Collection of Biostatistics Research Archive
COBRA Preprint Series

Regression Trees for Longitudinal Data

Madan Gopal Kundu[*]   Jaroslaw Harezlak[†]

[*]Indiana University Fairbanks School of Public Health, Department of Biostatistics, mgkundu@iupui.edu

[†]Indiana University Fairbanks School of Public Health, Department of Biostatistics

Regression Trees for Longitudinal Data

Madan Gopal Kundu and Jaroslaw Harezlak

Abstract

Often when a longitudinal change is studied in a population of interest we find that changes over time are heterogeneous (in terms of time and/or covariates' effect) and a traditional linear mixed effect model [Laird and Ware, 1982] on the entire population assuming common parametric form for covariates and time may not be applicable to the entire population. This is usually the case in studies when there are many possible predictors influencing the response trajectory. For example, Raudenbush [2001] used depression as an example to argue that it is incorrect to assume that all the people in a given population would be experiencing either increasing or decreasing levels of depression. In such cases, a group-averaged trajectory can mask important subgroup differences. Our aim is to identify and characterize longitudinally homogeneous subgroups based on the combination of baseline covariates. We achieve this goal by constructing regression tree through binary partitioning. We propose two steps procedure for binary partitioning: 1) first, choose the most significant partitioning variable and 2) then choose the best split by repetitive evaluation of a goodness of fit criterion at all the splits of chosen partitioning variable. To remedy for the problem of multiple testing, we propose a single test to identify the instability of parameter(s) in longitudinal models for a given partitioning variable. We obtain asymptotic results and examine finite sample behavior of our method through simulation studies. Finally, we apply our method to study the changes in brain metabolite levels of HIV infected patients.

Regression Trees for Longitudinal Data

Madan Gopal Kundu and Jaroslaw Harezlak

Indiana University Fairbank School of Public Health, Indianapolis, IN

September 30, 2013

Abstract

Often when a longitudinal change is studied in a population of interest we find that changes over time are heterogeneous (in terms of time and/or covariates' effect) and a traditional linear mixed effect model [18] on the entire population assuming common parametric form for covariates and time may not be applicable to the entire population. This is usually the case in studies when there are many possible predictors influencing the response trajectory. For example, Raudenbush [28] used depression as an example to argue that it is incorrect to assume that all the people in a given population would be experiencing either increasing or decreasing levels of depression. In such cases, a group-averaged trajectory can mask important subgroup differences. Our aim is to identify and characterize longitudinally homogeneous subgroups based on the combination of baseline covariates. We achieve this goal by constructing regression tree through binary partitioning. We propose two steps procedure for binary partitioning: 1) first, choose the most significant partitioning variable and 2) then choose the best split by repetitive evaluation of a goodness of fit criterion at all the splits of chosen partitioning variable. To remedy for the problem of multiple testing, we propose a single test to identify the instability of parameter(s) in longitudinal models for a given partitioning variable. We obtain asymptotic results and examine finite sample behavior of our method through simulation studies. Finally, we apply our method to study the changes in brain metabolite levels of HIV infected patients.

Keywords: Regression trees, Instability test, Longitudinal data, Mixed models, Score process, Brownian Bridge

1 Introduction

In longitudinal studies, repeated measurements of the outcome variable are often collected at irregular and possibly subject-specific time points. Parametric regression methods for analyzing such data have been developed by Laird and Ware [18] and Liang and Zeger [20] among others, and have been summarized by Diggle *et al.* [10]. Often the population under consideration is heterogeneous in terms of trend and covariate effect. Under such situation traditional mixed effect models (such as, linear mixed effect model) assuming a common parametric form for covariates and time might not be an appropriate option. If the population under consideration is diverse and there exists several distinct subgroups within it, the true parameter value(s) for longitudinal mixed effect model may vary between these subgroups. For example, Raudenbush [28] used a longitudinal depression study as an example to argue that it is incorrect to assume that all the people in a given population will be experiencing either increasing or decreasing levels of depression. In such instances, an assumption
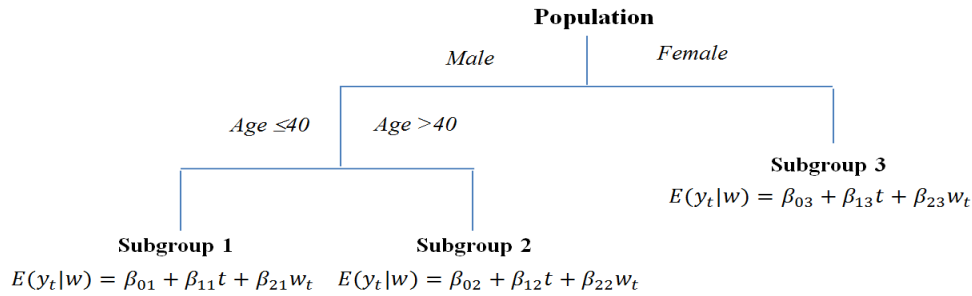
1

Figure 1: Sample longitudinal tree. The population consists of 3 subgroups and they differ in their longitudinal profiles. These subgroups are defined by the partitioning variables gender and age.

of common parametric form will mask important subgroup differences and will lead to erroneous conclusions. In our work, our interest is to identify meaningful and interpretable subgroups with differential longitudinal trajectories and/or differential covariate effect(s) on the response variable from such a heterogeneous population. We propose a regression tree construction technique with longitudinal data that (1) controls type I error at each split, and (2) is applicable in cases when measurements are taken at subject specific time-points.

When the longitudinal profile in a population depends on some baseline attributes, the most common strategy is to include these attributes (and their interaction terms) as covariates in the model. However, this strategy has some inherent drawbacks: (a) it can lead to overfitting due to inclusion of all possible interaction terms, especially when the number of potential baseline attributes is large, (b) functional form of the association with baseline attributes need to be known and correctly specified, and (c) it cannot capture nonlinear effect of baseline attributes. Our goal is to determine the most parsimonious model consisting of a number of homogeneous subgroups from a heterogeneous population profile without strict parametric restrictions or prior information. One of the popular technique to construct homogeneous subgroups is latent class modeling (LCM) [25]. LCM is a statistical method used to identify a set of discrete, mutually exclusive latent classes of individuals based on their responses. An alternative approach is to construct regression tree with longitudinal data[29]. Advantages of regression tree technique over LCM are: (1) it characterizes the subgroups in terms of partitioning variables and (2) number of subgroups need not to be known a-priori. In general, the thrust of any tree techniques is the extraction of meaningful subgroups characterized by common covariate values and homogeneous outcome. For longitudinal data, this homogeneity can pertain to the mean and/or covariance structure [29].

Throughout this article, we refer to the regression tree with longitudinal data as '*Longitudinal tree*'. Figure 1 displays a toy example for a longitudinal tree. This regression tree represents a heterogeneous population with three distinct subgroups in terms of their longitudinal profiles. These subgroups can be characterized by *gender* and *age*. Here, *gender* and *age* are baseline attributes. These baseline attributes are commonly referred as *partitioning variables*, because they partition heterogeneous population into homogeneous subgroups. In each of the three subgroups, the longitudinal trajectory depends on the covariates $w_1, \cdots, w_q$, but these subgroups are heterogeneous in terms of the true coefficients associated with their longitudinal profiles. Consider the following form of a linear

2

longitudinal mixed effect model

$$y_{it} = \beta_0^x + \beta_1^x t + \mathbf{w}_{it}^\top \boldsymbol{\beta}^x + \mathbf{z}_{it}^\top \mathbf{b}_i + \epsilon_{it} \tag{1.1}$$

where i is the subject index and y, t and $\mathbf{w}$ denote the outcome variable, time and the vector of measurements on scalar covariates $w_1, \cdots, w_q$, respectively. Let $X^{G_1}, \cdots, X^{G_s}$ include all potential baseline attributes that might influence the longitudinal trajectory in (1.1). The superscript x is added to the coefficients β_0, β_1 and $\boldsymbol{\beta}$ to reflect their possible dependence on these baseline attributes. Denote $\boldsymbol{\theta}^{x\top} = (\beta_0^x, \beta_1^x, \boldsymbol{\beta}^{x\top})$. With such a model, '*homogeneity*' refers to the situation when the coefficients' true values remain the same for all the individuals in the entire population, *i.e.* $\boldsymbol{\theta}^x = \boldsymbol{\theta}$. When the longitudinal changes in the population of interest are *heterogenous* there exists distinct subgroups differing in terms of the true values of the coefficients, *i.e.* $\boldsymbol{\theta}^x \neq \boldsymbol{\theta}$. $X^{G_1}, \cdots, X^{G_s}$ are the partitioning variable used in the regression tree construction.

