RESIDUALS DIAGNOSTICS FOR MODEL-BASED TREES FOR ORDERED RATING RESPONSES

Rosaria Simone [1]

[1] Department of Political Sciences, University of Naples Federico II, (e-mail: rosaria.simone@unina.it)

ABSTRACT: The contribution illustrates how selection of model-based trees can be supplemented by local diagnostics on a necessary condition for the correct specification of the baseline model, based on surrogate residuals' analysis. The procedure can support the choice of the baseline model or the tuning of pre-pruning conditions. Examples are given for MOB trees based on ordinal logit models.

KEYWORDS: Model-based tree, ordered data, surrogate residuals

1 Motivating framework

The contribution discusses the advantages of performing residuals diagnostics for ordinal data models (Liu & Zhang, 2018) in the setting of model-based classification trees. Specifically, a necessary condition for a model to be correctly specified is that surrogate residuals are uniformly distributed. The paper shows the procedure for model-based trees (Zeileis *et al.*, 2008) with ordinal logit models to tune pre-pruning conditions, to identify the nodes that should be preferably pruned, or to select the best tree in terms of the maintained local model. For illustration, we consider data from the 5th European Working Condition Survey carried out in 2010 and focus on $N = 972$ responses for Italy to the question 'Do you experience stress in your work?' on a $m = 5$ wording-type scale: 'Always', 'Most of the time', 'Sometimes', 'Rarely', 'Never' [*].[†]

[*] Coded from 1 to 5 for convenience

[†] To avoid bias in favour of variables with many splits, we consider as covariates dichotomous factors *Gender* (G) experience of *Insomnia* (I), experience of *Fatigue* (F), experience of *Depression* (D), presence of *Risk* (R) connected to the job stability, being the Household Breadwinner (B). The only non dichotomous covariate is the size of the *Household* (H) as number of components.

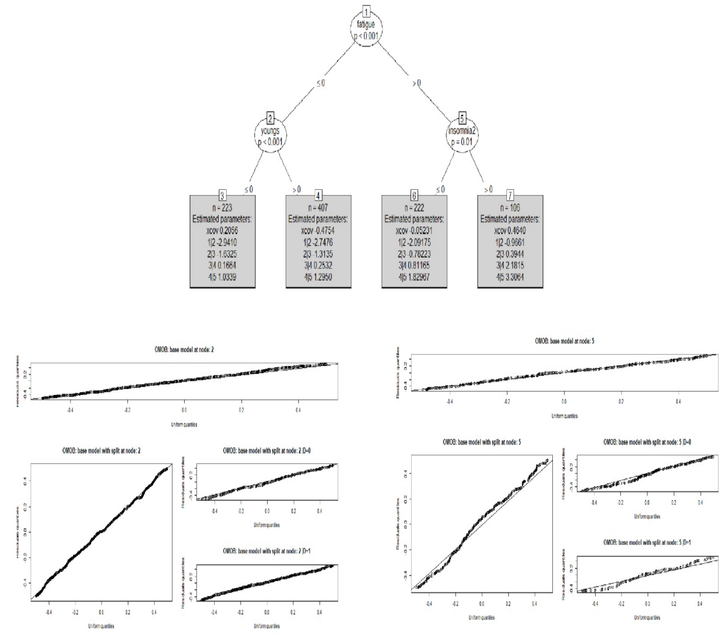

Figure 1. *MOB for $M : Stress \sim Gender$ (Top); Residuals' diagnostics for MOB based on M on perceived work-related stress (bottom)*

2 Residuals diagnostics of MOB trees for ordered responses

In the setting of MOB trees [‡], consider an ordered logit model $M: \; logit(Pr(R_i \leq j|x_i)) = \alpha_j - \beta_1 x_i, \; j = 1,\ldots,m$ as local maintained model. For instance, let $M : Stress \sim Gender$ (see the top panel in Figure 1). Then, Figure 1 (bottom) displays the uniform QQ plot of residuals at inner nodes and descendants, showing that the split at node 5 should be preferably pruned as M does not meet locally the necessary condition for being correctly specified.

