The Korean Journal of Applied Statistics (2011)
24*(5), 759–768*

DOI: http://dx.doi.org/10.5351/KJAS.2011.24.5.759

Tree-Structured Nonlinear Regression

Youngjae Chang[1] · Hyeonsoo Kim[2]

[1]Research Department, The Bank of Korea; [2]Research Department, The Bank of Korea

(Received June 2011; accepted August 2011)

Abstract

Tree algorithms have been widely developed for regression problems. One of the good features of a regression tree is the flexibility of fitting because it can correctly capture the nonlinearity of data well. Especially, data with sudden structural breaks such as the price of oil and exchange rates could be fitted well with a simple mixture of a few piecewise linear regression models. Now that split points are determined by chi-squared statistics related with residuals from fitting piecewise linear models and the split variable is chosen by an objective criterion, we can get a quite reasonable fitting result which goes in line with the visual interpretation of data. The piecewise linear regression by a regression tree can be used as a good fitting method, and can be applied to a dataset with much fluctuation.

Keywords: Regression tree, nonlinearity, piecewise regression, GUIDE.

1. Introduction

Many tree algorithms have been studied for decision theory or model fitting. A regression tree is a decision tree algorithm applied for regression problems. It is built through a process known as binary recursive partitioning. This is an iterative process of splitting data into partitions, and then splitting them further on each of the branches. A piecewise linear regression tree is a kind of regression tree. The piecewise linear regression tree discovers the relationships between a response variable and predictor variables at each node fitting a linear model that eventually uses the relationships to make predictions based on the mixture of the models at terminal nodes. Regression trees can be regarded as conditional regression methods since the model is fitted under the condition of predictors at each node. The condition is decided by splitting process at each node of the tree. The split point and split variables are chosen to get least sums of squared errors. Loh (2002) proposed a regression and classification tree algorithm, GUIDE(Generalized, Unbiased, Interaction Detection and Estimation (Loh, 2002)) which is a flexible regression tree method. The algorithm has little variable selection bias, and it can detect local interactions. GUIDE has split variable selection and split point search processes described above. GUIDE also has options to implement algorithms such as constant regression tree, quantile regression tree, multiple regression tree, and stepwise regression tree. All the regression models are built based on a tree structure and regarded as piecewise linear regression models. We use GUIDE to do our simulation and real data analysis. The results from

[1]Corresponding author: Economist, Research Department, The Bank of Korea, 110, Namdaemunno 3-ga, Chung-gu, Seoul 110-794, Republic of Korea. E-mail: yjchang@bok.or.kr

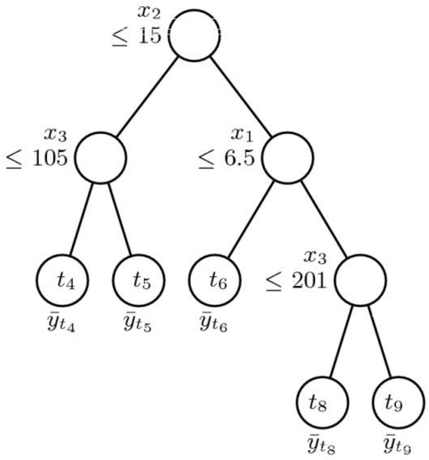

Figure 2.1. Example of a regression tree: At each intermediate node, a case goes to the left child node if and only if the condition is satisfied.

the analysis show that the piecewise linear regression by a regression tree can be used as a good fitting method for nonlinear regression problems.

This paper is organized as follows. In Section 2, a brief sketch of a regression tree algorithm and rationale for usefulness of regression tree as a nonlinearity measurement, followed by a simulation study and real data analysis. In Section 3, we conclude with summary of the results.

2. Regression Tree

A regression tree is a tree-structured solution in which a constant or a relatively simple regression model is fitted to the data in each partition as Loh (2008) pointed out. The regression tree is composed of nodes and branches which are related with split variables and split points. The tree grows with recursive partitioning.