In constructing a longitudinal tree through binary partitioning, one way to choose a partition is via maximizing improvement in goodness of fit criterion. For example, Abdolell *et al.* [1] chose deviance as goodness of fit criterion. They evaluated deviance at each split of a given partitioning variable and selected the partition with maximum reduction in deviance for the binary splitting. However, repetitive evaluation of goodness-of-fit criterion leads to the multiple testing problem. In order to avoid it, we follow a different strategy to identify the best split for partitioning. For the simplicity sake, we first explain our approach assuming that there is only a single partitioning variable, say X^G, with G cut-off points. In such case, we identify the best split in a two-step process as follows:

- **Step 1.** Perform an overall test to detect any evidence of heterogeneity of longitudinal model parameters across G cut-off points of X^G.
- **Step 2.** Given that there is a 'significant' evidence for heterogeneity, the split that provides maximal improvement in goodness of fit criterion is chosen as a partitioning point for the tree construction.

Note that we perform only a single statistical test in step 1. Step 2 is conditional on the outcome of step 1 and no additional statistical tests are done. Since single overall test replaces the multiple testing, we avoid the problem of multiplicity. Throughout this article, we refer to the overall test performed in step 1 as a *test for instability*. In the situations when more than one partitioning variable is present, we repeat the *instability* test for each variable and apply multiplicity correction based on the number of variables tested. We continue to the second step using the 'most significant' partitioning variable. Details of this algorithm are presented in Section 4.

In order to construct a test for *instability*, we borrow an idea from the time-series literature. In time-series context often the goal is to evaluate whether the parameter of a regression model is stable across different time points. This is often known as a *test for structural change* or *constancy of parameters* [e.g., 6, 27, 13]. We apply very similar idea to evaluate whether the true values of the parameter remains the same across the cut-off values of a partitioning variable in a mixed effects longitudinal model of interest.

In this paper we utilize the parameter instability test in multiple ways. First, in the case of continuous partitioning variables, the proposed test uses the results on score process derived by Hjort and Koning [13] in conjunction with the properties of Brownian motion and Brownian Bridge. Second,

3

for categorical partitioning variables with a small number of cut-off points, a test for parameter instability is derived in a straightforward way by employing asymptotic normality of the score functions. We derive the asymptotic properties of the instability test and explore its size and power through an extensive simulation study. Finally, we use these instability tests to construct an algorithm for regression trees with longitudinal data.

Among the tree based methods, classification and regression tree (CART) method [5, 31, 7] is probably the most popular one. Zeileis *et al.* [33] have extended the concept of CART methodology in the context of fitting cross-sectional regression models. Binary partitioning for longitudinal data has been proposed first by Segal [29]. However, Segal's implementation is restricted to longitudinal data with a regular structure, that is all the subjects have an equal number of repeated observations at the same time points [35]. In the Zhang [34]'s work, the multivariate adaptive splines are used to analyze longitudinal data. Their method, multivariate adaptive splines for the analysis of longitudinal data (MASAL), can be used to generate regression trees for longitudinal data. Abdolell *et al.* [1] have used deviance as a goodness-of-fit criterion for binary partitioning. They controlled the level of Type I error via permutation test taking into account testing multiplicity. However, permutation tests are computer intensive and the time taken to fit the models is intimidatingly high even for medium-sized data. Sela and Simonoff [30] as well as Galimberti and Montanari [12] merged the subgroup differences with the random individual differences. They constructed the regression tree through an iterative two-step process. In the first step, they obtained the random effects' estimates and in the second step, they constructed the regression tree ignoring the longitudinal structure. They repeat these two steps until the estimates of the random effect converge in the first step. Our proposed method of constructing a regression tree with longitudinal data provides an improvement over the existing methods in the following aspects: (1) it maintains the level of Type I error at each split, (2) it is applicable to the cases of measurements taken at subject-specific time points, (3) it does not merge group differences with the random subject effect components and (4) it reduces computational time.

The remainder of this paper is organized as follows. In Section 2 the longitudinal mixed effect models of interest are summarized. Tests for parameter instability for continuous and categorical partitioning variable cases are discussed separately in Section 3. Algorithm for constructing regression trees along with measures of improvement and a pruning technique are discussed in Section 4. Results from the simulation studies examining the performance of the instability test and the regression tree as a whole are reported in Section 5. An application of the longitudinal regression tree method is illustrated on the metabolite data collected from the chronically HIV-infected patients in Section 6. We present our conclusions in Section 7.

2 Preliminary

Let $\{y_{it}, \mathbf{w}_{it}\}$ be a set of measurements recorded on the ith subject ($i = 1, \ldots, N$) at time $t = (t_1, \ldots, t_{n_i})$, where y is a continuous scalar outcome; and $\mathbf{w}$ is the vector of measurements on scalar covariates $w_1, \cdots, w_q$. We assume that the functional form of dependence of y on these covariates is already known. In addition, for each individual, we observe a vector of attributes $(X_i^{G_1}, \cdots, X_i^{G_S})$ measured at baseline. We assume that $X^{G_1}, \cdots, X^{G_s}$ includes all the potential baseline attributes that might influence the longitudinal trajectory of y and its association with covariates $w_1, \cdots, w_q$.

4

Further, we don't have any idea about the functional form of influence of these baseline attributes. We use the variables $X^{G_1}, \cdots, X^{G_S}$ as the candidate partitioning variables to construct a longitudinal regression tree to discover meaningful and interpretable subgroups with differential changes in y characterized by the $X^{G_1}, \cdots, X^{G_S}$.

When the longitudinal profile is homogeneous in the entire population, we can fit the following traditional linear mixed effect model for all N individuals [18]

$$y_{it} = \beta_0 + \beta_1 t + \mathbf{w}_{it}^\top \boldsymbol{\beta} + \mathbf{z}_{it}^\top \mathbf{b}_i + \epsilon_{it}, \tag{2.1}$$

where $\epsilon_{it} \sim N(0, \sigma^2)$ and $\mathbf{b}_i$ is the vector of random effects pertaining to subject i and distributed as $N(0, \sigma^2 \mathbf{D})$. By '*homogeneity*' we mean that the true value of $\boldsymbol{\theta}^\top = (\beta_0, \beta_1, \boldsymbol{\beta}^\top)$ remains the same for all the individuals. In fact, (2.1) is the simplified version of model in (1.1) under homogeneity.

We follow the common assumptions made in longitudinal modeling that $\mathbf{z}_{it}$ is a subset of $[\mathbf{w}_{it}^\top \ t]^\top$; ϵ_{it} and $\mathbf{b}_i$ are independent; ϵ_{it} and $\epsilon_{i't'}$ are independent whenever $i \neq i'$ or $t \neq t'$ or both, and $\mathbf{b}_i$ and $\mathbf{b}_{i'}$ are independent if $i \neq i'$. Here, $\mathbf{w}_{it}^\top \beta$ is the fixed effect term and $\mathbf{z}_{it}^\top \mathbf{b}_i$ is the standard random effects term. For the ith subject, we rewrite the Eq. (2.1) as follows

$$\mathbf{y}_i = \mathbf{w}_i \boldsymbol{\theta} + \mathbf{z}_i \mathbf{b}_i + \epsilon_i, \tag{2.2}$$

where $\mathbf{y}_i^\top = (y_{i1}, \cdots, y_{in_i})$, $\mathbf{w}_i$ is the design matrix consisting of the intercept, time (t) and covariates ($\mathbf{w}$). n_i is the number of observations obtained from the ith individual. The score function for estimating $\boldsymbol{\theta}$ under (2.2) is [see e.g., 9]

$$\mathbf{u}(\mathbf{y}_i, \boldsymbol{\theta}) = \frac{d}{d\boldsymbol{\theta}} l(\mathbf{y}_i, \boldsymbol{\theta}) = \frac{1}{\sigma^2} \mathbf{w}_i^\top \mathbf{V}_i^{-1} (\mathbf{y}_i - \mathbf{w}_i \boldsymbol{\theta})$$

where $\mathbf{V}_i = \mathbf{I} + \mathbf{z}_i \mathbf{D} \mathbf{z}_i^\top$ and $\mathbf{e}_i = \mathbf{y}_i - \mathbf{w}_i \boldsymbol{\theta}$. Further, its variance is

$$\text{Var}\left[\mathbf{u}(\mathbf{y}_i, \boldsymbol{\theta})\right] = \mathbf{J}(\boldsymbol{\theta}) = -E\left[\frac{d}{d\boldsymbol{\theta}} \mathbf{u}(\mathbf{y}_i, \boldsymbol{\theta})\right] = \frac{1}{\sigma^2} \mathbf{w}_i^\top \mathbf{V}_i^{-1} \mathbf{w}_i$$

Likelihood estimate of $\boldsymbol{\theta}$ obtained using all the observation from N subjects is valid only if all the individuals under considerations are homogeneous. If the individuals are not homogeneous in terms of $\boldsymbol{\theta}$ then the likelihood estimate obtained considering all the subjects together are misleading; the extent and direction of ambiguity in the estimate will depend on the nature and proportion of heterogeneity in the sampled individuals. Therefore, it is important to decide first whether the true value of $\boldsymbol{\theta}$ remains the same for all the subjects or not. In the next section, we describe a way to test whether the true value of $\boldsymbol{\theta}$ remains the same across all the values of a given partitioning variable.