Then, consider model $M : Stress \sim Breadwinner$ and the corresponding MOB with `minsplit=50`, `maxdepth=4` (see Figure 2 - left). Uniform QQ plots of residuals' at tree nodes are displayed in Figure 3, showing that - except for node 3 and its descendants - there is poor evidence for M being correctly specified locally. Then, modifying the pre-pruning condition on

[‡] (`partykit` R package)

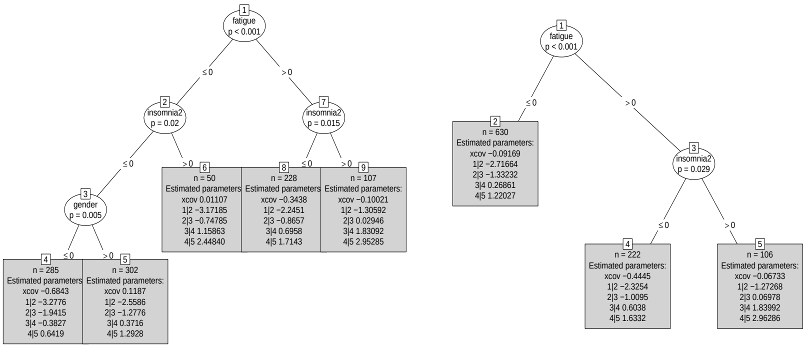

Figure 2. *MOB tree for M : Stress ∼ Breadwinner with* `minsplit = 50` *(left) and* `minsplit=100` *(right)*

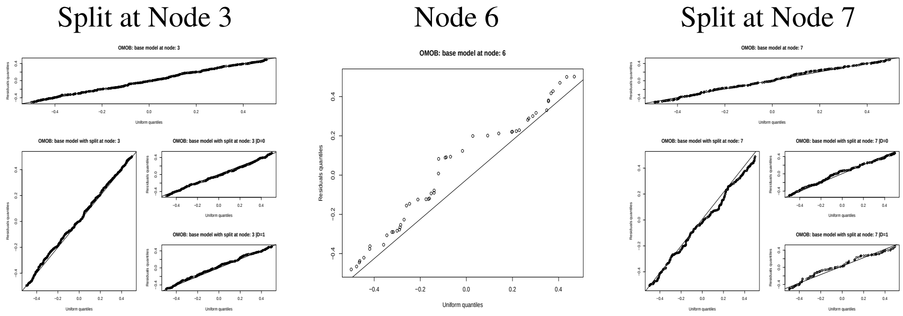

Figure 3. *QQ plot of surrogate residuals for a MOB tree based on M : Stress ∼ Breadwinner (*`minsplit=50`*)*

minimum sample size required to attempt a split (`minsplit = 100`) yields a reduced MOB tree (see Figure 2 - right), with evidence that the necessary condition for being correctly specified is overall satisfied (see Figure 4).

3 Concluding remarks

Residuals diagnostics in the setting of model-based trees can be successfully exploited also for trees based on CUB models (Cappelli *et al.*, 2019) to select the baseline model or the best performing partitioning criterion. The proposed procedure can be further integrated within model selection in order to focus only on models for which the necessary condition for being correctly specified

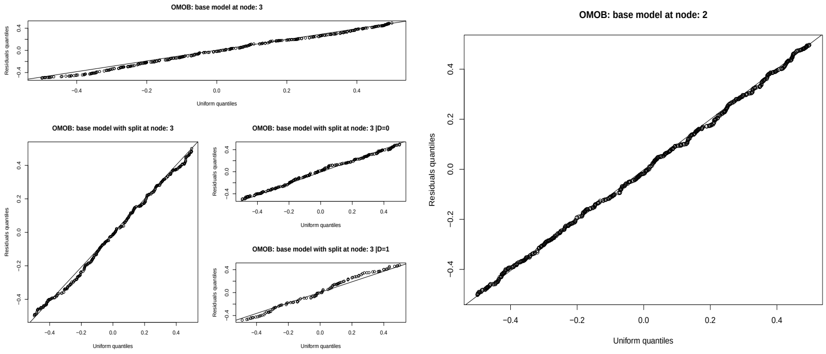

Figure 4. *Uniform QQ plot for local diagnostics on model* M : $Stress \sim Breadwinner$ *(minsplit=100)*

can be maintained. For instance, local uncertainty diagnostics of Binomial classification trees for rating data has been advanced in Simone, 2023. Further studies will investigate the impact of residual diagnostics on the derivation of variable importance measures from model-based tree ensembles.

References

CAPPELLI, C., SIMONE, R., & DI IORIO, F. 2019. CUBREMOT: a tool for building model-based trees for ordinal responses. *Expert Systems with Applications*, **124**, 39–49.

LIU, D., & ZHANG, H. 2018. Residuals and Diagnostics for Ordinal Regression Models: A Surrogate Approach. *Journal of the American Statistical Association*, **113(522)**, 845–854.

SIMONE, R. 2023. Uncertainty diagnostics of Binomial Regression Trees for Ordered Rating Data. *Journal of Classification*, **40**, 79–105. 10.1007/s00357-022-09429-5.

ZEILEIS, A., HOTHORN, T., & HORNIK, K. 2008. Model-Based Recursive Partitioning. *Journal of Computational and Graphical Statistics*, **17**, 492–514.