Regression tree-based diagnostics for linear multilevel models

Jeffrey S. Simonoff
New York University

May 11, 2011

Longitudinal and clustered data

Panel or longitudinal data, in which we observe many individuals over multiple periods, offers a particularly rich opportunity for understanding and prediction, as we observe the different paths that a variable might take across individuals. Clustered data, where observations have a nested structure, also reflect this hierarchical character. Such data, often on a large scale, are seen in many applications:

- test scores of students over time
- test scores of students across classes, teachers, or schools
- blood levels of patients over time
- transactions by individual customers over time
- tracking of purchases of individual products over time

Longitudinal data

I will refer to such data as longitudinal data here, but all of the content applies equally to other clustered data. The analysis of longitudinal data is especially rewarding with large amounts of data, as this allows the fitting of complex or highly structured functional forms to the data.

We observe a panel of *individuals* $i = 1, ..., I$ at times $t = 1, ..., T_i$. A single observation period for an individual (i, t) is termed an *observation*; for each observation, we observe a vector of covariates, $\mathbf{x}_{it} = (x_{it1}, ..., x_{itK})'$, and a response, y_{it}.

Longitudinal data models

Because we observe each individual multiple times, we may find that the individuals differ in systematic ways; e.g., y may tend to be higher for all observation periods for individual i than for other individuals with the same covariate values because of characteristics of that individual that do not depend on the covariates. This pattern can be represented by an "effect" specific to each individual (for example, an individual-specific intercept) that shifts all predicted values for individual i up by a fixed amount:

$$y_{it} = Z_{it}\mathbf{b}_i + f(x_{it1}, \ldots, x_{itK}) + \varepsilon_{it}.$$

Mixed effects models

- If f is linear in the parameters and the b_i are taken as fixed or potentially correlated with the predictors, then this is a linear *fixed effects model*.
- If f is linear in the parameters and the b_i are assumed to be random (often Gaussian) and uncorrelated with the predictors, then the model is a linear *mixed effects model*.

Conceptually, random effects are appropriate when the observed set of individuals can be viewed as a sample from a large population of individuals, while fixed effects are appropriate when the observed set of individuals represents the only ones about which there is interest.

Linear model and goodness-of-fit

The most commonly-used choice of f is unsurprisingly the linear model

$$y_{it} = Z_{it}\mathbf{b}_i + X_{it}\boldsymbol{\beta} + \varepsilon_{it},$$

assuming errors ε that are normally distributed with constant variance. This model has the advantage of simplicity of interpretation, but as is always the case, if the assumptions of the model do not hold inferences drawn can be misleading. Such model violations include *nonlinearity* and *heteroscedasticity*. If specific violations are assumed, tests such as likelihood ratio tests can be constructed, but omnibus goodness-of-fit tests would be useful to help identify unspecified model violations.

Regression trees and goodness-of-fit

The idea discussed here is a simple one that has (perhaps) been underutilized through the years: since the errors are supposed to be unstructured if the model assumptions hold, examining the residuals using a method that looks for unspecified structure can be used to identify model violations. A natural method for this is a **regression tree**.

Miller (1996) proposed using a CART regression tree (Breiman, Friedman, Olshen, and Stone, 1984) for this purpose in the context of identifying unmodeled nonlinearity in linear least squares regression, terming it a **diagnostic tree**. They note that evidence for a signal left in the residuals (and hence a violation of assumptions) comes from a final tree that splits in the growing phase and is not ultimately pruned back to its root node.

Proposed method

Su, Tsai, and Wang (2009) altered this idea slightly by simultaneously including both linear and tree-based terms in one model, terming it an **augmented tree**, assessing whether the tree-based terms are deemed necessary in the joint model. They also note that building a diagnostic tree using squared residuals as a response can be used to test for heteroscedasticity.

We propose adapting the diagnostic tree idea to longitudinal/clustered data.

Proposed method

- Fit the linear mixed effects model.
- Fit an appropriate regression tree to the residuals from this model to explore nonlinearity.
- Fit an appropriate regression tree to the absolute residuals from the model to explore heteroscedasticity (squared residuals are more non-Gaussian and lead to poorer performance).

A final tree that splits from the root node rejects the null model.

Trees for longitudinal and clustered data

There has been a limited amount of work on adapting regression trees to longitudinal/clustered data. Segal (1992) and De'Ath (2002) proposed the use of multivariate regression trees in which the response variable was the vector $\mathbf{y}_i = (y_{i1}, ..., y_{iT})$. At each node, a vector of means, $\mu(g)$, is produced, where $\mu_t(g)$ is the estimated value for y_{it} at node g. Galimberti and Montanari (2002) and Lee (2005, 2006) proposed similar types of tree models. Unfortunately, these tree estimators have several weaknesses, including the inability to be used for the prediction of future periods for the same individuals.