3 Test for parameter instability

In this section, we utilize the ideas introduced by Hjort and Koning [13] to test for the constancy of model parameters over time in time-series context. Our goal is to test whether the true value of $\boldsymbol{\theta}$ remains the same across all distinct values of a given partitioning variable. We refer to this test as a *test for parameter instability*. The testing strategy is described in this section with a single

partitioning variable. For multiple partitioning variables, the test needs to be repeated for each of them with an adjustment for multiple testing.

Let $X^G \in \{X^{G_1}, \cdots, X^{G_S}\}$ be any partitioning variable with G ordered cut-off points $c_{(g)}, g = 1, \cdots, G; c_{(1)} \leq \cdots \leq c_{(G)}$ and $\boldsymbol{\theta}_{(g)}$ be the true value of $\boldsymbol{\theta}$ when $X^G = c_{(g)}$. Assume that there are m_g subject with $X^G = c_{(g)}$. We denote the cumulative number of subjects with $X^G \leq c_{(g)}$ by M_g. That is, $M_g = \sum_{j=1}^g m_j$ and $M_G = \sum_{j=1}^G m_j = N$. We want to conduct an omnibus test,

$$H_0 : \boldsymbol{\theta}_{(g)} = \boldsymbol{\theta}_0 \quad H_1 : \boldsymbol{\theta}_{(g)} \neq \boldsymbol{\theta}_0.$$

Here, H_0 indicates the situation when parameter $\boldsymbol{\theta}$ remains constant (that is, *homogeneity*) and H_1 corresponds to the situation of parameter instability (that is, *heterogeneity*) . The two tests described in this section utilize the following properties of score function under H_0:

- A1: $E_{H_0}[\mathbf{u}(\mathbf{y}_i, \boldsymbol{\theta}_0)] = 0;$
- A2: $\text{Var}_{H_0}[\mathbf{u}(\mathbf{y}_i, \boldsymbol{\theta}_0)] = \mathbf{J}(\boldsymbol{\theta}_0) = \mathbf{J};$
- A3: $\mathbf{u}(\mathbf{y}_i, \hat{\boldsymbol{\theta}})|_{H_0} \rightarrow^d N[0, \hat{\mathbf{J}}],$

where $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimate of $\boldsymbol{\theta}$ and $\hat{\mathbf{J}} = \mathbf{J}(\hat{\boldsymbol{\theta}})$. We discuss the instability test separately for the categorical and continuous variables X^G.

3.1   Instability test with a categorical partitioning variable

It is straightforward to obtain a test for parameter instability using the properties A1–A3 when the partitioning variable, X^G, is categorical with a small number of categories (that is, $G \ll N$). Since the score functions $\mathbf{u}(\mathbf{y}_i, \hat{\boldsymbol{\theta}})$ are independent, we have under H_0, the following quantity

$$\chi^2_{cat} = \sum_{g=1}^G \left[\sum_{i=1}^N I(X_i^G = c_{(g)})\mathbf{u}(\mathbf{y}_i, \hat{\boldsymbol{\theta}})\right]^\top \left[m_g \hat{\mathbf{J}}\right]^{-1} \left[\sum_{i=1}^N I(X_i^G = c_{(g)})\mathbf{u}(\mathbf{y}_i, \hat{\boldsymbol{\theta}})\right]$$

is asymptotically distributed as χ^2 with $(G-1)p$ degrees of freedom where p is the dimension of $\boldsymbol{\theta}$. Here, $I(\cdot)$ is the indicator function. The reduction in p degrees of freedom is due to the estimation of p dimensional $\boldsymbol{\theta}$ from the data.

3.2   Instability test with continuous partitioning variable

Here, we first review the results obtained by Hjort and Koning [13] and then we propose the test for instability with a single continuous partitioning variable. We begin by defining the following *score process*

$$\mathbf{W}_N(t, \boldsymbol{\theta}_0) = N^{-1/2} \sum_{i=1}^{M_g} \mathbf{u}(\mathbf{y}_i, \boldsymbol{\theta}_0) \qquad t \in [t_g, t_{g+1})$$

where $t_g = \dfrac{M_g}{N}$. Using multivariate version of Donsker's theorem and Cramér-Wold theorem [see e.g. 3] it can be shown that

$$\mathbf{W}_N(t, \boldsymbol{\theta}_0) \rightarrow_d \mathbf{Z}(t)$$

6

where $\mathbf{Z}(t)$ is the zero-mean Gaussian process with $\text{cov}[\mathbf{Z}(t), \mathbf{Z}(s)] = \min(t, s)\mathbf{J}(\boldsymbol{\theta}_0)$. Note that Z is a linear transformation of p independent Brownian motions. Since, $\boldsymbol{\theta}_0$ is unknown in practice, we define the following *estimated score process* replacing $\boldsymbol{\theta}_0$ by $\hat{\boldsymbol{\theta}}$

$$\mathbf{W}_N(t, \hat{\boldsymbol{\theta}}) = N^{-1/2} \sum_{i=1}^{M_g} \mathbf{u}(\mathbf{y}_i, \hat{\boldsymbol{\theta}})$$

By applying Taylor series expansion it is straightforward to show that

$$\mathbf{W}_N(t, \hat{\boldsymbol{\theta}}) \doteq \mathbf{W}_N(t, \boldsymbol{\theta}_0) - t\, \mathbf{W}_N(1, \boldsymbol{\theta}_0)$$

where $A_n \doteq B_n$ means that $A_n - B_n$ tends to zero in probability. In the case of linear mixed effects models, this relationship is exact as the second derivative of the score function is equal to 0. That is, $\mathbf{W}_N(t, \hat{\boldsymbol{\theta}}) = \mathbf{W}_N(t, \boldsymbol{\theta}_0) - t\, \mathbf{W}_N(1, \boldsymbol{\theta}_0)$. Consequently,

$$\mathbf{W}_N(t, \hat{\boldsymbol{\theta}}) \rightarrow_d \mathbf{Z}^0(t) = \mathbf{Z}(t) - t \cdot \mathbf{Z}(1)$$

The limit process $\mathbf{Z}^0(t)$ is a p-dimensional process with covariance function $\text{cov}[\mathbf{Z}^0(t), \mathbf{Z}^0(s)] = s(1 - t)J(\boldsymbol{\theta}_0)$ for $s < t$. We can go on to the construction of *canonical monitoring process* $\mathbf{M}_N(t, \hat{\boldsymbol{\theta}})$, and under H_0,

$$\mathbf{M}_N(t, \hat{\boldsymbol{\theta}}) = \hat{\mathbf{J}}^{-1/2}\mathbf{W}_N(t, \hat{\boldsymbol{\theta}}) \rightarrow_d \mathbf{W}^0(t)$$

where $\mathbf{W}^0(t) = (W_1^0(t), \cdots, W_p^0(t))$ is a vector with p independent standard Brownian Bridges as component processes. In other words, kth component of $\mathbf{M}_N(t, \hat{\boldsymbol{\theta}})$ is distributed as a standard Brownian Bridge, $W^0(t)$. That is,

$$M_N(t, \hat{\theta}_k) \rightarrow_d W^0(t)$$

The above weak convergence continues to hold for any 'reasonable' functionals (including supremum) of $M_N(t, \hat{\theta}_k)$ [see e.g. 8, pp 509, Theorem 1]. At this point, Hjort and Koning [13] proposed several functionals of $M_N(t, \hat{\theta}_k)$ as possible test statistics and suggested to approximate their distribution functions through simulation for comparison purpose. For example, they stated

$$\max_{0 \leq t \leq 1} ||M_N(t, \hat{\theta}_k)||^2 \rightarrow_d \max_{0 \leq t \leq 1} ||W^0(t)||^2$$

and suggested to use $\max_{0 \leq t \leq 1} ||M_N(t, \hat{\theta}_k)||^2$ as test statistic. Instead we propose to use D_k as defined below as a test statistic

$$D_k \equiv \max_{0 \leq t \leq 1} |M_N(t, \hat{\theta}_k)| = \max_{1 \leq j \leq N-1} |M_N(t, \hat{\theta}_k)| \rightarrow_d \max_{0 \leq t \leq 1} |W^0(t)| \equiv D \qquad (3.1)$$

The primary reason for preferring $\max_{0 \leq t \leq 1} |M_N(t, \hat{\theta}_k)|$ over the $\max_{0 \leq t \leq 1} ||M_N(t, \hat{\theta}_k)||^2$ is that the limiting distribution of the former is known. The use of D_k as a test statistic eliminates the additional simulation work approximating the limiting distribution and thus making the testing process much more computationally efficient. The resulting reduction in the computation time is significant in the context of regression tree construction with the longitudinal data. D has distribution function [3]

$$F_D(x) = 1 + 2 \sum_{l=1}^{\infty} (-1)^l \exp\left(-2\, l^2 x^2\right).$$

7

Although this expression involves an infinite series, this series converges very rapidly. Usually a few terms suffice for very high accuracy. This result can be used to formulate a test for instability of parameters at α level of significance as follows: (1) Calculate the value of the process D_k for each parameter $k = 1, \cdots, p$ and obtain the raw p-values. (2) Adjust the p-values according to a chosen multiple testing procedure. (3) Reject H_0 if the adjusted p-value for any of the processes, D_k, is less than α.

