).

3. GENERALIZED RESIDUALS IN CUB MODELS

In a previous work, Di Iorio and Piccolo (2009) discussed generalized residuals for CUB models for uncertainty and feeling covariates. They defined, for $i=1,2,...,n$, the quantities:

$$e_i^{(\pi)} = (1-\pi_i(\hat{\beta}))\left(1-\frac{1}{m\, p_i(\hat{\theta})}\right) \tag{4}$$

$$e_i^{(\xi)} = \left(1-\frac{1-\hat{\pi}}{m\, p_i(\hat{\theta})}\right)(m-r_i-(m-1)\,\xi_i(\hat{\gamma})) \tag{5}$$

which may be considered as ML generalized residuals with respect to π_i and ξ_i, respectively. They are obtained from first-order conditions of ML equations and may be useful to investigate where and how significant covariates affect single probabilities. Notice that these quantities assume a number of different values depending on the number of modalities of covariates.

Hereafter, we introduce a different definition of residuals for CUB models which is more similar to the standard regression framework. For simplicity, we consider the common case where only the feeling parameter ξ is explained by covariates $\boldsymbol{W}$. Thus, the discussion is limited to CUB*(0,q)* models.

We define residuals as: $e_i = R_i - E(R_i \mid W=w_i;\, \boldsymbol{\theta})$ and their estimates as:

$$\hat{e}_i = r_i - E(R_i \mid W = w_i;\theta), \qquad i = 1,2,...,n.$$

Since the expectation of a CUB model is: $E(R) = \pi(m-1)\left(\frac{1}{2}-\xi\right)+\frac{(m+1)}{2}$, given that covariates affect only the feeling parameter ξ, from (1), the previous residuals may be expressed as:

$$\hat{e}_i = r_i - \left[\hat{\pi}(m-1)\left(\frac{1}{2}-\xi_i(\hat{\gamma})\right)+\frac{(m+1)}{2}\right], \qquad i = 1,2,...,n. \tag{6}$$

If we compare last expression with (4) and (5), the difference between the two approaches in defining residuals should be evident. In fact, in (6) the evaluation

of $\hat{e}_i$ is yet conditioned to $\hat{\pi}$ but the explicit reference to $p_i(\hat{\theta})$ in denominators – as it happens in definitions (4) and (5) – disappeared. Thus, small probabilities are not so influential on the residual computations.

Notice that:

$$E(e_i) = E(R_i - E(R_i \mid W = w_i; \theta)) = E(R_i) - E(R_i \mid W = w_i; \theta) = 0;$$

$$V(e_i) = Var(R_i) = (m-1)\left[\pi \frac{e^{-A_i}}{(1+e^{-A_i})^2} + \frac{1-\pi}{4}\left(\frac{m+1}{3} + \pi(m-1)\left[\tanh\left(-\frac{A_i}{2}\right)^2\right]\right)\right];$$

where A_i=-$\boldsymbol{w}_i$ γ. Thus, the residuals (6) are heteroscedastic.

4. BINNED RESIDUAL PLOTS

In order to avoid the discrete pattern in the residual plot, which is the consequence of the ordinal nature of responses, we introduce a binned residual plot for CUB models according to similar proposals for dichotomous data.

Suppose that X is the variable we use to define J bins for plotting averaged residuals belonging to the selected j-th bin, for $j=1,2,...,J$. Such variable may be a subjects' covariate (or a combination of them) but also a probability computed by the estimated CUB models.

To be consistent, we define the i-th residual belonging to the j-th bin as:

$$e_{i[j]}, \quad i = 1,2,...,n_j; \quad j = 1,2,...,J.$$

Typically, there is some arbitrariness in choosing the number of bins: we want enough points so that the averaged residuals are not so noisy, but it helps to have many bins to see more local patterns in the residuals. In this regard, it is a common practice to specify n_i as a constant value ($\sqrt{n}$, say) but some bins may posses different size; so, our notation is taking this generalization into account.

Another approach would be to apply a nonparametric smoothing procedure such as *lowess* (locally weighted scatterplot smoothing: Cleveland, 1979). It combines the simplicity of linear least squares regression with the flexibity of nonlinear regression by fitting simple models to local subsets of data to build a function that describes the deterministic variation in the data.

The graphical device we are discussing about is a representation of the average of n_j observed realizations of $e_{i[j]}$ for each group (bin) selected according to the ordered values of a prefixed variable X. Then, the binned residuals are given by:

$$\overline{e}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} e_{i[j]}, \quad j = 1,2,...,J.$$

This random variable is characterized by:

$$E(\overline{e}_j) = 0; \qquad Var(\overline{e}_j) = \frac{1}{n_j}\sum_{i=1}^{n_j} \sigma^2{}_{i[j]}, \quad j = 1,2,...,J,$$

where we denoted by $\sigma^2{}_{i[j]}$ the variance of a CUB random variable (Piccolo, 2003) conditioned by the observed value of the X variable for the i–th subject.