Figure 2.1 shows an example of a regression tree, where the root node contains all the training observations, and the training data are recursively partitioned by values of the input variables until reaching the terminal nodes (t_4, t_5, t_6, t_8 and t_9) where the predictions are made.

Originally, the CART(Classification And Regression Tree) algorithm was proposed by Breiman *et al.* (1984); however, the algorithm showed variable selection bias when dealing with categorical variables of many categories (Loh, 2002). Strobl *et al.* (2007) also pointed out the bias in Random Forests importance measures, which is basically a bagging version of CART. GUIDE(Generalized, Unbiased, Interaction Detection and Estimation (Loh, 2002)) is an advanced flexible regression tree method. The algorithm has little variable selection bias, and it can detect local interactions. A recent study shows that the performance of GUIDE is quite good in terms of prediction accuracy among the many machine learning algorithms (Kim *et al.*, 2006). Chang (2010) applied the regression tree algorithm to analyze the factors that affect the Business Survey Index related with macroeconomic variables. Its prediction accuracy is comparable to Random Forests (Breiman, 2001) based on bagging approach (Breiman, 1996). We apply GUIDE to measure the nonlinearity of data because it has a procedure to tell whether data fit the linear function well or not. This can be possible because of a chi-squared test procedure. The chi-square statistic is the sum of the contributions from each of the individual cells. Every cell in a table contributes something to the overall chi-squared statistic. If a given cell differs markedly from the expected frequency, then the contribution of that cell to the overall chi-squared statistic is large. If a cell is close to the expected frequency for that cell, then the contribution of that cell to the overall chi-squared statistic is low. A large chi-squared statistic indicates that somewhere in the table, the observed frequencies differ markedly from the expected frequencies. In GUIDE, the chi-square test statistics are related with residuals from linear regression. Details will be introduced in the GUIDE algorithm.

2.1. Review of the GUIDE algorithm

Loh (2002) proposed the GUIDE algorithm, which has negligible selection bias and relatively low computational cost. GUIDE is also known as a smart data mining tool with flexible model fitting methods at each node. The procedure for fitting a stepwise linear regression model in GUIDE is as follows:

1. Let t denote the current node. Use stepwise regression that allows addition and deletion of variables. The default values of F-to-enter and F-to-remove are 4.00 and 3.99, respectively.
2. Do not split a node if the R^2 of the fitted model is greater than 0.99 or the current node has less than $2n_0$ observations, where n_0 is a previously specified number.
3. For each observation, define the class variable Z by the sign of its residual for each observation. That is, Define $Z = 1$ if the observation is associated with a positive residual. Otherwise, define $Z = 0$.
4. Construct a $2 \times m$ cross-classification table for each predictor variable X. The rows of the table are the values of Z, while the columns of the table are 4 intervals at the sample quartiles if X is a numerical variable ($m = 4$). If X is a categorical variable, its m distinct values form the columns of the table. Compute a p-value for the chi-squared test for each X based on the table.
5. In addition to the above "curvature" tests, perform chi-squared tests to detect interactions between pairs of same type variables (*i.e.*, numerical variable pairs, categorical variable pairs) or between pairs of different types of predictors. If a pair of variables from these interaction tests gives the smallest p-value, the split variable is one of two variables depending on the composition of the pairs.
6. Select the split variable X from the previous steps. Let t_L and t_R denote the left and right subnodes of t.
   - If X is a numerical variable, search for the split point that gives the lowest total of the sums of squared residuals in t_L and t_R, provided that the number of observations at each node is at least n_0 or user-specified value.
   - If X is a categorical variable, search for the split of the form $X \in C$, which gives the lowest weighted sum of the variances of Z in t_L and t_R, provided that the number of observations at each node is at least n_0. Here C is a subset of the values taken by X, and weights are proportional to sample sizes.
7. After splitting has stopped, prune the tree with a test sample or by cross-validation.