3.3 Instability test for multiple partitioning variables

The testing strategy discussed in Sections 3.1 and 3.2 for a single partitioning variable depends only on the predictor variable type, either categorical or continuous. However, in practice, we expect to have more than one partitioning variable. Let there be S partitioning variables: $\{X_1^{G_1}, \cdots, X_S^{G_S}\}$. In that case we need to perform the instability test for each of the partitioning variables $X_1^{G_1}, \cdots, X_S^{G_S}$ subject to adjustment for multiplicity of type I errors. Let the p-values after multiplicity adjustment be $p_1, \cdots, p_S$, respectively and $p_{min} = \min\{p_1, \cdots, p_S\}$. Candidate partitioning variable with the smallest p-value (p_{min}) is chosen as a partitioning variable if p_{min} is smaller than the nominal significance level. For further discussion please see Section 4.

3.4 Power under the alternative hypothesis

We consider the following form of Pitman's local alternatives in the vicinity of H_0

$$\boldsymbol{\theta}_{(g)} = \boldsymbol{\theta}_0 + \boldsymbol{\delta} \circ \mathbf{h}\left(\frac{c_{(g)}}{c_{(G)}}\right)\frac{1}{\sqrt{N}} + O\left(\frac{1}{N}\right) \tag{3.2}$$

where $\boldsymbol{\delta} = (\delta_1, \cdots, \delta_p)^\top$ is the vector containing degrees of departure from the null hypothesis and $\mathbf{h} = (h_1, \cdots, h_p)^\top$ is the vector containing magnitudes of departure. The operation $\circ$ denotes the point-wise multiplication, i.e.,

$$\boldsymbol{\delta} \circ \mathbf{h}\left(\frac{c_{(g)}}{c_{(G)}}\right) = \left[\delta_1 h_1\left(\frac{c_{(g)}}{c_{(G)}}\right), \cdots, \delta_p h_p\left(\frac{c_{(g)}}{c_{(G)}}\right)\right]^\top$$

Theorem 3.1. *Under* (3.2), *the limiting distribution for the* χ^2_{cat} *is a non-central chi-square distribution*

$$\chi^2_{cat} \longrightarrow_d \chi'^2\left[(G-1)p, \ \sum_{g=1}^{G}\lambda_g^2\right]$$

where

$$\lambda_g = \mathbf{J} \cdot m_g \mathbf{h}\left(\frac{c_{(g)}}{c_{(G)}}\right) \cdot \frac{1}{\sqrt{N}}$$

Theorem 3.2. *Under* (3.2), *the limiting distribution for the canonical monitoring process is as follows*

$$\mathbf{M}_N(t, \hat{\boldsymbol{\theta}}) \longrightarrow_d \mathbf{J}^{1/2} \cdot t_g \cdot \boldsymbol{\delta} \circ (\bar{\mathbf{h}}_g - \bar{\mathbf{h}}) + \mathbf{W}^0(t) \qquad t \in [t_g, t_{g+1})$$

where,

$$\bar{\mathbf{h}}_g = \frac{1}{M_g}\sum_{j=1}^{g} m_j \mathbf{h}\left(\frac{c_{(j)}}{c_{(G)}}\right) \qquad \bar{\mathbf{h}} = \bar{\mathbf{h}}_G$$

8

Proofs of these theorems are provided in the Appendix. Briefly, we first approximate the density function using Taylor series expansion and then proceed in the way analogous to the one discussed in Section 3.2.

4   Longitudinal Regression Tree

4.1   Algorithm

Smaller p-values from the instability test indicate greater amount of evidence towards instability. Intuitively, splits in the tree should be based on the partitioning variable that shows higher evidence towards instability of the parameters. Therefore, we propose the following algorithm in order to construct a regression tree for longitudinal data.

Step 1. Perform the instability test for each partitioning variable separately at a prespecified level of significance α. The level of significance for performing instability test is subject to adjustment for multiple comparisons in order to maintain the level of type I error at each non-terminal node.

Step 2. Stop if no partitioning variable is significant at level α. Otherwise, choose the partitioning variable with the smallest p-value and proceed to step 3.

Step 3.   Consider all cut-off points of the chosen partitioning variable. At each cut-off point, calculate the improvement in the goodness of fit criterion (e.g., deviance). With X^G as the chosen partitioning variable, the improvement in goodness of fit criterion can be obtained at the cut-off point $c_{(g)}$ in the following steps:

    a. Split the data in two parts. One group will include the observations from subjects with $X^G \leq c_{(g)}$ and the other group will have the observations from subjects with $X^G > c_{(g)}$.

    b. Fit the longitudinal model on (i) all the individuals in the node, (ii) the individuals with $X^G \leq c_{(g)}$ and (iii) the individuals with $X^G > c_{(g)}$. Calculate the goodness of fit criterion from each of these three models. Call them as GOF_{all}, GOF_I and GOF_{II}, respectively.

    c. Calculate the improvement in goodness of fit criterion as $\text{GOF}_I + \text{GOF}_{II} - \text{GOF}_{all}$.

Step 4.   Choose the cut-off value that provides the maximum improvement in goodness of fit criterion and use this cut-off for binary splitting.

Step 5. Follow the Steps 1-4 for each non-terminal node.

The above strategy for construction of regression tree with longitudinal data has two major advantages over the currently existing algorithms. First, the level of significance is maintained at each node. Second, there are huge savings in computation time as we are evaluating the improvement in selected goodness of fit criterion at the cut-off points of the chosen partitioning variable only.

9

4.2 Improvement

A measure of improvement due to regression tree can be provided in terms of likelihood function based criterion. For example, Akaike Information criterion (AIC) for a tree T can be obtained as

$$\mathrm{AIC}_T = 2 \sum_{k=1}^{|T|} l_k - 2 \cdot |T| \cdot p$$

where $|T|$ denotes the number of terminal nodes in T, l_k is the log-likelihood in kth terminal node and p is the number of estimated parameters in each node. If we denote the AIC obtained from the traditional linear mixed effect model at root node (that is, common parametric form for covariates and time for the entire population) by AIC_0, the improvement due to regression tree can be measured as

$$\text{Improvement } (T) = \mathrm{AIC}_T - \mathrm{AIC}_0$$

Since the overall model fitted to all the data is nested within the regression tree based model, a likelihood ratio test or test for deviance can be constructed as well to evaluate the overall significance of a given regression tree.

4.3 Pruning

The improvement in regression tree comes at a cost of adding complexity to the model. If we can summarize complexity of a tree by number of terminal nodes, the cost adjusted AIC of a regression tree T can be defined as follows

$$\mathrm{AIC}_T(\gamma) = \mathrm{AIC}_T - \gamma(|T| - 1), \quad \gamma > 0$$

where γ be the *average complexity* for each terminal node. As a result, the tree T will be selected if

$$\mathrm{AIC}_T - \gamma(|T| - 1) > \mathrm{AIC}_0$$

or

$$\gamma < \frac{\mathrm{AIC}_T - \mathrm{AIC}_0}{|T| - 1} \equiv \gamma_T \tag{4.1}$$

That is, the tree T will be chosen as long as γ_T does not exceed some pre-set level of *average complexity*, γ_0; otherwise, we have to prune the tree T to bring γ_T below γ_0.

5 Simulation

We have explored the performance of instability test for continuous partitioning variables and the performance of proposed regression tree algorithm as a whole through simulation studies. The first two simulation studies evaluate the performance of instability test with continuous partitioning variable (as discussed in section 3.2). The third simulation study is aimed to explore the performance of the regression tree algorithm in section 4.1.

5.1 Performance of instability test with continuous partitioning variable

Let X^G be continuous partitioning variable with ordered cut-off points as $c_{(1)} \leq \cdots \leq c_{(G)}$. We first investigated the size of the test and then obtained the size-corrected power.

Table 1: Summary of the test statistic D_k based on $10,000$ Monte-Carlo samples generated according to the model 5.1. The values in the parenthesis below the α in the first column are the critical values from the limiting distribution and rejection is determined based on this critical value. The top row for each α level reflects the size of the test obtained by the Monte-Carlo simulation.