Thus, to obtain a *binned residual plot* we order data with respect to the X variable into categories (bins), and then we plot the average residual versus the average value of the variable for each bin. Specifically, each point of the *binned residual plot* is specified by an abscissa computed on some average of X and an ordinate computed on the realizations of $\overline{e}_j$.

The asymptotic standard-error bounds, within which one would expect about 95% of the binned residuals to fall if the model is assumed to be true, are:

$$\pm 1.96 \frac{\sigma^2{}_{i[j]}}{\sqrt{n_j}}, \qquad j = 1,2,...,J.$$

5. SOME EMPIRICAL EVIDENCE

To verify the effectiveness of the proposal we implement the analysis of binned residual plots for CUB models on a real data set. Specifically, we analyze a subset of ISFOL-2006 data by considering 20184 subjects which expressed their perception of subjective survival probabilities to 75 years; such data have been transformed in a qualitative assessment about the personal perception to survive by means of a Likert scale with $m=7$. For illustrative purposes, we refer to models fitted to such data and already discussed in Di Iorio and Piccolo (2009), Iannario and Piccolo (2010).

In the estimated CUB models, only the *Gender*, the deviations and the squared deviations of logarithm of *Age* turned out to be relevant covariates for explaining the feeling parameter, and the main results are reported in Table 1. As we can see, the inclusion of covariates for feeling parameter ξ does not sensibly modify the estimation of $\hat{\pi}$, whereas significantly improves the goodness of fit measures (as shown by the increase of estimated log-likelihood functions and the corresponding reductions of BIC).

If we apply the proposed analysis to such models, we obtain different binned residual plots. We report in Figure 1 (left panel) the average residuals $\overline{e}_j$ for a CUB*(0,1)* model with deviations of logarithm of *Age* and (in the right panel) the average residuals of a CUB(0,2) model which includes both deviations and squared deviations of logarithm of *Age*. Each bin contains $\sqrt{n} = 142$ residuals, in both plots.

TABLE 1

CUB models for subjective survival probabilities to age 75

Models	Parameter Estimates		Log-likel.	BIC
CUB(0,0)	$\hat{\pi} = 0.867(0.005)$	$\hat{\xi} = 0.163(0.001)$	-30383	60782.3
CUB(0,1)	$\hat{\pi} = 0.865(0.005)$	$\hat{\gamma}_0 = -1.642(0.011)$	-30365	60754.5
Log(Age)		$\hat{\gamma}_1 = -0.135(0.023)$		
CUB(0,2)	$\hat{\pi} = 0.867(0.005)$	$\hat{\gamma}_0 = -1.522(0.015)$	-30310	60652.7
Log(Age)		$\hat{\gamma}_1 = -0.125(0.024)$		
[Log(Age)]²		$\hat{\gamma}_2 = -0.682(0.065)$		
CUB(0,3)	$\hat{\pi} = 0.868(0.005)$	$\hat{\gamma}_0 = -1.598(0.020)$	-30291	60622.8
Log(Age)		$\hat{\gamma}_1 = -0.108(0.024)$		
[Log(Age)]²		$\hat{\gamma}_2 = -0.616(0.066)$		
Gender		$\hat{\gamma}_3 = 0.121(0.020)$		

In the left panel of Figure 1, we observe a quadratic (parabolic) structure, which disappears in the right panel after the introduction of deviations of squared logarithm of age. Specifically, we observe that about 10% of the binned residuals corresponding to small values of the deviance of log(*Age*) exceeds the upper level of confidence bound.

