Computational Statistics and Data Analysis xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Tree-based varying coefficient regression for longitudinal ordinal responses

Q1 Reto Bürgin *, Gilbert Ritschard

National Center of Competence in Research LIVES, Switzerland
Institute of Demographic and Life Course Studies, University of Geneva, Switzerland

ARTICLE INFO

Article history:
Received 11 June 2014
Received in revised form 5 January 2015
Accepted 5 January 2015
Available online xxxx

Keywords:
Recursive partitioning
Varying coefficient models
Mixed models
Generalized linear models
Longitudinal data analysis
Ordinal regression
Statistical learning

ABSTRACT

A tree-based algorithm for longitudinal regression analysis that aims to learn whether and how the effects of predictor variables depend on moderating variables is presented. The algorithm is based on multivariate generalized linear mixed models and it builds piecewise constant coefficient functions. Moreover, it is scalable for many moderators of possibly mixed scales, integrates interactions between moderators and can handle nonlinearities. Although the scope of the algorithm is quite general, the focus is on its usage in an ordinal longitudinal regression setting. The potential of the algorithm is illustrated by using data derived from the British Household Panel Study, to show how the effect of unemployment on self-reported happiness varies across individual life circumstances.[1]

1. Introduction

Regression analysis for longitudinal responses addresses a wide range of applications, particularly in medical and social sciences. Siddall et al. (2003), for example, analyze long-term effects of injuries on repeatedly measured pain. Likewise, Oesch and Lipps (2013) use repeatedly measured well-being to examine the impact of the transition from employment to unemployment.

When carrying out longitudinal regression analysis, researchers are specifically interested in the impact of moderator variables on selected regression coefficients in order to enhance insights on the studied relation and/or to control for confounding variables. For example, the effect of an injury could depend on age, while that of unemployment could vary across social groups. Herein, we propose a method to learn such moderation in longitudinal data. The method combines a mixed model approach with a regression tree approach. Although the proposed method applies generally in the multivariate generalized linear mixed model (MGLMM) setting, we focus on its usage with longitudinal ordinally scaled responses such as pain or well-being.

The remainder of the article is organized as follows. Sections 1.1 and 1.2 introduce the framework used in the present study and related works. Section 2 explains the method in detail. Section 3 illustrates its potential by using an empirical example and simulation studies and, finally, Section 4 concludes, including addressing the limitations of the proposed method and the software implementation.

* Correspondence to: University of Geneva (CH), Bd du Pont d'Arve 40, Switzerland. Tel.: +41 22 379 98 72.
E-mail addresses: Reto.Buergin@unige.ch (R. Bürgin), Gilbert.Ritschard@unige.ch (G. Ritschard).

[1] R-codes and datasets are available online as supplementary files (see Appendix B).

http://dx.doi.org/10.1016/j.csda.2015.01.003
0167-9473/

1.1. Framework

The proposed algorithm extends multivariate generalized linear mixed models (e.g. Tutz and Hennevogl, 1996) by allowing the fixed coefficients to vary as nonparameterized functions of some moderator variables $Z_1, \ldots, Z_L$. Let $\mathbf{y}_{it}$ denote the $R \times 1$ response vector of individual i at time $t, i = 1, \ldots, N, t = 1, \ldots, N_i$. Denote by $\mathbf{X}_{it}$ and $\mathbf{W}_{it}$ the $Q \times P_{\boldsymbol{\beta}}$ and $Q \times P_{\mathbf{b}_i}$ design matrices associated with fixed coefficients $\boldsymbol{\beta}$ and (individual-specific) random coefficients $\mathbf{b}_i$, respectively. Further, denote by $\mathbf{z}_{it}$ the $L \times 1$ vector of moderators, also called *effect modifiers* in the literature (e.g. Hastie and Tibshirani, 1993). MGLMMs link the $Q \times 1$ predictor vector $\boldsymbol{\eta}_{it}$ with the conditional expectation $\boldsymbol{\mu}_{it} = E(\mathbf{y}_{it}|\mathbf{b}_i; \mathbf{X}_{it}, \mathbf{W}_{it}, \mathbf{z}_{it})$ via $\boldsymbol{\mu}_{it} \in \mathbb{R}^R \mapsto \boldsymbol{\eta}_{it} = \mathbf{g}(\boldsymbol{\mu}_{it}) \in \mathbb{R}^Q$, where $\mathbf{g}$ is a known link function. We aim to fit predictor functions of the form

$$\mathcal{M} : \boldsymbol{\eta}_{it} = \mathbf{X}_{it}\boldsymbol{\beta}(\mathbf{z}_{it}) + \mathbf{W}_{it}\mathbf{b}_i, \quad \mathbf{b}_i \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_b). \tag{1}$$