$\alpha(\%)$		N				
		50	100	200	500	1000
1.25	% of rejection	0.54	0.56	0.89	1.02	0.95
(1.5930)	$(1-\alpha)$100th percentile	1.4760	1.4938	1.5366	1.5643	1.5504
1.67	% of rejection	0.75	0.85	1.10	1.33	1.29
(1.5472)	$(1-\alpha)$100th percentile	1.4447	1.4532	1.4891	1.5147	1.4986
2.50	% of rejection	1.20	1.46	1.77	2.04	1.94
(1.4802)	$(1-\alpha)$100th percentile	1.3722	1.3998	1.4180	1.4392	1.4412
5.00	% of rejection	2.78	3.35	3.48	4.07	4.19
(1.3581)	$(1-\alpha)$100th percentile	1.2497	1.2924	1.2934	1.3154	1.3287
10.00	% of rejection	5.66	7.14	7.19	8.37	8.53
(1.2238)	$(1-\alpha)$100th percentile	1.1236	1.1585	1.1629	1.1901	1.1857
20.00	% of rejection	13.05	14.73	15.83	16.97	17.14
(1.0728)	$(1-\alpha)$100th percentile	0.9859	1.0045	1.0194	1.0350	1.0373

5.1.1 Size of the test

In order to examine the size of the test we have considered a longitudinal model with single mean parameter. We generated observations for N subjects at $t = 0, 1, 2, 3$ from the following model

$$X^G = c_{(g)}: \quad y_{it} = \beta_0 + b_i + \epsilon_{it} \tag{5.1}$$

with $\beta_0 = 2$, $b_i \sim N(0, 0.5^2)$ and $\epsilon_{it} \sim N(0, 0.2^2)$. The observations for X^G were generated for each simulation separately from uniform(0,300). For each N, $10,000$ Monte-Carlo samples were generated and the test statistic D_k (see Eq. (3.1)) was calculated for each sample separately. The null hypothesis of parameter stability is rejected at $\alpha\%$ level of significance when D_k exceeds the $(1-\alpha) \times 100$th percentile of the limiting distribution.

The observed percentiles and the percentage of rejected null hypotheses are summarized in Table 1. We can make following observations: 1) the type I error of test does not exceed the nominal level, 2) the size of the test approaches to the desired significance level α with the increase in the sample size N, and 3) the test is under-sized for smaller sample sizes. The severe problem with the size of the test for smaller sample size can be explained as follows. Calculation of test statistic, D_k, involves σ^2 and $\mathbf{V}_i$. However, in practice, the true values of σ^2 and $\mathbf{V}_i$ are unknown and we replace them by their estimates. A consistent estimator (e.g. ML- or REML-based) approaches the true value with an increasing sample size. However, the estimates might be biased for smaller sample sizes. To be precise, for smaller sample size, σ^2 and $\mathbf{V}_i$ may remain underestimated and this leads to smaller value of D_k which in turn results in a smaller size of the test. However, bias in estimation of σ^2 and $\mathbf{V}_i$ fades away with the increase in N and this increases the size of the test. We observe this trend in Table 1 as the size of test approaches the nominal level of type I error with the increase in sample size. However, the size of test remains smaller than nominal level even for the reasonably large N.

11

Table 2: Estimated power (%) of instability test with continuous partitioning variable obtained in the simulation described in Section 5.1. Figures corresponding to β_1 and β_0 represent the rejection percentage of H_0 associated with β_1 and β_0, respectively. The 'Overall' figures represent the percentage of at least one rejection out of the two.

N			δ					
		0	0.25(−0.25)	0.50(−0.50)	0.75(−0.75)	1.00(−1.00)	1.2(−1.2)
50	β_1	1.4	1.4(1.4)	1.6(1.6)	1.9(1.9)	2.3(2.3)	2.4(2.3)
	β_0	1.6	4.4(4.3)	16.9(16.6)	41.9(42.0)	70.2(70.6)	86.9(87.0)
	Overall	2.9	5.6(5.5)	17.9(17.6)	42.6(42.5)	70.5(70.8)	87.0(87.1)
100	β_1	1.5	1.6(1.6)	2.0(2.1)	2.5(2.6)	3.0(3.0)	3.2(3.2)
	β_0	1.7	(5.2(5.3)	18.7(19.7)	44.4(46.0)	72.9(73.9)	88.9(89.0)
	Overall	3.1	6.6(6.7)	19.8(20.8)	45.0(46.6)	73.1(74.2)	89.0(89.1)
200	β_1	1.8	1.9(1.8)	2.2(2.2)	2.7(2.7)	3.3(3.3)	3.5(3.4)
	β_0	1.9	5.6(5.3)	20.7(19.8)	47.5(46.8)	75.7(75.2)	90.1(89.8)
	Overall	3.6	7.4(6.8)	21.9(21.0)	48.2(47.4)	76.0(75.4)	90.6(89.9)
500	β_1	2.1	2.1(2.2)	2.7(2.5)	3.2(3.2)	3.6(3.7)	3.9(4.0)
	β_0	1.8	6.1(6.0)	21.4(20.1)	48.1(48.2)	76.6(76.6)	91.1(91.1)
	Overall	3.7	7.8(7.8)	22.8(22.2)	48.8(49.1)	77.0(77.0)	91.3(91.2)

The reduced size has been also reported in other tests based on the Brownian Bridge process. For example, Kolmogorov Smirnov test for normality (which also uses the Brownian Bridge as limiting distribution) is conservative [21, 23, 4]. As N exceeds 500, the size of the test is close to the nominal level of significance. As a remedy for smaller sample sizes, one might consider using a liberal α level or small sample distribution for D_k obtained through simulation.

5.1.2 Power

We generate observations for N subjects at $t = 0, 1, 2, 3$ from the following model to evaluate performance of instability test for X^G

$$X^G = c_{(g)}: \quad y_{it} = \beta_{0(g)} + \beta_{1(g)}t + b_i + \epsilon_{it},$$

$$\beta_{0(g)} = \beta_0 \qquad \beta_{1(g)} = \beta_1 + \delta \cdot \frac{c_{(g)}}{c_{(G)}}$$

We set $\beta_0 = 1$ and $\beta_1 = 2$. b_i, ϵ_{it} and X^G were generated similarly as before in Section 5.1.1. In this simulation, the parameter β_1 is not stable unless $\delta = 0$. We deal with two parameters: β_0 and β_1, thus we will have two Brownian bridge processes. We adjusted the p-values according to the Hochberg's step-up procedure [14]. We have chosen Hochberg's step-up procedure because it is relatively less conservative than the Bonferroni procedure [15]. However, in principle, any multiple comparison procedure can be applied here.

The results based on 10,000 simulation are displayed in Table 2. As the absolute value of δ deviates from zero, the power increases. The power of test is close to 80% and approaching the 90% mark as $|\delta| > 1$. The sign of δ does not influence the power of the test. Sizes of the test are very much in
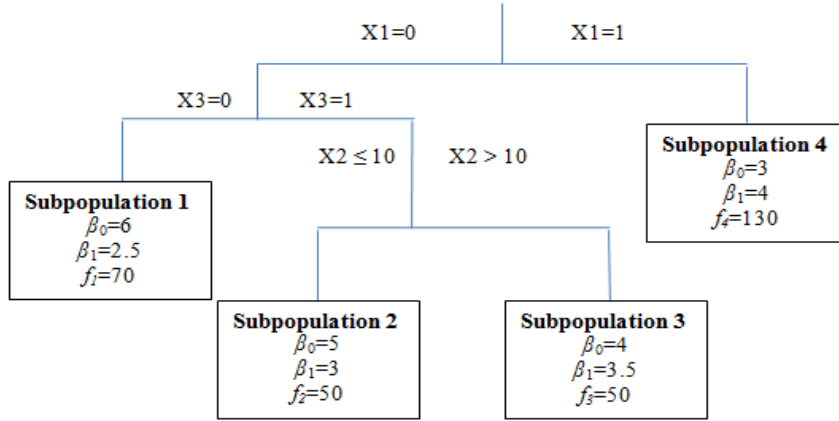
12

Figure 2: True tree structure for the simulation described in section 5.2

agreement with the first simulation study. As observed previously, the test is mildly conservative in the current simulation scenario as the observed level of type I error is consistently slightly below the nominal value $\alpha = 0.05$.

5.2 Performance of regression tree for longitudinal data

In this simulation, our goal is to assess the improvement in estimation due to regression tree algorithm when the population under consideration is truly heterogeneous. We have simulated observations for $N = 300$ subjects and these subjects come from one of the four different subgroups. Description of these subgroups is displayed in the form of a tree structure in Figure 2. The subgroups can be defined in terms of the partitioning variables X_1, X_2 and X_3. In rth subgroup ($r = 1, \cdots, 4$), the values for continuous response variable y were generated at $t = 0, 1, 2, 3$ according to following model:

$$y_{it} = \beta_{0k} + \beta_{1k} t + b_i + \epsilon_{it}; \quad i = 1, \cdots, f_r,$$

where $b_i \sim N(0, 4)$ and $\epsilon_{it} \sim N(0, 1)$. As displayed in Figure 2, the true values of β_1 were set at 2.5, 3.0, 3.5 and 4.0 and for β_0, the true values were set at 6, 5, 4 and 3, for the four subgroups, respectively. Further, observations were generated for $f_1 = 70$ individuals in subgroup 1, $f_2 = 50$ individuals in subgroup 2, $f_3 = 50$ individuals in subgroup 3, and $f_4 = 130$ individuals in subgroup 4. In order to study the performance of our algorithm constructing the longitudinal regression tree, we calculated the mean absolute deviation (MAD) in β_0 and β_1 in rth subgroup for each simulation as defined below

$$\text{MAD}(\hat{\beta}_{0r}) = \frac{1}{f_r} \sum_{j \in S_r} |\beta_{0r} - \hat{\beta}_{0j}| \qquad \text{MAD}(\hat{\beta}_{1r}) = \frac{1}{f_r} \sum_{j \in S_r} |\beta_{1r} - \hat{\beta}_{1j}|,$$

where β_{0r} and β_{1r} are the true values of β_0 and β_1 in the rth subgroup and $\hat{\beta}_{0j}$ and $\hat{\beta}_{1j}$ are the corresponding estimates for the jth individual applyig longitudinal tree and then fitting mixed model in each subgroup. S_r is the set of indices for all individuals in the rth subgroup while f_r denotes their number.

