Linear regression pathmox segmentation tree: the case of visitors' satisfaction to attend a Spanish football match at the stadium

Cristina Davino, Giuseppe Lamberti

1. Introduction

Segmentation trees have been attracting a great deal of attention as model comparison tools, with research mainly motivated by the fact that segmentation trees allow identification of partitions of data characterised by different dependency structures. Few algorithms have been proposed by the statistical community that combine model estimation and segmentation trees, outside the MOdel-based recursive partitioning (MOB) procedure proposed by Zelies *et al.* (2008). In a new approach we generalize the pathmox algorithm developed by Lamberti *et al.* (2016) to the context of linear regression models, using a model comparison test to identify the most significant partitions (i.e., sub-groups) in data. Further developments of the proposed approach will involve extensions to other contexts such as quantile regression.

2. State-of-the-art

Analysis of a dependency model can be furthered by assessing whether a model and/or the impact of regressors on dependent variables differ if heterogeneity is observed. In other words, it may be interesting to assess differences between a global model estimated on the whole set of observations and models based on sub-groups identified on the basis of known categorical variables external to the model. These variables may identify partitions characterised by a dependency structure heterogeneity. The most popular approaches to comparing regression models rely on comparative statistical testing or on recursive methods. The comparison approach consists of comparing coefficients related to a model common to all the data (i.e., a restricted model representing a homogeneous situation) and another model that reflects the interactions between categorical and predictor variables (i.e., an unrestricted model corresponding to a heterogeneous situation). The comparison approach, which allows for analysis of one categorical variable at a time, is reflected in the F-tests developed by Chow (1960) and Lebart *et al.* (1979), based on an assumption of the normality of the residuals of the two models. Comparison is done by calculating restricted deviance (SSR_0) and unrestricted deviance (SSR_1). The latter will be lower if interaction between categorical and predictor variables is significant. Under the null hypothesis, if both types of deviance are equal, then the categorical variables produce no differences in model coefficients. This null hypothesis is tested by computing an F–statistic:

$$F = \frac{(SSR_0 - SSR_1)/(n-2p)}{SSR_1/p} \tag{1}$$

The recursive approach, based on multiple model comparisons, ranks variables that produce differences in the model coefficients. The outcome is a tree where each node represents a model. Partitions are obtained by comparing the effect of each categorical variable on the model coefficients and choosing the partitions that produce the biggest differences. This approach requires a criterion to quantify differences in the model coefficients. In case of the MOB

Cristina Davino, University of Naples Federico II, Italy, cristina.davino@unina.it, 0000-0003-1154-4209
Giuseppe Lamberti, Autonomous University of Barcelona, Spain, giuseppe.lamberti@uab.cat, 0000-0002-8666-796X

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Cristina Davino, Giuseppe Lamberti, *Linear regression pathmox segmentation tree: the case of visitors' satisfaction to attend a Spanish football match at the stadium*, pp. 137-140, © 2021 Author(s), CC BY 4.0 International, DOI 10.36253/978-88-5518-461-8.26, in Bruno Bertaccini, Luigi Fabbris, Alessandra Petrucci (edited by), *ASA 2021 Statistics and Information Systems for Policy Evaluation. Book of short papers of the on-site conference*, © 2021 Author(s), content CC BY 4.0 International, metadata CC0 1.0 Universal, published by Firenze University Press (www.fupress.com), ISSN 2704-5846 (online), ISBN 978-88-5518-461-8 (PDF), DOI 10.36253/978-88-5518-461-8

procedure this criterion is based on a fluctuation test that measures coefficient instability (Zelies and Hornick, 2007) as caused by a categorical variable. High instability points to a significant effect of the variable. Tree partitions are defined according to the variables that produce the highest instability.