The fixed coefficients $\boldsymbol{\beta}(\cdot) = (\beta_1(\cdot), \ldots, \beta_{P_{\boldsymbol{\beta}}}(\cdot))^\top$ of $\mathcal{M}$ are *varying coefficients* that state that the linear effects of the elements of matrix $\mathbf{X}_{it}$ on the expectation of $\mathbf{y}_{it}$ are nonparameterized functions of $\mathbf{z}_{it}$. In the predictor function $\mathcal{M}$, the intercept coefficients are included in $\boldsymbol{\beta}(\cdot)$. Such *varying intercepts* are functions of $\mathbf{z}_{it}$ and estimate the direct effects of $\mathbf{z}_{it}$ on $E(\mathbf{y}_{it}|\cdot)$. In contrast to fixed coefficients, the individual-specific random coefficients $\mathbf{b}_i$ do not depend on $\mathbf{z}_{it}$ in $\mathcal{M}$. Such random coefficients are used to take into account the correlation between repeated responses and could include individual-specific intercepts or slopes over time. As stated in (Eq. (1)), we assume here that the random coefficients are normally, identically and independently distributed with $\mathrm{E}(\mathbf{b}_i) = \mathbf{0}$ and $\mathrm{Var}(\mathbf{b}_i) = \boldsymbol{\Sigma}_b$.

MGLMMs include models with density functions of the multivariate exponential family that, with random coefficients $\mathbf{b}_i$, have the general form

$$f(\mathbf{y}_{it}|\mathbf{b}_i; \boldsymbol{\beta}, \phi) = \exp\left\{\frac{\mathbf{y}_{it}^\top\boldsymbol{\theta}_{it} - b(\boldsymbol{\theta}_{it})}{\phi} + c(\mathbf{y}_{it}, \phi)\right\}, \tag{2}$$

with ϕ the dispersion parameter and $b(\cdot)$ and $c(\cdot)$ family-specific functions. $\boldsymbol{\theta}_{it}$ is the so-called vector of natural parameters. It is here defined as $\boldsymbol{\theta}_{it} = \mathbf{d}(\boldsymbol{\mu}_{it}) = \mathbf{d}(\mathbf{g}^{-1}(\mathbf{X}_{it}\boldsymbol{\beta}(\mathbf{z}_{it}) + \mathbf{W}_{it}\mathbf{b}_i))$, with $\mathbf{d}(\cdot)$ a known, vector-valued function. MGLMMs include, for instance, several univariate models such as the (Gaussian) linear mixed model or the Poisson mixed model. Here, we restrict the consideration of specific models to that of the cumulative logit mixed model, which really requires the multivariate form above.

The cumulative logit mixed model (CLMM). The cumulative logit model (e.g. McCullagh, 1980) is a popular and conceptually simple model for ordinal response variables Y taking ordered categorical values r in $\{1, \ldots, R\}$. It is motivated (e.g. Tutz, 2012) by assuming that Y is a coarse version of a latent continuous variable $Y^* = f(\cdot) + \varepsilon$, with $f(\cdot)$ a function of predictors and ε the error with distribution $\varepsilon \overset{i.i.d.}{\sim} \mathrm{Logistic}(0, 1)$. The connection between the observed ordinal and the latent variable is defined as: $Y = r \Leftrightarrow \theta_{r-1} < Y^* \leq \theta_r$; with $-\infty = \theta_0 < \theta_1 < \cdots < \theta_R = \infty$ the *threshold* coefficients.