13

Table 3: Description of the mixed models used in section 5.2 for the comparison with regression tree algorithm (Model 1). All models include random intercepts to account for the subject-specific effects.

	Predictors
Model 2	t
Model 3	t, X_1, X_2, X_3
Model 4	t, X_1, X_2, X_3, X_1X_2, X_1X_3, X_2X_3
Model 5	t, X_1, X_2, X_3, X_1X_2, X_1X_3, X_2X_3, $X_1X_2X_3$
Model 6	t, X_1, X_2, X_3, tX_1, tX_2, tX_3
Model 7	t, X_1, X_2, X_3, X_1X_2, X_1X_3, X_2X_3, tX_1, tX_2, tX_3, tX_1X_2, tX_1X_3, tX_2X_3
Model 8	t, X_1, X_2, X_3, X_1X_2, X_1X_3, X_2X_3, $X_1X_2X_3$, tX_1, tX_2, tX_3, tX_1X_2, tX_1X_3, tX_2X_3, $tX_1X_2X_3$

Table 4: Summary of the results for the simulation described in section 5.2

	Par.	Subpop 1		Subpop 2		Subpop 3		Subpop 4		Overall	
		mad0	mad1	mad0	mad1	mad0	mad1	mad0	mad1	mad0	mad1
Model 1	8*	0.291	0.083	0.309	0.093	0.315	0.088	0.159	0.033	0.241	0.064
Model 2	2	1.802	0.899	0.802	0.399	0.204	0.101	1.198	0.601	1.107	0.554
Model 3	5	1.439	0.899	0.615	0.399	0.320	0.101	0.893	0.601	0.879	0.554
Model 4	8	1.345	0.899	0.641	0.399	0.277	0.101	0.895	0.601	0.855	0.554
Model 5	9	1.345	0.899	0.628	0.399	0.281	0.101	0.896	0.601	0.854	0.554
Model 6	8	0.404	0.185	0.202	0.061	0.585	0.292	0.432	0.200	0.413	0.189
Model 7	14	0.286	0.070	0.306	0.114	0.300	0.105	0.291	0.074	0.294	0.085
Model 8	16	0.294	1.822	0.309	1.225	0.299	2.181	7.089	2.256	3.242	1.970

mad0= Average MAD($\hat{\beta}_0$), mad1= Average MAD($\hat{\beta}_1$), Par.: No.of parameters.

Model 1: Subgroups are extracted using regression tree algorithm and mixed model with time slope and random intercept fitted separately in each subgroup.

Model 2 - 8: Description is given in Table 3

$\star$ - In Model 1, 81% of time regression tree with 4 subgroups were extracted.

14

The simulation results are summarized in Table 4 based on 1000 simulations in each case. In each simulation, regression tree was constructed with the following specifications: (1) the overall significance level of instability test was set at 5%, (2) minimum node size for further split was set at 40, and (3) minimum terminal node size was set at 20. Recall that we are considering four subgroups in the current simulation. The tree algorithm extracted exactly four subgroups in 81% of the cases and five subgroups in 16% of the cases. Only in 1.6% instances the tree algorithm extracted six subgroups and there were only 1.3% instances when three subgroups were extracted. For the comparison purposes, we considered seven linear mixed models (Model 2 - Model 8). These models are described in Table 3. The application of the longitudinal regression tree algorithm (Model 1) shows comparatively larger improvements in the estimation of the coefficients in all four subgroups. Both the $\mathrm{MAD}(\hat{\beta}_0)$ and $\mathrm{MAD}(\hat{\beta}_1)$ are considerably smaller in Model 1 compared to the Models 2—8. The improvement in estimation of coefficients in regression tree is attributed to its ability to extract homogeneous subgroups and then fitting mixed model separately within each group. On the contrary, Models 2—8 assume either additive (Models 2—3) or an interaction (Models 4—8) mixed effect model for the entire population assuming parametric form for both covariates and time. These models do not capture the complexity for the heterogenous subgroups and overestimate it for the homogenous subgroups.

Inclusion of the interaction terms in the model does not necessarily take into account subgroup heterogeneity in the presence of continuous partitioning variable. For example, in Models 4 and 5 common slope is assumed for the entire population, but include interaction terms in the baseline effect; still, the absolute deviation in estimating β_0 is almost 2.5 times higher compared to that of in regression tree. Similarly, the Models 6 – 8 include interaction terms for both baseline and longitudinal effects, but again the absolute deviations in estimating β_0 and β_1 are higher compared to what we have obtained with the longitudinal regression tree.

Model 6 including the interaction terms with t and the partitioning variables is probably the most commonly used model in practice. However, the application of the regression tree algorithm offers a considerable improvement in the estimation compared to Model 6. Models 6 – 7 provide some improvement over regression tree in some of the subgroups. However, these improvements are comparatively rare and largely influenced by the fact how the subgroups are defined. We would close this section pointing out, apart from providing improvement in estimation, the regression tree algorithm also identifies the meaningful subgroups defined by the partitioning variables which would remain unidentified otherwise.

6 Application

We applied the longitudinal regression tree algorithm to study the changes in relative concentration of glutamine and glutamate (Glx) relative to creatine (Cr) in a longitudinal study of late stage HIV patients. Glx have been reported as an important metabolite for maintenance and promotion of basic cell function such as protein synthesis, function of different organs including kidney, liver and intestine and also plays important role in inter-organ transportation [e.g., see 26]. We considered a total of $\sum_{i=1}^{N} n_i = 757$ observations from $N = 231$ subjects. The longitudinal observations for each subject were within 4 years from baseline. The number of observations per subject ranged from 2 to 7 with median number of observations equal to 3.
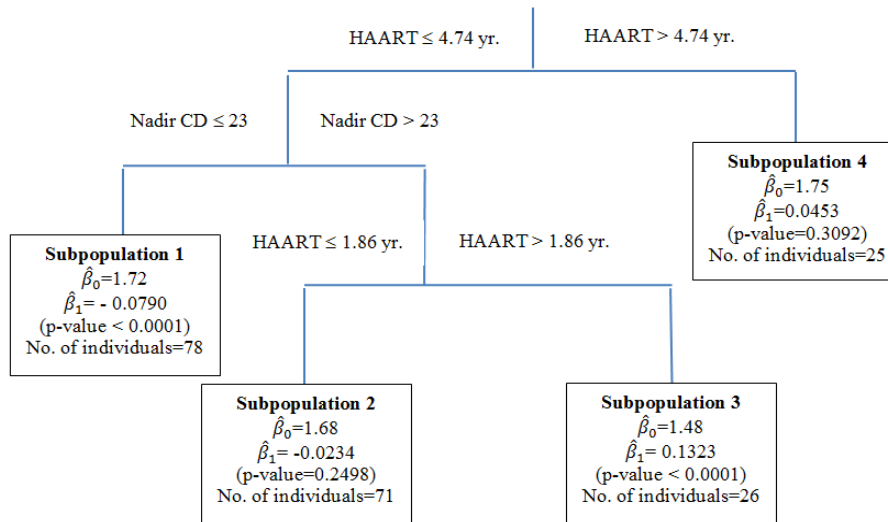
15

Figure 3: Regression tree obtained from concentration of glutamine and glutamate (Glx) relative to creatine (Cr) data along with the estimated parameters in each node as discussed in section 6. The p-value in each node corresponds to the estimate of the slope β_1.

For the construction of regression tree we used baseline measurements of several clinical and demographic variables including sex, race, education, age, CD4 count, nadir CD4 count, duration of HIV, years of highly active antiretroviral therapy (HAART), plasma HIV RNA count, antiretroviral CNS penetration-effectiveness (CPE) score and AIDS dementia complex (ADC) stage as partitioning variables. In each node we consider fitting the following model separately

$$y_{it} = \beta_0 + \beta_1 t + b_i + \epsilon_{it}$$

where y_{it} indicates the measurement of the relative concentration of Glx to Cr (Glx/Cr) from the ith individual at time t (in years) and b_i is the subject specific intercept. Regression tree was constructed with the following specifications: (1) the significance level for individual instability test was set to 10%, (2) the minimum node size for further split was set to 50, and (3) the minimum terminal node size was set to 25. Figure 3 displays the estimated longitudinal regression tree with the estimates of β_0 and β_1 for each terminal node or subpopulation. We chose larger level of significance over the standard 5% level due to our simulation studies finding showing that the proposed instability test is conservative for smaller sample sizes.