3. Pathmox in a nutshell

Pathmox (Lamberti *et al.*, 2016), developed to detect heterogeneity in models, is a recursive algorithm based on segmentation trees. While pathmox was introduced in the context of partial least square structural equation modelling, it can be generalized to other contexts when a suitable test for comparing models is available. The algorithm applies binary segmentation principles to produce a tree with different models in each node. It starts by fitting a global model to all the data (i.e., the tree root) and identifies models with the most significant differences in child nodes. The most different models are identified by minimizing the sum of the squares of the residuals of the models estimated in each child node. The available data are recursively partitioned according to categorical variables – not included in the model – that yield the most significant differences in the child nodes. Partitions are identified using a test that determines the degree of difference between two compared sub-models. Finally, pathmox avoids overfitting using stopping rules based on maximum depth, minimum node size and non-significance of the partitioning criterion. As the partitioning criterion we propose the hypothesis test as proposed by Lebart *et al.* (1979) and Chow (1960) to compare two linear regression models.

4. Visitors' satisfaction to attend a Spanish football match: a pathmox application

We applied the pathmox approach in an empirical analysis to measure the visitors' satisfaction to attend a Spanish football match at the stadium. The sample consisted of visitors aged 18 years and older who attended Barcelona Football Club home matches during the 2017/2018 season. Visitors were selected using a no-random selection based on convenience. Three hours before matches started, randomly selected visitors were approached by seven researchers, who had previously reviewed and resolved any doubts regarding the questionnaire. The visitors were told about the purpose of the research and were asked to collaborate. If they agreed, they were asked to supply an email address to receive an online version of the questionnaire to be completed after the match. The questionnaire was available in the Catalan, Spanish, English and French languages to avoid bias due to the understanding of the questions by tourists.We offered the possibility of accessing the questionnaire through a QR code if they did not want to give an email address. Finally to encourage participation, respondents were entered in a lottery to win an authentic Barsa football shirt. A total of 944 visitors were invited to take part; the response rate of 38.45% meant that 362 usable questionnaires were collected. Men represented almost three-quarters (71.27%) of the respondents (women 28.72%), and nearly half (48.34%) were aged $\leq$30 years (34.52%, 31-45 years, and 16.71% $\geq$46 years). Involvement was strong, moderate, and slight in 32.50%, 48.76%, and 18.45% of the respondents, respectively. The percentage of tourist was 40,88% (no tourist 59.12%). 69.06% indicated that it was not the first time that they went to the Camp Nou Stadium.

The questionnaire was designed with closed questions answered on a 5-point Likert scale and aimed to measure the *visitors' satisfaction* in terms of perceived benefits of attending a Barcelona Football Club match (adapted from Ahrholdt et al., 2017; Oliver, 2010), *image of the football team* measured as visitors' perception of the attributes, players, management, and condition of the club (adapted from Beccarini, and Ferrand. 2006), and *stadium service quality* measured as visitors' perception of service performance, based on evaluations of several service dimensions as tickets price, accessibilities, stadium facilities (adapted from Ahrholdt et al., 2017). The following categorical variables, reflecting specific visitors' characteristics, were considered as potential sources of heterogeneity: *gender*, *age* ($\leq$30, 31-45, $\geq$46 years), *involvement* (strong, moderate, slight), *tourist* (yes or not), and *first time at the stadium* (yes or not).

Pathmox analysis results are reported in Figure 1 and Table 1. We set maximum depth to two levels, bounded the final number of segments to a maximum of four and set the minimum admissible node size to 10% of the total sample. The significance threshold for the partitioning algorithm was p=0.05. The pathmox algorithm identified *involvement* as the variable with the greatest power, distinguishing between not involved– slight – (Node 2) and involved – strong and moderate – (Node 3). Not involved visitors were differentiated according to the variable *tourist* defining two terminal nodes: Node 4 (no tourist) and Node 5 (tourist). Involved visitors were further differentiated according to *age*: visitors aged $\leq$30 years form one group (Node 6) while visitors aged $>$30 years (Node 7) form another. On the basis of involvement combined with age and tourist, we could characterise partitions and assign labels to sub-groups. Thus, Node 4 can be defined as the group of not involved-local visitors, Node 5 as not involved-tourist visitors, Node 6 as younger-involved visitors, and Node 7 as older-involved visitors. Finally, the global model coefficients were compared with the coefficients for the four models estimated for the sub-samples identified by the terminal nodes (Table 1), showing that, in terms of satisfaction, not involved-local visitors primarily valued the image of the football team, not involved-tourist visitors valued more the quality of the stadium, younger-involved valued both image and quality in a similar way, and the older-involved valued again primarily the image of the football team.