This algorithm is similar to the piecewise multiple linear regression tree algorithm, except for the model fitting step. There is little difference between the two methods for the simple linear regression

Table 2.1. Contingency table of signs of residuals for model (2.1)

Sign of residuals	$0.0 \sim 0.25$	$0.25 \sim 0.5$	$0.5 \sim 0.75$	$0.75 \sim 1.0$
$+$	4	2	3	3
$-$	7	4	4	3

problems because the variable selection does not play and important role in this case. Note that chi-squared tests are applied to test whether the data form a curvature structure or not in the Step 4 and Step 5. Following the test results, split process goes on or stops without further splitting. The last pruning procedure could make the results a little bit different, since it may remove the deepest splitted terminal nodes based on the cross-validation results.

2.2. Simulation study

We are now illustrating simple examples to show how the curvature test is implemented. The simulation study helps us understand the test procedure described in the previous section clearly. We can see that the procedure makes the nonlinear fitting more reliable.

We simulated a small dataset of 30 observations as follows.

$$y_i = \mu(x_i) + \epsilon_i, \quad i = 1, \ldots, n$$

with a mean function,

$$\mu(x_i) = x_i, \quad 0 \le x_i \le 1, \tag{2.1}$$

$$\mu(X_i) = \begin{cases} x_i, & 0 \le x_i < 0.25, \\ 2x_i, & 0.25 \le x_i < 0.75, \\ 10x_i, & 0.75 \le x_i \le 1, \end{cases} \tag{2.2}$$

where x_i's are $i.i.d$ from $U(0,1)$ with $\epsilon_i \sim N(0,1)$.

A scatter plot of simulation model (2.1) and a corresponding contingency table are presented as the following.

We see that the chi-squared test statistic for curvature test of the model (2.1) is 4.13 which is calculated in the following way.

$$\chi^2 = \frac{\sum(O_i - E_i)^2}{E_i}, \tag{2.3}$$

where O_i denotes the observed counts for each cell in the contingency table, E_i denotes the expected counts in the contingency table. The degrees of freedom is $(4-1)(2-1) = 3$ in this case. Note that the chi-squared statistic is lower than the critical value, which means we cannot reject the null hypothesis that data are equally distributed across the contingency table. We can regard the equal distribution of residuals as evidence of the linearity of the data we fit with a linear regression model.

A scatter plot of simulation model (2.2) and a corresponding contingency table follow. We can see that a linear model cannot fit the data well. The chi-squared statistic for the contingency table is relatively large. The chi-squared test statistic for curvature test of the model (2.2) is 8.93 which is greater than the critical value for chi-square distribution with degrees of freedom 3. It implies that we reject the null hypothesis that residuals are equally distributed along the fitted line.

We can regard this result as statistically significant existence of curvature structures in the data. We can also interpret the result as the existence of nonlinearity.