The regression tree algorithm extracted 4 subgroups and these subgroups can be characterized by the number of "Years on HAART therapy" and the "Nadir CD4 count". Improvement in deviance due to application of regression tree algorithm was 332 (log-likelihoods were -390 vs. -556; with 6 degrees of freedom). Figure 3 shows that individuals in Subgroup 3 characterized by HAART therapy for prolonged period (≥ 1.86 years) in combination with nadir CD4 count greater than 23 experienced significant increase (p-value< 0.0001) in the relative concentration of Glx over time. We also observe positive slope in Subpopulation 4, but it was smaller than that of in Subpopulation 3 and was not significantly different from zero. On the contrary, patients with a combination of the smaller nadir CD4 count and fewer years of HAART therapy in Subgroup 1 experience significant decrease (p-value< 0.0001) in the Glx/Cr over time. Increase in nadir CD4 count had positive effect on the
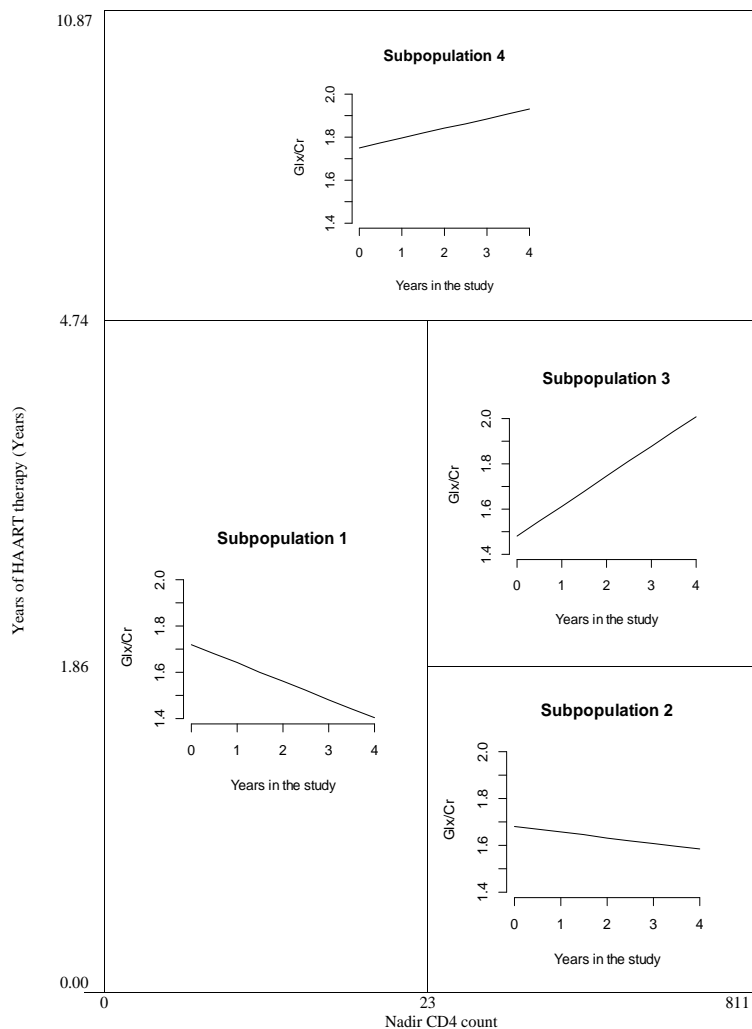
16

Figure 4: Visual representation of the fitted longitudinal profiles in each subgroup as estimated by the proposed longitudinal regression tree algorithm (see Figure 3). Separate panels show the estimated longitudinal profiles for the subgroups defined by the cutoff points for the "Nadir CD4 counts" (x-axis) and the cutoff points for the "Years of HAART therapy" (y-axis).

17

progression of Glx/Cr, however, Glx/Cr continued to decrease as long as number of HAART therapy remained less than 1.86 years. In summary, we have shown that prolonged period of HAART therapy is the key factor that influences increase of Glx/Cr over time and greater nadir CD4 count gives boost to this process. All these interpretable subgroups along with a significant improvement in overall model fit show that the underlying population is heterogenous. Thus considering a traditional linear mixed effect model for the entire population is not defensible. Figure 4 displays the longitudinal profiles in each of the 4 subgroups.

7 Discussion

In this paper we proposed an algorithm for the construction of regression tree with longitudinal data. The superiority of our method over the other similar algorithm is two folds. First, it maintains the type I error at each splitting through performing a single instability test. Second, it reduces the computation time substantially as we first choose the partitioning variable and then evaluate the goodness of fit criterion at all cut-off points of the selected partitioning variable only. In this way, we avoid the calculation of the goodness of fit criterion at the cut-off points of the partitioning variables other than the chosen one. Our interest to construct regression tree from longitudinal data is different from latent class modeling (LCM) in two aspects. Unlike LCM, the number of subgroups is not known a priori in our case. Moreover, the subgroups determined by the tree algorithm can be defined by the partitioning variables which is not the case with LCM.

The homogeneity of the longitudinal trajectories in the population under consideration is not always justifiable. However, this is the basic assumption underlying the traditional linear mixed effects model fitting. For example, in an observational study where patients are recruited at different stages of the disease progression, it would be very hard to believe that all the patients enroll in the study with similar baseline values for the disease marker or have similar rates of progression. This might be true even for the controlled studies (e.g., clinical trials) with diverse population characteristics. In such cases, the proposed regression tree algorithm can be useful to explore the diversity in a population in terms of disease progression trajectories which would be otherwise remain uncovered. There is a plethora of evidence for the heterogeneity of longitudinal profiles; for example Leuchter *et al.* [19] reported heterogeneity in progression of depression in a double-blind randomized trial. Other reported examples include heterogeneous trend in aggressive behavior among different classes of students [17, 16], differential math achievement among different dropout groups [24], and varying age-crime curve among different birth cohorts [22].

The most common strategy to incorporate the effect of baseline attributes in a traditional linear mixed effect model is to include these attributes and their interaction terms as covariates in the model. However, feasibility of this approach is limited to the situations where (1) there is a limited number of baseline attributes, (2) the functional form of the attributes to be included in the model is known and (3) the effect of continuous attributes is linear. On the contrary, the proposed longitudinal tree algorithm enables us to discover the most parsimonious model for the homogenous subgroups in a population without any functional form restriction. For example, in 5.2, the true model requires only 8 parameters to be estimated whereas the traditional mixed effect model considering all possible interactions of the covariates (Model 8) requires 16 parameters to be estimated and yet performs poorly compared to the Longitudinal tree (Model 1) (see Tables 4 and 3)

18

Both the instability test and the tree algorithm discussed in this paper are based on the score process. This increases the utility of the proposed method beyond the application to the mixed effects longitudinal models studied in this paper. We can apply a similar algorithm in other scenarios as long as we can obtain (or approximate) an expression for the score function and the Hessian matrix in a tractable form. For example, we can apply our method in the generalized linear mixed effects model (GLMM) where score function is difficult to obtain, but can be approximated. With the binary response it would be analogous to the construction of a classification tree with the longitudinal data. Another extension, we currently work on is in the context of regression tree construction with multiple response variables, both in cross-sectional and longitudinal setting.

One of the drawbacks of the proposed method is an underestimation of the nominal test size, especially for the small sample sizes. As already mentioned in Section 5.1.1, this finding is consistent with other score type tests that use Brownian Bridge as limiting process. One way to address this issue is by increasing the nominal type I error level. A more principled approach to address this problem would be to find the exact distribution through a simulation study. As an follow-up work, it would be interesting to compare the results of the parameter instability test for continuous partitioning variable (and, regression tree in general) between the exact and the limiting distributions. We end our conclusions by discussing the possibility of sup-Wald type test [e.g. see 2] as an alternative to the score test. In general, Wald test has higher power compared to score test [11], however, the former is often criticised for not maintaining the type I error. Further, we are not aware of any result on the convergence of the test statistic distribution used in sup-Wald type tests. Unavailability of limiting distribution for sup-Wald type test makes it infeasible to use in construction of a longitudinal tree.