Sela and Simonoff (2009) proposed a tree-based method that accounts for the longitudinal structure of the data while avoiding these difficulties.

“EM”-type algorithm

Consider again a general mixed effects model

$$y_{it} = Z_{it}\mathbf{b}_i + f(x_{it1}, ..., x_{itK}) + \varepsilon_{it}.$$

If the random effects, $\mathbf{b}_i$, were known, the model implies that we could fit a regression tree to $y_{it} - Z_{it}\mathbf{b}_i$ to estimate f via a tree structure. If the fixed effects, f, were known, then we could estimate the random effects using a traditional random effects linear model with fixed effects corresponding to the fitted values, $f(x_i)$. This alternation between the estimation of different parameters is reminiscent of (although is not) the EM algorithm, as used by Laird and Ware (1982); for this reason, we call the resulting estimator a Random Effects/EM Tree, or **RE-EM Tree**.

Estimation of a RE-EM Tree

1. Initialize the estimated random effects, $\hat{\mathbf{b}}_i$, to zero.
2. Iterate through the following steps until the estimated random effects, $\hat{\mathbf{b}}_i$, converge:
    2.1 Estimate a regression tree approximating f, based on the target variable, $y_{it} - Z_{it}\hat{\mathbf{b}}_i$, and predictors, $\mathbf{x}_{it\cdot} = (x_{it1}, ..., x_{itK})$, for $i = 1, ..., I$ and $t = 1, ..., T_i$. The tree is originally overgrown, and then pruned back using the one-SE rule of Breiman et al. (1984). Use this regression tree to create a set of indicator variables, $I(\mathbf{x}_{it\cdot} \in g_p)$, where g_p ranges over all of the terminal nodes in the tree.

    2.2 Fit the linear random effects model, $y_{it} = Z_{it}\mathbf{b}_i + I(\mathbf{x}_{it\cdot} \in g_p)\mu_p + \varepsilon_{it}$ using ML or REML. Extract $\hat{\mathbf{b}}_i$ from the estimated model using the Empirical Bayes estimates.

Estimation of a RE-EM Tree

This algorithm has several advantages over other approaches.

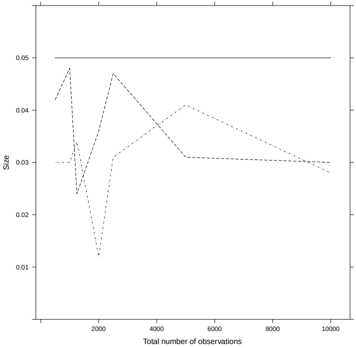
- The fitting of the regression tree uses built-in methods for missing data.
- Different numbers of time points for different individuals are easily handled, as is prediction of response values for future time points.
- The fixed effects portion of the model can be based on time-varying or nonvarying predictors.
- The fitting of the random effects portion of the model can be based on either independence within individuals, or a specified autocorrelation structure, thus allowing for complex correlation structure within individuals.
- Multilevel hierarchies are easily handled.

Structure of simulations

We use limited Monte Carlo simulations to investigate the properties of the method. We examine number of individuals I ranging from 50 to 200 with number of time points T ranging from 10 to 100 (implying number of observations $I \times T$ ranging from 500 to 20,000). Simulations show that properties are driven by the number of observations, not I or T separately. The null linear model is based on 5 normally-distributed predictors with mean 10 and standard deviation 1, $\beta' = (1, 2, -3, 4, -5)$, with the null model including 5 additional predictors with zero slopes; $\sigma_{\varepsilon}^2 = \sigma_b^2 = 1$.

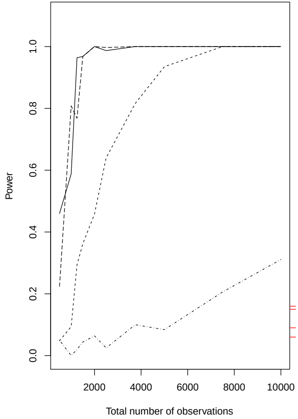
Size of tests

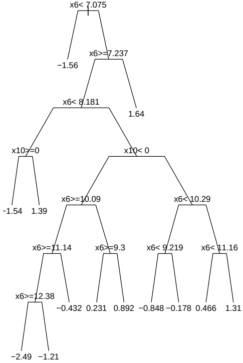
Even though the growing/pruning rules for the tree are not designed to directly control Type I error, it turns out that they do at a roughly .05 level, resulting in a generally conservative test.