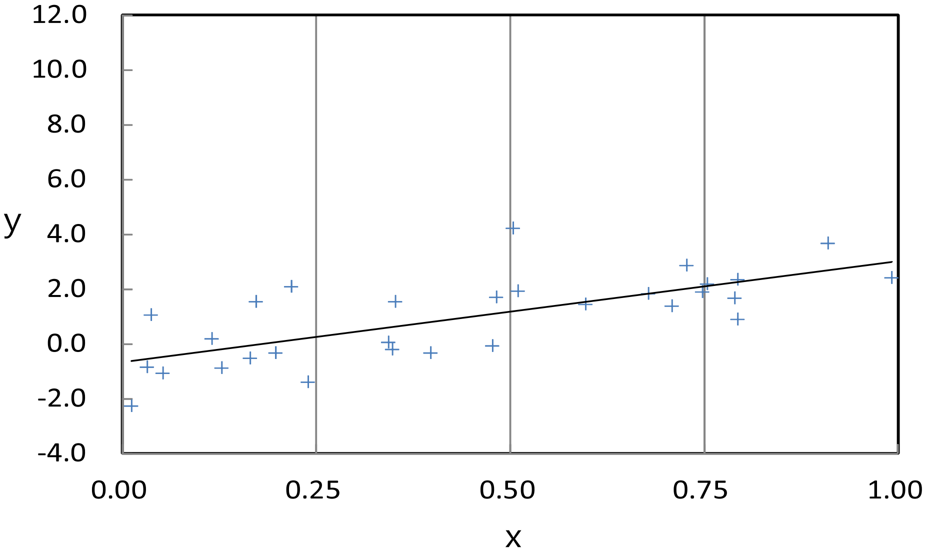

Figure 2.2. Scatter plot of simulation model (2.1)

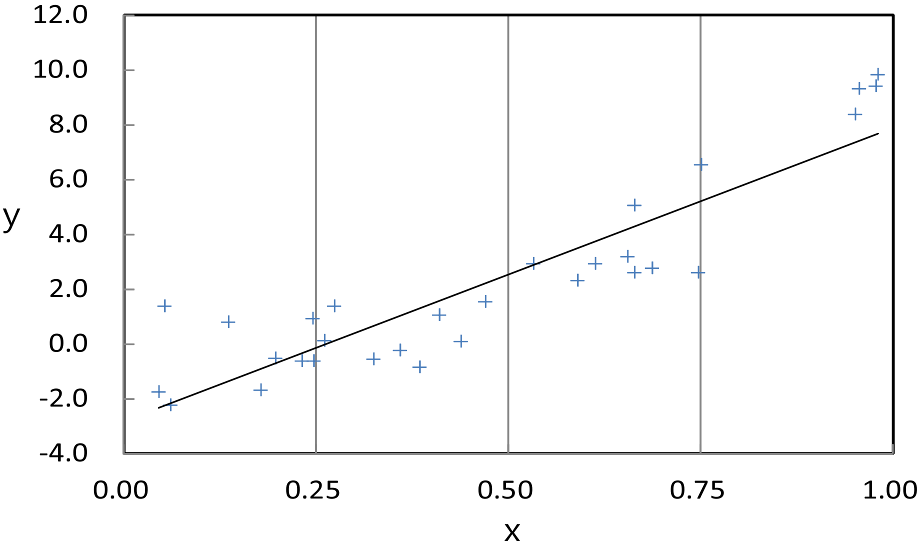

Figure 2.3. Scatter plot of simulation model (2.2)

We intentionally divide the domain three pieces and fit each linear model respectively. The results are presented in Figure 2.4 and Table 2.3. The piecewise regression model is presented with dotted lines. The chi-squared test statistic for the curvature test of the model (2.2) related with piecewise linear regression is 2.53, lower than 7.82, the critical value for chi-square distribution with degrees of freedom 3.

We also fit the data with a piecewise linear regression tree. We use GUIDE to grow a regression tree. Figure 2.5 shows a GUIDE piecewise linear least-squares regression tree model with stepwise

Table 2.2. Contingency table of signs of residuals for model (2.2)

Sign of residuals	0.0 ∼ 0.25	0.25 ∼ 0.5	0.5 ∼ 0.75	0.75 ∼ 1.0
+	5	2	2	5
−	4	6	6	0