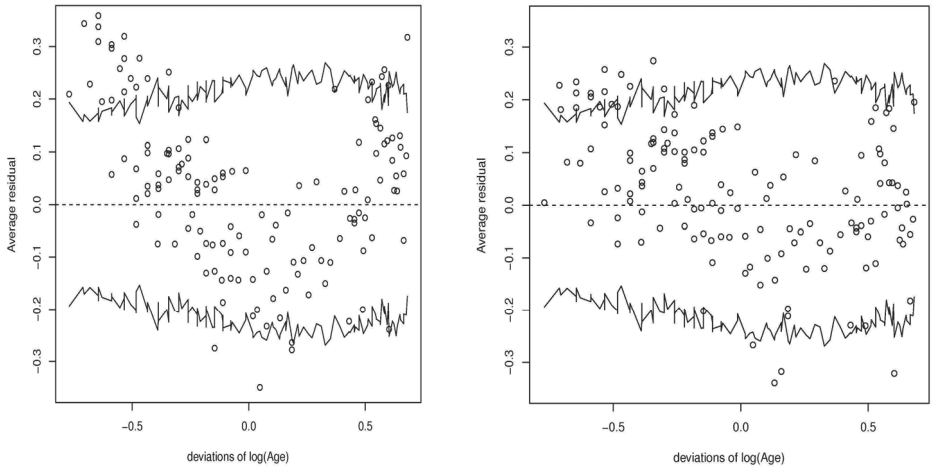

Figure 1 – Binned residuals for ξ parameter with covariates *Age* (left panel) and *Age*² (right panel).

A further inspection of these data shows that the pattern definitely improves after the introduction of *Gender* (left panel, Figure 2), as confirmed also by the results of the last model of Table 1. The points in the plot do not show further patterns and the number of points out of the confidence limits are dramatically reduced.

In addition, in the right panel of Figure 2, we report another kind of residual plot in which we observe the relationship between average residuals and expected values of the rating scores. In the left panel, a moderate amount of residuals falls outside of the dotted 95% confidence bands for the residual plot, whereas in the right panel just few binned residuals are located out of the confidence bands. Thus, such representations summarize and support the quality of the estimation.

Finally, if we compare the graphical devices for the same data set reported by Di Iorio and Piccolo (2009), and based on (4) and (5) definitions of residuals, it seems evident that Figures 1-2 depict in a sharper manner the information obtained by the model diagnostics step.

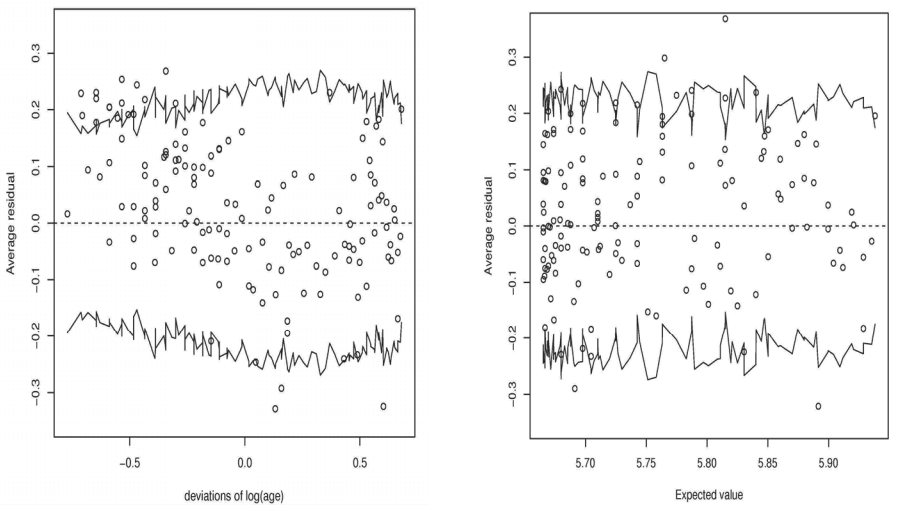

Figure 2 – Binned residuals for ξ parameter vs average of *Age* and expected values.

6. CONCLUDING REMARKS

In this work, we have discussed the role of residuals obtained by fitting CUB models to ordinal data as diagnostic tools for improving the interpretation and checking the effect of covariates. The binned residual plot seems a convenient trick to avoid the discreteness nature of such data which hides useful relationships and prevent from the detection of influential behaviors.

In this regard, further studies are necessary in order to explore related issues, as the effect of different size of bins since the arbitrariness in choosing the number of bins (see Gelman and Hill, 2007, pag. 97) may affect some conclusions. Investigations of several data sets are important but also simulation studies should be performed in order to test the real effectiveness of such new diagnostic tools.

Department of Theory and Methods of Human and Social Sciences, Statistical Sciences Section
University of Naples Federico II

FRANCESCA DI IORIO
MARIA IANNARIO

REFERENCES

A. AGRESTI, (2010), *Analysis of ordinal categorical data*, 2nd edition, J. Wiley & Sons, New Jersey.

R. ARBORETTI, S. BONNINI, M. IANNARIO, F. SOLMI, (2011), *Permutation test approach for the analysis of rating data*, Proceeding of SIS Meeting, Bologna.

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

D. J. BARTHOLOMEW, P. TZAMOURANI, (1999), *The goodness of fit of latent trait models in attitude measurement*, "Sociological Methods and Research", 27:525-546.