Different slopes

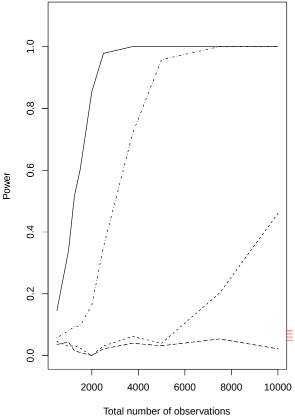
$$E(y) = E_0(y) \pm_{x_{10}} \alpha x_6, \alpha = .25(.25)1$$

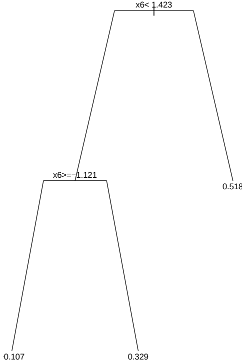




Quadratic term

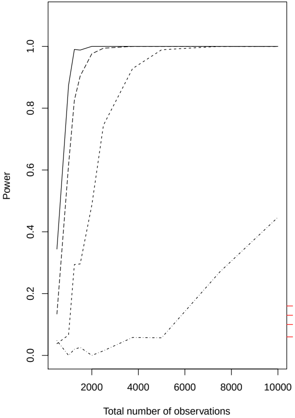
$$E(y) = E_0(y) \pm \alpha x_6^2, \alpha = .05(.05).2, E(x_6) = 0$$

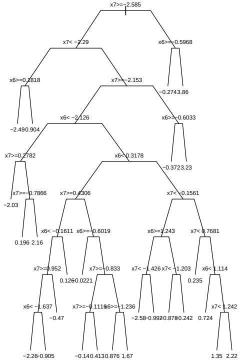




Product term

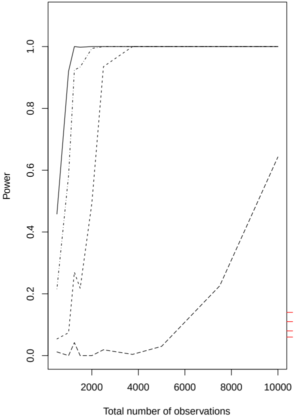
$$E(y) = E_0(y) \pm \alpha x_6 x_7, \alpha = .25(.25)1, E(x_6) = E(x_7) = 0$$

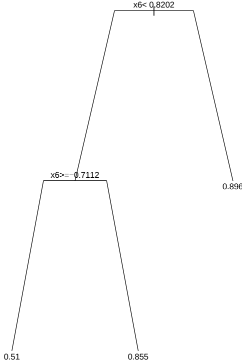




Heteroscedasticity related to nonpredictor

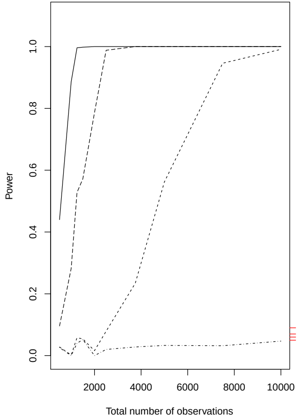
$$\sigma_y^2 = |x_6|^\alpha, \alpha = .125(.125).5$$

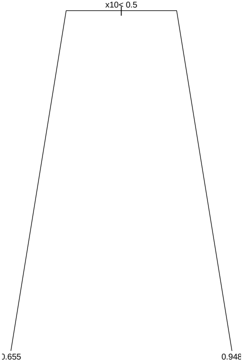




Heteroscedasticity related to subgroups

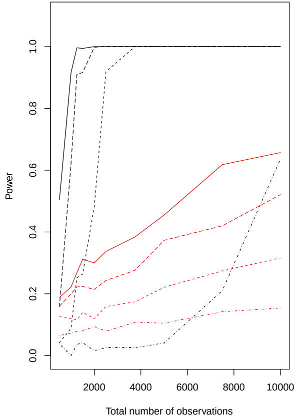
$$\sigma_y^2 = 1 \pm_{x_{10}} \alpha, \alpha = .0625(.0625).25$$

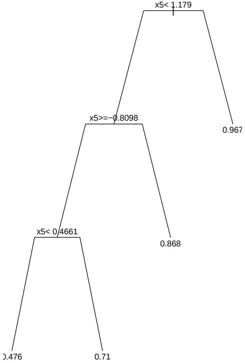




Heteroscedasticity related to predictor

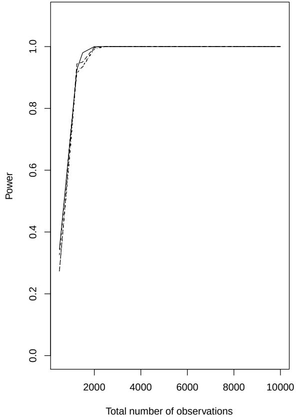
$$\sigma_y^2 = |x_5|^\alpha, \alpha = .125(.125).5\, E(x_5) = 0$$