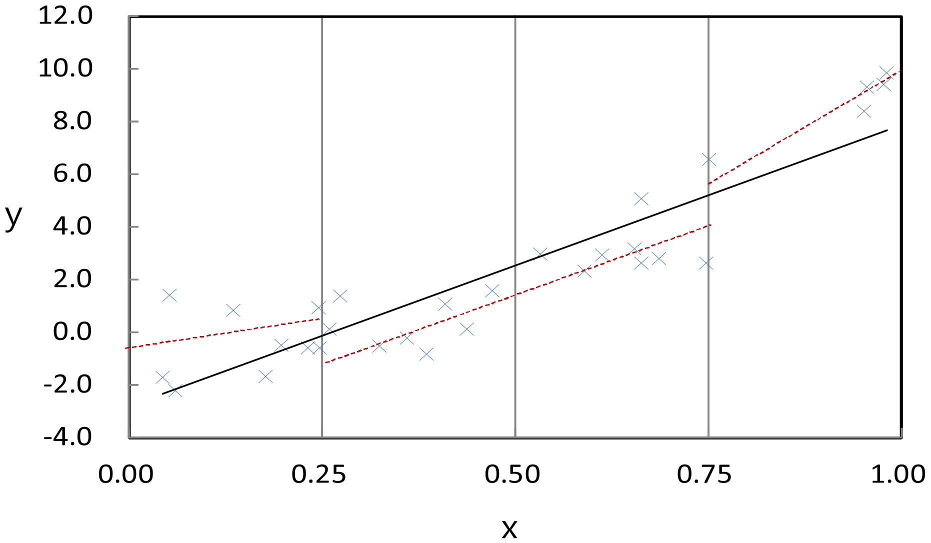

Figure 2.4. Scatter plot of simulation model (2.2) and corresponding piecewise linear model

Table 2.3. Contingency table of signs of residuals for model (2.2) with piecewise linear regression

Sign of residuals	0.0 ∼ 0.25	0.25 ∼ 0.5	0.5 ∼ 0.75	0.75 ∼ 1.0
+	3	4	4	3
−	6	4	4	2

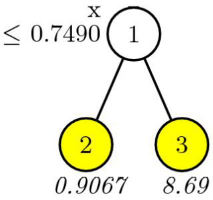

Figure 2.5. A piecewise linear regression tree for model (2.2)

variable selection. At each intermediate node, a case goes to the left child node if and only if the condition is satisfied. Number in italics beneath a leaf is the sample mean of y. The split point of the regression tree is 0.749, which is the precise split point of the simulation model. However, the split point of 0.25 is not identified in the tree. This is due to the error terms which make the difference of two slopes (1 and 2) indistinguishable. Error terms with standard deviation one may be too noisy in this case. Even if the tree does not detect the intended one split point, the regression tree looks reasonable.

We also did a simulation study with a little bit complicated data. We generate data based on sin

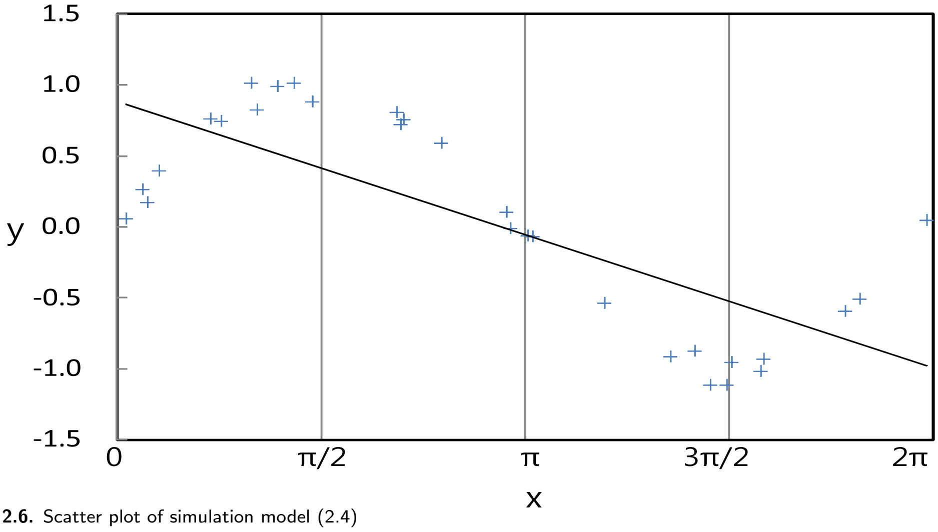

Figure 2.6. Scatter plot of simulation model (2.4)