A Proofs

A.1 Proof of Theorem 3.1

Proof. Using Taylor series expansion we can write

$$f(\mathbf{y}, \boldsymbol{\theta}_{(g)}) \doteq f(\mathbf{y}, \boldsymbol{\theta}_0) \left\{ 1 + \mathbf{u}(\mathbf{y}, \boldsymbol{\theta}_0)^\top \boldsymbol{\delta} \circ \mathbf{h}\Big(\frac{c_{(g)}}{c_{(G)}}\Big) \frac{1}{\sqrt{N}} \right\}$$

Consequently,

$$
\begin{aligned}
E_{\boldsymbol{\theta}_g}[\mathbf{u}(\mathbf{y}, \boldsymbol{\theta}_0)] &= \int u(\mathbf{y}, \boldsymbol{\theta}_0) f(\mathbf{y}, \boldsymbol{\theta}_{(g)}) dy = E_{\boldsymbol{\theta}_0}[\mathbf{u}(\mathbf{y}, \boldsymbol{\theta}_0)] + \mathbf{J} \cdot \boldsymbol{\delta} \circ \mathbf{h}\Big(\frac{c_{(g)}}{c_{(G)}}\Big) \frac{1}{\sqrt{N}} \\
&= \mathbf{J} \cdot \boldsymbol{\delta} \circ \mathbf{h}\Big(\frac{c_{(g)}}{c_{(G)}}\Big) \frac{1}{\sqrt{N}}
\end{aligned}
\tag{A.1}
$$

It can be shown that

$$\text{cov}_{H_1}[\mathbf{W}_N(t, \boldsymbol{\theta}_0)] = \text{cov}_{H_0}[\mathbf{W}_N(t, \boldsymbol{\theta}_0)] + O\left(\frac{1}{N}\right) \doteq \mathbf{J} \tag{A.2}$$

Proof of Theorem 3.1 follows from the definition of non-central chi-square distribution.

$\square$

19

A.2 Proof of Theorem 3.2

Proof. Using (A.1) and (A.2),

$$E_{H_1}[\mathbf{W}_N(t, \boldsymbol{\theta}_0)] = \mathbf{J} \frac{1}{N} \sum_{i=1}^{M_g} \delta \circ \mathbf{h}\left(\frac{c_{(g)}}{c_{(G)}}\right) = \mathbf{J} \cdot t_g \cdot \boldsymbol{\delta} \circ \bar{\mathbf{h}}_g \qquad t \in [t_g, t_{g+1})$$

This time using the FCLT along with Cramer-Wold device we can show that

$$\mathbf{W}_N(t, \boldsymbol{\theta}_0) \longrightarrow_d \mathbf{J} \cdot t_g \cdot \delta \circ \bar{\mathbf{h}}_g + \mathbf{Z}(t) \qquad t \in [t_g, t_{g+1})$$

Therefore, for $t \in [t_g, t_{g+1})$,

$$\mathbf{W}_N(t, \hat{\boldsymbol{\theta}}) = \mathbf{W}_N(t, \boldsymbol{\theta}_0) - t_g\, \mathbf{W}_N(1, \boldsymbol{\theta}_0) + o_p(1) \longrightarrow_d \mathbf{J} \cdot t_g \cdot \delta \circ (\bar{\mathbf{h}}_g - \bar{\mathbf{h}}) + \{\mathbf{Z}(t) - t \cdot \mathbf{Z}(1)\}$$

Thus under H_1,

$$\mathbf{M}_N(t, \hat{\boldsymbol{\theta}}) = \hat{\mathbf{h}}^{-1/2}\mathbf{W}_N(t, \hat{\boldsymbol{\theta}}) \longrightarrow_d \mathbf{J}^{1/2} \cdot t_g \cdot \boldsymbol{\delta} \circ (\bar{\mathbf{h}}_g - \bar{\mathbf{h}}) + \mathbf{W}^0(t) \qquad t \in [t_g, t_{g+1})$$

$\square$

Acknowledgements

References

[1] Abdolell, M. and LeBlanc, M. and Stephens, D. and Harrison, R. (2002). Binary partitioning for continuous longitudinal data: categorizing a prognostic variable. *Statistics in medicine*, **21(22),** 3395–3409.

[2] Andrews, D. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrica: Journal of the Econometric Society*, 821–856.

[3] Billingsley, P. (2009) Convergence of probability measures, volume 493. Wiley-Interscience.

[4] Birnbaum, Z. (1952). Numerical tabulation of the distribution of Kolmogorov's statistic for finite sample size. *Journal of the American Statistical Association*, **47(259),** 425–441.

[5] Breiman, L. and Friedman, J. and Stone, C. and Olshen, R. (1984) Classification and regression trees, Chapman & Hall/CRC.

[6] Brown, R. and Durbin, J. and Evans, J. (1975). Techniques for testing the constancy of regression relationships over time. *Journal of the Royal Statistical Society. Series B (Methodological)*, 149–192.

20

[7] Clark, L. and Pregibon, D. (1992) ., "Tree-based models" in Statistical Models in S, eds. Chambers J. and Hastie T., New York, NY: Chapman and Hall.

[8] CsõRGõ, M. (2002). A glimpse of the impact of pál erd# ous on probability and statistics. *Canadian Journal of Statistics*, **30(4),** 493–556.

[9] Demidenko, E. (2004) Mixed models: theory and applications, volume 518, Wiley-Interscience.

[10] Diggle, P. and Heagerty, P. and Liang, K. and Zeger, S. (2002) Analysis of longitudinal data, volume 25, Oxford University Press, USA.

[11] Engle, R. (1981) "Wald, Likelihood Ratio, and Lagrange Mul- tiplier Tests in Econometrics" in Handbook of Econometrics, eds. Griliches Z. and Intriligator M., Amsterdam: North Holland.

[12] Galimberti, G. and Montanari, A. (2002). Regression trees for longitudinal data with time-dependent covariates. *Classification, clustering and data analysis*, 391–398.

[13] Hjort, N. and Koning, A. (2002). Tests for constancy of model parameters over time. *Journal of Nonparametric Statistics*, **14(1-2),** 113–132.

[14] Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, **75(4),** 800–802.

[15] Hochberg, Y. and Tamhane, A. (1987). Multiple comparison procedures, *John Wiley & Sons, Inc.*.

[16] Ialongo, N. and Werthamer, L., Kellam, S., and Brown, C., and Wang, S. and Lin, Y (1999). Proximal impact of two first-grade preventive interventions on the early risk behaviors for later substance abuse, depression, and antisocial behavior. *American journal of community psychology*, **27(5),** 599–641.

[17] Kellam, S., Rebok, G., Ialongo, N. and Mayer, L. (1994). The course and malleability of aggressive behavior from early first grade into middle school: Results of a developmental epidemiologically-based preventive trial. *Journal of Child Psychology and Psychiatry*, **35(2),**259–281.

[18] Laird, N. and Ware, J. (1982). Random-effects models for longitudinal data. *Biometrics*, **38,** 963–974.

[19] Leuchter, A., Cook, I., Witte, E., Morgan, M. and Abrams, M. (2002). Changes in brain function of depressed subjects during treatment with placebo. *American Journal of Psychiatry*, **159(1),** 122–129.

[20] Liang, K.Y. and Zeger, S. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73(1),** 13–22.

[21] Lilliefors, H. (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, **62(318),** 399–402.

[22] Loughran, T. and Nagin, D. (2006). Finite sample effects in group-based trajectory models. *Sociological methods & research*, **35(2),** 250–278.

21

[23] Massey Jr, F. (1951). The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American statistical Association*, **46(253),** 68–78.

[24] Muthén, B. (2004). Latent variable analysis: The handbook of quantitative methodology for the social sciences (D. Kaplan ed.), *Sage Publications*, 345–68.

[25] Muthén, B. and Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics*, **55(2),** 463–469.

[26] Newsholme, P. and Procopio, J. and Lima, M. and Pithon-Curi, T. and Curi, R. (2003) Glutamine and glutamatetheir central role in cell metabolism and function. *Cell biochemistry and function*, **21(1),** 1–9.

[27] Nyblom, J. (1989). Testing for the constancy of parameters over time. *Journal of the American Statistical Association*, **84(405),** 223–230.

[28] Raudenbush, S. (2001). Comparing personal trajectories and drawing causal inferences from longitudinal data. *Annual review of psychology*, **52(1),** 501–525.

[29] Segal, M. (1992). Tree-structured methods for longitudinal data. *Journal of the American Statistical Association*, **87(418),** 407–418.

[30] Sela, R. and Simonoff, J. (2012). RE-EM trees: a data mining approach for longitudinal and clustered data. *Machine learning*, **86(2),** 169–207.

[31] Verbyla, D. (1987). Classification trees: a new discrimination tool. *Canadian Journal of Forest Research*, **17(9),** 1150–1152.

[32] Zeileis, A. and Hothorn, T. and Hornik, K. (2010). party with the mob: Model-based Recursive Partitioning in R. *Relation*, **10(1.15),** 4593.

[33] Zeileis, A. and Hothorn, T. and Hornik, K. (2008). Numerical tabulation of the distribution of Kolmogorov's statistic for finite sample size. *Journal of Computational and Graphical Statistics*, **17(2),** 492–514.

[34] Zhang, H. (1997). Multivariate adaptive splines for analysis of longitudinal data. *Journal of Computational and Graphical Statistics*, **87(418),** 74–91.

[35] Zhang, H. and Singer, B. (1999). Recursive partitioning in the health sciences, *Springer Verlag.*