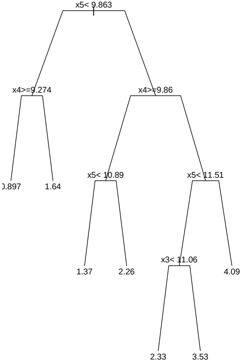




Heteroscedasticity related to expected response

$$\sigma_y^2 = |E(y)|^{\alpha}, \alpha = .0625(.0625).25$$



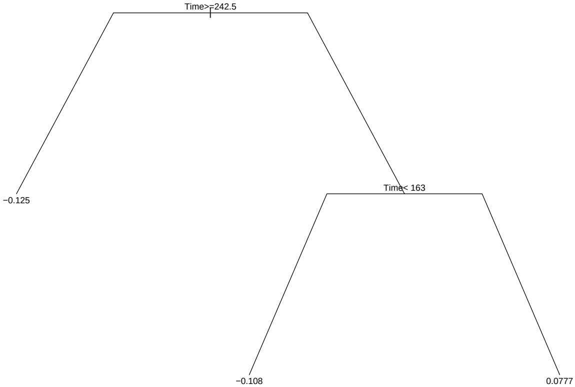


Spruce tree growth

Diggle, Liang, and Zeger (1994) and Venables and Ripley (2002) discuss a longitudinal growth study. The response is the log-size of 79 Sitka spruce trees, two-thirds of which were grown in ozone-enriched chambers, measured at five time points.

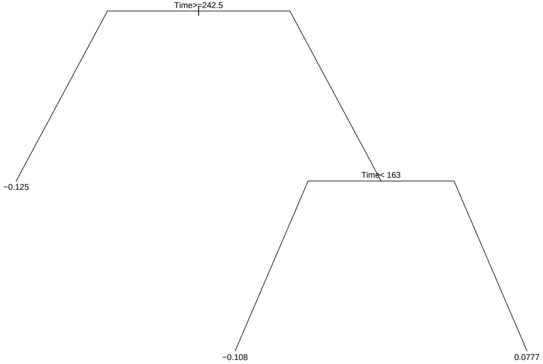
First, a linear model based on treatment status and time is fit, but the tree-based nonlinearity test indicates lack of fit related to time.

Test of fit of linear model



Test of fit of different slopes model

A natural alternative model is one allowing for different slopes for the treatment and control groups, but that does not correct the lack of fit.



Treating time as categorical

As the diagnostic trees suggest, the problem is in the linear formulation of the effect of time. If time is treated as a categorical predictor, the apparent lack of fit disappears, as the diagnostic tree has no splits.

An additional interaction of the treatment and (categorical) time effects is statistically significant, but has higher *AIC* and *BIC* values than the additive model, reinforcing that from a practical point of view the fit of the simpler model is adequate.

Heteroscedasticity diagnostic trees for all models do not split.

Transaction data set

We also apply the diagnostic trees to a dataset on third-party sellers on Amazon Web Services aiming to predict the prices at which software titles are sold based on the characteristics of the competing sellers (Ghose, 2005; Sela and Simonoff, 2009).

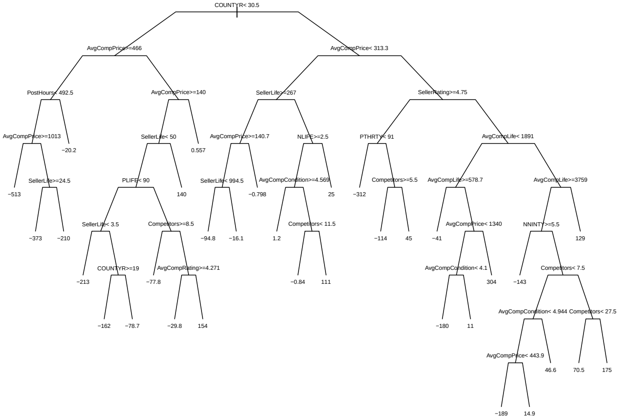
The data consist of 9484 transactions for 250 distinct software titles; thus, there are $I = 250$ individuals in the panel with a varying number of observations T_i per individual.