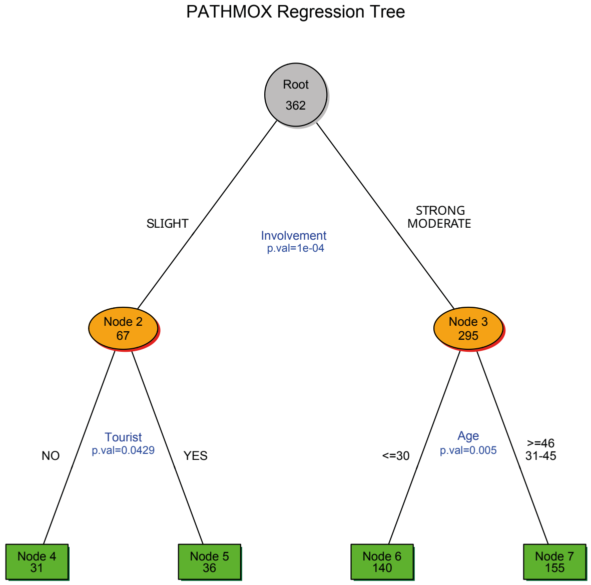

Figure 1: Pathmox tree

	Image of football team	Stadium service quality
Global model	0.493*	0.253*
Node 4: not involved-local	0.892*	0.468*
Node 5: not involved-tourist	0.319*	0.750*
Node 6: younger-involved	0.287*	0.243*
Node 7: older-involved	0.548*	0.166*

* indicates significance according to the t-test (p-value $<$0.05)

Table 1: Coefficient comparison for global and terminal nodes.

5. Discussion and conclusion

Our results suggest that pathmox can be used to compare regression models, opening up a future research line in other contexts such as quantile regression. From a decision-making perspective, the paper contributes evidence exemplifying how an apparently representative global model can in fact mask different relationships between variables due to heterogeneous data, underlining the importance of accounting for heterogeneity when defining new polices. While the algorithm allows partitions to be identified where differences between model coefficients are greatest, it has the limitation that no overall significance criterion is considered once each partition is identified. This important aspect needs to be considered in a future version of the algorithm. Note that pathmox aims to identify the most significantly different sub-groups, unlike a classic decision tree where the objective is to obtain the best prediction based on splitting observations into sub-groups. Therefore, the only similar method is the MOB proposed by Zelies *et al.* (2008), which, however, uses a different criterion to identify the best partitions. A comparison of both approaches will be a natural next step in our research.

References

Ahrholdt, D.C., Gudergan, S.P. and Ringle, C.M. (2017). Enhancing service loyalty: the roles of delight, satisfaction, and service quality. *Journal of Travel Research*, **56**, pp. 436–450.

Beccarini, C., and Ferrand, A. (2006). Factors affecting soccer club season ticket holders' satisfaction: the Influence of club image and fans' motives. *European Sport Management Quarterly*, **6**(1), pp. 1–22.

Chow, G.C. (1986). Test of equality between sets of coefficients in two linear regressions. *Econometrica*, **28**, pp. 591–605.

Lamberti, G., Aluja, T, Sanchez, G. (2016). The Pathmox approach for PLS path modeling. *Applied Stochastic Models in Business and Industry*, **32**, pp. 453–468.

Lebart, L., Morineau, A, Feenelon, J.P.(1979). *Traitement des donnees statistiques*. Dunod, Paris.

Oliver, R.L. (2010) *Satisfaction: a behavioral perspective on the consumer*. Armonk, NY: M.E. Sharpe.

Zeileis, A., Hornik, K., (2007). Generalized M-fluctuation tests for parameter instability. *Statistica Neerlandica*, **61**, pp. 488–508.

Zeileis, A., Hothorn, T., Hornik, K., (2008). Model-Based Recursive Partitioning. *Journal of Computational and Graphical Statistics*, **17**, pp. 492–514.