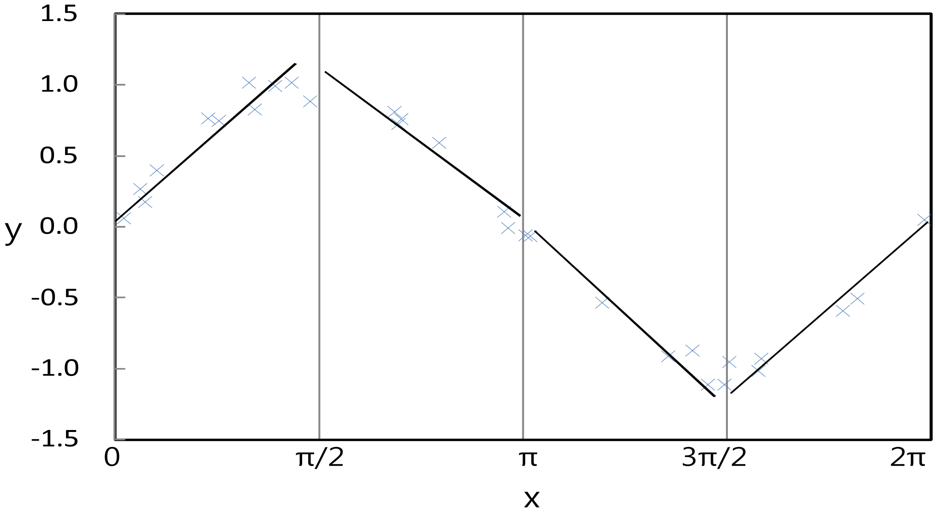

Figure 2.7. Scatter plot of simulation model (2.4) and corresponding piecewise linear model

function. That is,

$$\mu(x_i) = \sin(x_i), \quad 0 \leq x_i \leq 2\pi \tag{2.4}$$

where x_i's are $i.i.d$ from $U(0, 2\pi)$ with $\epsilon_i \sim N(0, 0.1)$.

We easily see that the data cannot be captured with a usual linear function. Figure 2.6 shows that a straight linear function does not fit the data well. However, we can fit the curve well with several piecewise linear functions if we split the domain adequately.

Figure 2.7 shows that four pieces of linear regression models fit the simulated sine data very well. The corresponding chi-squared test statistic of the Table 2.5 is relatively low(3.07) compared to the

Table 2.4. Contingency table of signs of residuals for model (2.4)

Sign of residuals	$0.0 \sim \pi/2$	$\pi/2 \sim \pi$	$\pi \sim 3/2\pi$	$3/2\pi \sim 2\pi$
$+$	7	6	2	3
$-$	4	0	5	3

Table 2.5. Contingency table of signs of residuals for model (2.4) with piecewise linear regression

Sign of residuals	$0.0 \sim \pi/2$	$\pi/2 \sim \pi$	$\pi \sim 3/2\pi$	$3/2\pi \sim 2\pi$
$+$	5	4	3	3
$-$	6	2	4	3

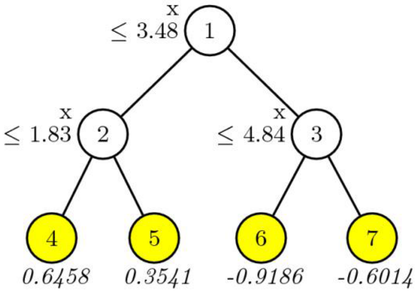

Figure 2.8. A piecewise linear regression tree for model (2.4)

previous case(9.47) of Table 2.4.