Transaction data set variables

- Target variable: the price premium that a seller can command (the difference between the price at which the good is sold and the average price of all of the competing goods in the marketplace). We also examine the log of this variable.
- Predictor variables
  - The seller's own reputation (total number of comments, the number of positive and negative comments received from buyers, the length of time that the seller has been in the marketplace)
  - The characteristics of its competitors (the number of competitors, the quality of competing products, and the average reputation of the competitors, and the average prices of the competing products).

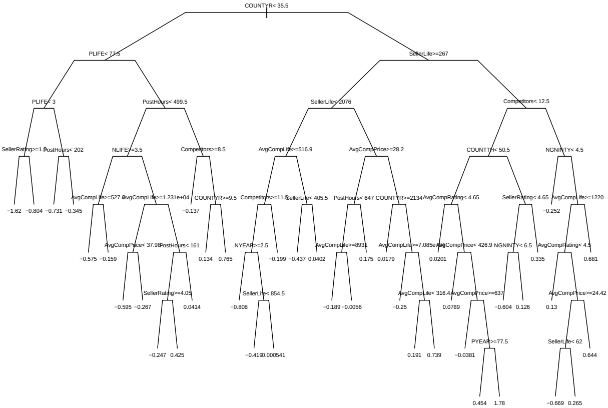
Test of fit of linear model

A linear model is clearly inadequate.



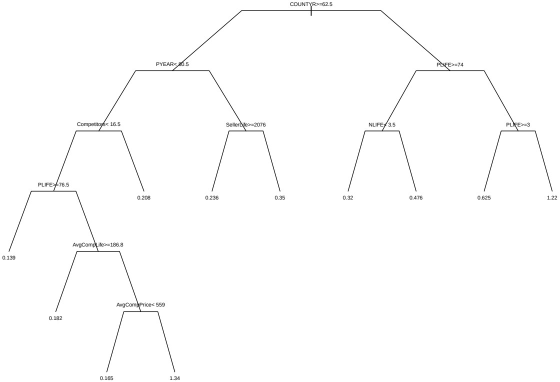
Test of fit of log-linear model

A log-linear model is also clearly inadequate.

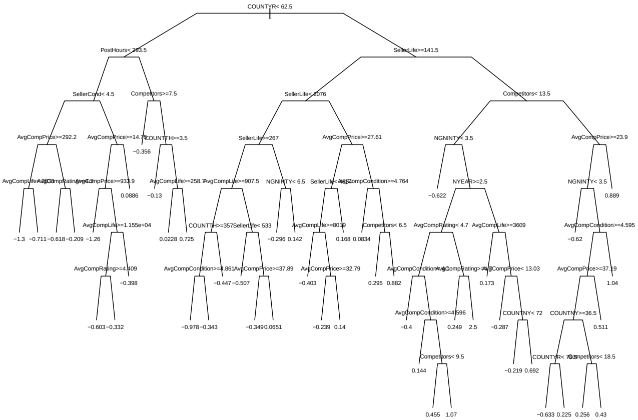


Heteroscedasticity test of log-linear model

There is apparent heteroscedasticity in the log-linear model, although we recognize that the lack of fit can affect this test.

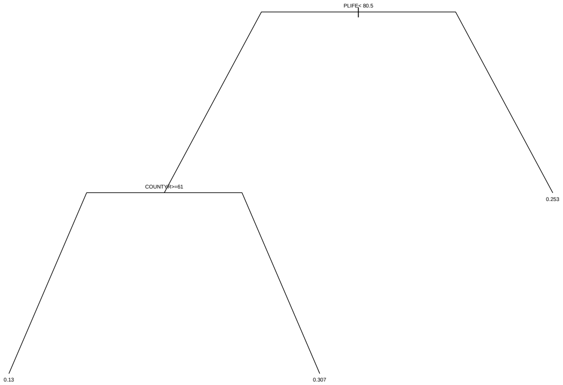


RE-EM tree for log price premium



Heteroscedasticity test of log price premium RE-EM tree

Heteroscedasticity is apparently much reduced when a tree model is used.



Conclusion

- Goodness-of-fit diagnostic trees can be constructed for longitudinal and clustered data based on the RE-EM tree idea.
- Versions to assess potential nonlinearity (based on residuals) and heteroscedasticity (based on absolute residuals) correspond to roughly .05 level tests, and demonstrate effective power for identifying different types of model violations.
- The diagnostic trees are not meant to replace examination of residuals or more focused (and powerful) tests of specific model violations; rather, they are an omnibus tool to add to the data analyst's toolkit to try to help identify unspecified mixed effects model violations.

Background information and R code

A paper describing the RE-EM tree method is available at `http://archive.nyu.edu/handle/2451/28094`.

The R package `REEMtree` used to construct RE-EM trees is available from CRAN (for versions of R starting with 2.12.2).