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

R. BENDER, A. BENNER, (2000), *Calculating ordinal regression models in SAS and S-Plus,* "Biometrical Journal", 42:677-699.

S. BONNINI, D. PICCOLO, L. SALMASO, F. SOLMI, (2011), *Permutation inference for a class of mixture models*, "Communications in Statistics. Theory and Methods", forthcoming.

A. C. CAMERON, F. A. G. WINDMEIJER, (1997), *An R-squared measure of goodness of fit for some common nonlinear regression models*, "Journal of Econometrics", 77:329-342.

W. S. CLEVELAND, (1979), *Robust locally weighted regression and smoothing scatterplots*, "Journal of the American Statistical Association", 74:829-836.

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

C. DANIEL, F. S. WOOD, (1971), *Fitting equation to data*, J. Wiley & Sons, New York.

A. D'ELIA, D. PICCOLO, (2005), *A mixture model for preference data analysis*, "Computational Statistics & Data Analysis", 49:917-934.

F. DI IORIO, D. PICCOLO, (2009), *Generalized residuals in CUB models: definition and applications*, "Quaderni di Statistica", 11:73-88.

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

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

A. GELMAN, J. HILL, (2007), *Data Analysis using Regression and Multilevel/Hierarchical Models*, Cambridge University Press, New York.

T. HASTIE, (1987), *A closer look at deviance*, "The American Statistician", 41:16-20.

M. IANNARIO, (2009), *Fitting measures for ordinal data models*, "Quaderni di Statistica", 11:39-72.

M. IANNARIO, (2010), *On the identifiability of a mixture model for ordinal data*, "METRON", LXVIII: 87-94.

M. IANNARIO, (2012), *Modelling shelter choices in a class of mixture models for ordinal responses*, "Statistical Methods and Applications", 21, 1-22.

M. IANNARIO, D. PICCOLO, (2010), *Statistical modelling of subjective survival probabilities*, "GENUS", LXVI:17-42.

M. IANNARIO, D. PICCOLO, (2012), *CUB models: Statistical methods and empirical evidence*, in: R.S. KENETT, S. SALINI (eds.) "Modern Analysis of Customer Surveys: with application using R", J. Wiley & Sons, New York, 231-258.

K. JÖRESKOG, I. MOUSTAKI, (2001), *Factor analysis of ordinal variables: A comparison of three approaches*, "Multivariate Behavioral Research", 36:347-387.

J. M. LANWEHER, D. PREGIBON, A. C. SHOEMAKER, (1984), *Graphical methods for assessing logistic regression models,* "Journal of the American Statistical Association", 79:61-83.

G. LETI, (1979), *Distanze e indici statistici*, La Goliardica, Roma.

B. G. LINDSAY, K. ROEDER, (1992), *Residual diagnostics for mixture models*, "Journal of the American Statistical Association", 87:785-792.
I. LIU, B. MUKHERJEE, T. SUESSE, D. SPARROW, S. K. PARK, (2009), *Graphical diagnostics to check model misspecification for the proportional odds regression model.* "Statistics in Medicine", 28:412-429.
P. MCCULLAGH, (1980), *Regression models for ordinal data (with discussion)*, "Journal of the Royal Statistical Society, Series B", 42:109-142.
P. MCCULLAGH, J. A. NELDER, (1998), *Generalized linear models*, 2nd edition, Chapman & Hall, London.
D. PICCOLO, (2003), *On the moments of a mixture of uniform and shifted binomial random variables*, "Quaderni di Statistica", 5:85-104.
D. PICCOLO, (2006), *Observed information matrix for MUB models*, "Quaderni di Statistica", 8:33-78.
D. PREGIBON, (1981), *Logistic regression diagnostics*, "The Annals of Statistics", 9:705-724.
H. PURSCHA, (1994), *Partial residuals in cumulative regression models for ordinal data*, "Statistical Papers", 35:273-284.
S. WEISBERG, (1980), *Applied linear regression*, J. Wiley & Sons, New York.
J. S. SIMONOFF, (2003), *Analyzing categorical data*, Springer, New York.

SUMMARY

Residual diagnostics for interpreting CUB models

CUB models represent a new approach for the analysis of categorical ordinal data. The relevant domain of study is the specification and estimation of the behaviour of respondents when faced to ratings by analysing the relationship among ordinal scores and observed covariates. The increasing use of such models suggests to delve into the issue of appropriate residuals to be used for diagnostic purposes. In fact, the *discreteness* of the response variable discourages the use of standard regression paradigms. In this context, we propose the extension and implementation of a specific graphical methodology, known as binned residual plots, in order to check the adequacy of fitted CUB models and/or infer about improvements of the maintained model. Such proposals have been exemplified through the analysis of real data.