We fit the simulated data with a GUIDE piecewise linear regression tree. Figure 2.5 shows that there are three splits and four terminal nodes. The tree matches the curvature, or the nonlinear structure of the data very well. Note that the split points are 1.83, 3.48, and 4.84, which are close to $\pi/2$, π, and $3/2\pi$. A little difference is caused by the error term in the simulation model. Number in italics beneath a leaf is the sample mean of y.

2.3. Real data example for detecting structural break points

We apply our nonlinear fitting method for the real data. We consider the some price variables in the international commodity market showing significant fluctuation recently. Dubai oil price from January 1970 to June 2011 and CRB index from January 1994 to May 2011 are used. We are interested in the structural breaks of the time series data. It could be possible by a regression tree splitting the time span of the time series data. We can apply the piecewise regression tree method for this purpose, regrading the problem as a kind of linear regression with time being an independent variable.

If the tree can split the time span of economic time series data as it does in the simulation study, it would be helpful for the economic analysis reflecting the change of economic environment as time goes on.

Figure 2.11 and 2.12 are the regression trees of Dubai oil price and CRB index respectively. As expected, time spans are nicely splitted at the critical points of the real data. We may analyze the economic atmosphere or the relationships between macro-economic variables at the specific period determined by the regression tree. With the help of regression tree, we can split the time span objectively, and do analyses by the splitted pieces of the regression models.