The cumulative logit mixed model has been introduced by Hedeker and Gibbons (1994), and Tutz and Hennevogl (1996) exemplified it as a special case of MGLMMs. Here, the CLMM with varying coefficients is defined as follows: Let $\mathbf{y}_{it} = (y_{it1}, \ldots, y_{itR})^\top$ be the response vector of individual i at time t, which is coded as $y_{itr} = 1$ if $Y_{it} = r$ and $y_{itr} = 0$ if $Y_{it} \neq r$. Assume that $\mathbf{y}_{it}$ is an outcome of a multinomial distribution with the conditional probabilities $E(\mathbf{y}_{it}|\mathbf{b}_i; \mathbf{x}_{it}, \mathbf{w}_{it}, \mathbf{z}_{it}) = \boldsymbol{\pi}_{it}$, with $\mathbf{x}_{it}$ and $\mathbf{w}_{it}$ the predictor vectors to be incorporated into the design matrices $\mathbf{X}_{it}$ and $\mathbf{W}_{it}$. The CLMM links the predictor $\boldsymbol{\eta}_{it}$ with the conditional probabilities $\boldsymbol{\pi}_{it}$ via $\eta_{itq} = g_q(\boldsymbol{\pi}_{it}) = \log((\pi_{it1} + \cdots + \pi_{itq})/(1 - \pi_{it1} - \cdots - \pi_{itq})) = \mathrm{logit}(P(Y_{it} \leq q))$ for $q = 1, \ldots, Q = R - 1$. The predictor function is defined as

$$\mathcal{M}_{\mathrm{CLMM}} : \begin{bmatrix} \eta_{it1} \\ \vdots \\ \eta_{itQ} \end{bmatrix} = \begin{bmatrix} 1 & & \mathbf{x}_{it}^\top \\ & \ddots & \\ & 1 & \mathbf{x}_{it}^\top \end{bmatrix} \boldsymbol{\beta}(\mathbf{z}_{it}) + \begin{bmatrix} 1 & \mathbf{w}_{it}^\top \\ \vdots & \vdots \\ 1 & \mathbf{w}_{it}^\top \end{bmatrix} \mathbf{b}_i, \tag{3}$$

where the qth row determines the logits of responding with $\{1, \ldots, q\}$ rather than with $\{q+1, \ldots, R\}$. The first Q elements of $\boldsymbol{\beta}(\cdot)$ are the varying intercepts, or *varying thresholds* $\theta_1(\cdot), \ldots, \theta_{R-1}(\cdot)$ in terms of the latent variable motivation, that take into account the direct effects of the moderators $\mathbf{z}_{it}$. In order to maintain the order $P(Y_{it} \leq 1) \leq \cdots \leq P(Y_{it} \leq Q)$, these intercepts must satisfy $\beta_1(\mathbf{z}_{it}) \leq \cdots \leq \beta_Q(\mathbf{z}_{it})\ \forall\ (i, t)$. Further, stacking the vectors $\mathbf{x}_{it}^\top$ and $(1, \mathbf{w}_{it}^\top)$ in the design matrices constraints the corresponding effects to be identical for all Q cumulative logits. This constraint, which considerably simplifies the model, is commonly called the *proportional odds assumption* (e.g. McCullagh, 1980) or *parallelism*. For the direct effects of $\mathbf{z}_{it}$, the proportional odds assumption is relaxed in $\mathcal{M}_{\mathrm{CLMM}}$ since the corresponding varying intercepts are logit-specific. Therefore, $\mathcal{M}_{\mathrm{CLMM}}$ can be seen as a *partial proportional odds model* (e.g. Tutz, 2012, Chap. 9.1.3). Note that if $R = 2$, $\mathcal{M}_{\mathrm{CLMM}}$ is equivalent to a logistic mixed model.

The unknown varying coefficients $\boldsymbol{\beta}(\cdot)$ of the predictor function $\mathcal{M}$ (Eq. (1)) are proposed to be approximated by a piecewise constant function, based on *model-based recursive partitioning*, which is conceptually similar to *regression trees* (e.g. Breiman et al., 1984). These two approaches can be distinguished by their aims: regression trees attempt to discover differences in the mean, while model-based recursive partitioning aims to discover differences in the model coefficients. While recursive partitioning has certain drawbacks, particularly that it is a heuristic and may be instable regarding small changes in the data, its advantages for statistical learning are hardly covered by the alternative methods to date (cf. Hastie

Please cite this article in press as: Bürgin, R., Ritschard, G., Tree-based varying coefficient regression for longitudinal ordinal responses. Computational Statistics and Data Analysis (2015), http://dx.doi.org/10.1016/j.csda.2015.01.003