Accepted Manuscript

Unbiased regression trees for longitudinal and clustered data

Wei Fu, Jeffrey S. Simonoff

PII: S0167-9473(15)00043-2
DOI: http://dx.doi.org/10.1016/j.csda.2015.02.004
Reference: COMSTA 6035

To appear in: *Computational Statistics and Data Analysis*

Received date: 12 February 2014
Revised date: 8 December 2014
Accepted date: 6 February 2015

Please cite this article as: Fu, W., Simonoff, J.S., Unbiased regression trees for longitudinal and clustered data. *Computational Statistics and Data Analysis* (2015), http://dx.doi.org/10.1016/j.csda.2015.02.004

Unbiased Regression Trees for Longitudinal and Clustered Data

Wei Fu[a], Jeffrey S. Simonoff[a,*]

[a]*New York University Leonard N. Stern School of Business, NY, USA*

Abstract

A new version of the RE-EM regression tree method for longitudinal and clustered data is presented. The RE-EM tree is a methodology that combines the structure of mixed effects models for longitudinal and clustered data with the flexibility of tree-based estimation methods. The RE-EM tree is less sensitive to parametric assumptions and provides improved predictive power compared to linear models with random effects and regression trees without random effects. The previously-suggested methodology used the CART tree algorithm for tree building, and therefore that RE-EM regression tree method inherits the tendency of CART to split on variables with more possible split points at the expense of those with fewer split points. A revised version of the RE-EM regression tree corrects for this bias by using the conditional inference tree as the underlying tree algorithm instead of CART. Simulation studies show that the new version is indeed unbiased, and has several improvements over the original RE-EM regression tree in terms of prediction accuracy and the ability to recover the correct tree structure.

Keywords:
clustered data, longitudinal data, mixed models, regression tree

1. Introduction

The regression tree is a nonparametric method for estimating a regression function. Assume the data set consists of a response variable y and one or more predictor (covariate) variables $\mathbf{X} = (X_1, X_2, ..., X_k)$. The regression tree algorithm splits the data set into subsets based on the values of its covariate variables $\mathbf{X}$. The process is repeated on each derived subset recursively until halted based on some stopping rule. One common approach is to split until the subset at a node has all the same value of the response variable y or predictor

*Corresponding author; 44 West 4th Street, New York NY USA 10012-1126; telephone 1-212-998-0452; fax 1-212-995-4003

Email addresses: wfu@stern.nyu.edu (Wei Fu), jsimonof@stern.nyu.edu (Jeffrey S. Simonoff)