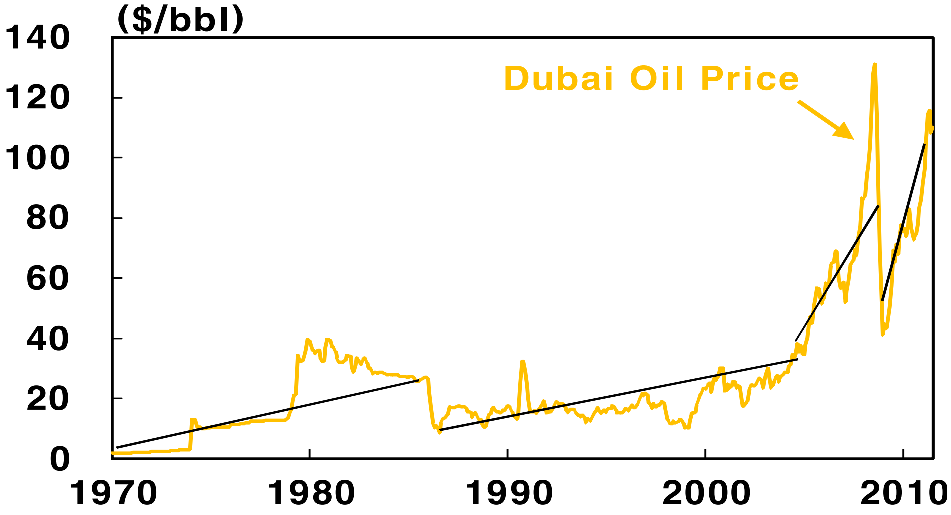

Figure 2.9. Dubai oil price

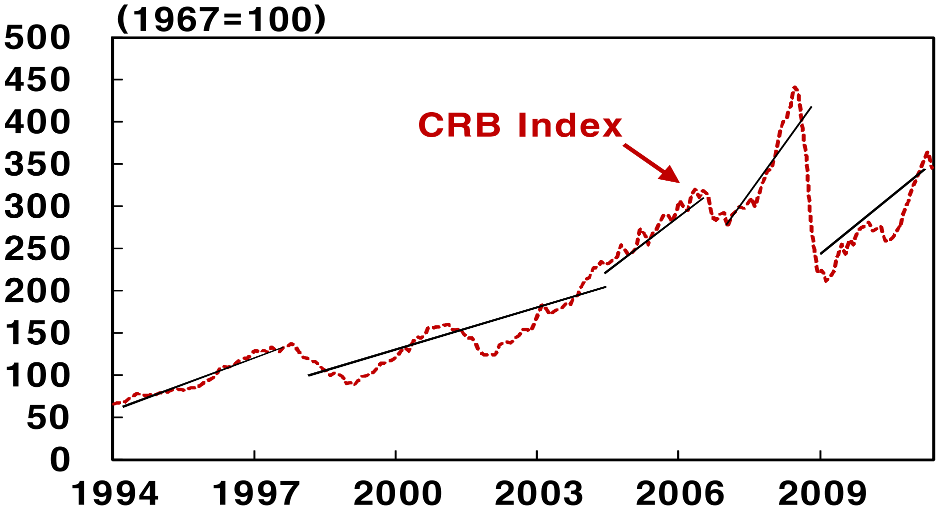

Figure 2.10. CRB index

3. Conclusion and Future Work

We did a simulation study to see how well the nonlinear regression could be implemented by a regression tree. The simulation study shows that structural changes are detected pretty well by piecewise linear regression trees. We also fit a couple of piecewise linear regression trees to real datasets. The trees capture the nonlinear patterns of the data by time very well. Based on our study, we can think about more useful application of regression trees. We can extend simple linear regression problems to multiple linear cases. That is to say, we can do a simulation study concerning the curvature structure with multiple independent variables and see whether the splits

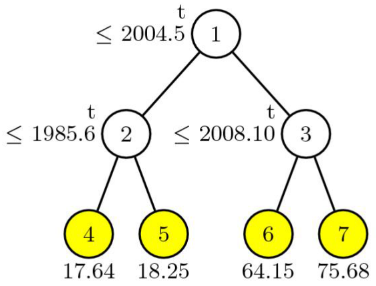

Figure 2.11. A piecewise linear regression tree for Dubai Oil Price

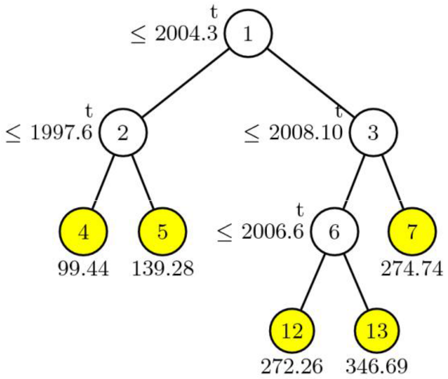

Figure 2.12. A piecewise linear regression tree for CRB index

are reasonably decided; in addition, it may be of interest to forecast problems with regression trees. If we can get the other independent variables which affect a dependent variable with some time lags, we can build a multiple regression model for the relationships between the variables that can be regarded as a forecasting tree. Due to the tree-structure, we can also do conditional forecasting based on the splits in the tree.

References

Breiman, L. (1996). Bagging predictors, *Machine Learning*, **24**, 123–140.

Breiman, L. (2001). Random Forests, *Machine Learning*, **45**, 5–32.

Breiman, L., Friedman, J., Stone, C. and Olshen, R. A. (1984). *Classification and Regression Trees*, 1st Edition, Chapman & Hall/CRC.

Chang, Y. (2010). The analysis of factors which affect business survey index using regression trees, *The Korean Journal of Applied Statistics*, **23**, 63–71.

Kim, H., Loh, W.-Y., Shih, Y.-S. and Chaudhuri, P. (2006). A visualizable and interpretable regression model with good prediction power, *IIE Transactions*, Special Issue on Data Mining and Web Mining.

Loh, W.-Y. (2002). Regression trees with unbiased variable selection and interaction detection, *Statistica Sinica*, **12**, 361–386.

Loh, W.-Y. (2008). Regression by parts: Fitting visually interpretable models with GUIDE, In *Handbook of Data Visualization*, C. Chen, W. Härdle, and A. Unwin, Eds. Springer, 447–469.

Strobl, C., Boulesteix, A.-L., Zeileis, A. and Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution, *BMC Bioinformatics*, **8**, 25.