ECONOMIC FORECASTING BY A PIECEWISE REGRESSION TREE METHOD

YOUNGJAE CHANG

Korea National Open University
E-mail: yjchang@knou.ac.kr

Abstract- Real GDP growth rate is an indicator which can be used to measure the overall economic achievement of a country during the certain amount of time. Policy makers are very interested in GDP growth rate, especially, forecasting the growth rate of a country. We propose the application of regression tree methods to do an economic forecasting. The method is quite easier than complex models due to flexible model fitting algorithm, and gives us reasonable forecasting results.

Index Terms- Data Mining, Forecasting, GDP Growth Rate, Regression Tree

I. INTRODUCTION

Real GDP growth rate is an indicator which can be used to measure the overall economic achievement of a country during the certain amount of time. Policy makers are very interested in GDP growth rate, especially, forecasting the growth rate of a country. It is important for their purpose because they can prepare the policy in advance based on the forecasts. Many modeling methods have been applied to the problem of economic forecasting. We propose the application of regression tree methods to do an economic forecasting. The method is quite easier than complex models due to flexible model fitting algorithm, and gives us reasonable forecasting results. This paper is organized as follows. Chapter II covers the basic principle of regression tree method and its application. Chapter III presents major results followed by chapter IV which contains some discussions.

II. METHODS

A. Regression tree

The tree model is a representative analytical method of data mining that performs the classification analysis or regression analysis by plotting the analysis process into a tree structure. When the response variable is a categorical variable, it is called a decision tree or classification tree. In the other hand, if the response variable of a tree model is a numerical variable, we call it a regression tree. Therefore we can regard the regression tree as a kind of tree structured regression model. In general, a regression tree is defined as a tree structured solution in which a constant or a regression model is fitted in each partition of data. From the machine learning perspective, we can define a regression tree as a recursive partitioning algorithm which gives predicted value at the terminal nodes. A simple example of a regression tree is presented in Figure 1.

Figure 1 is a regression tree of which response variable is the business sentiment index in Chang (2010). The predictors are sales index, facility investment index, and KOSPI index which is the Korean stock market index. Every index is transformed to an increase rate over the previous year. At each intermediate node, a case goes to the left child node if and only if the condition is satisfied. Equation (1) is a brief representation of the regression tree in Figure 1. We can see that the structure of regression tree has some combination of variables, which can be interpreted as several interaction terms in equation (1).

$$y = \beta_0 + \beta_1 I(Sales < 10.2) * I(FI < 15.87) + \beta_2 I(Sales \leq 10.2) * I(FI > -15.87) * I(kospi \leq -0.79) + \cdots \quad (1)$$

Flexible model fitting including mixture of interaction terms is well known strong point of regression trees. We do not worry about the combination of predictors for the interaction terms since a regression tree itself selects split variables through the algorithm. We have only to interpret the tree once the tree is finally grown.

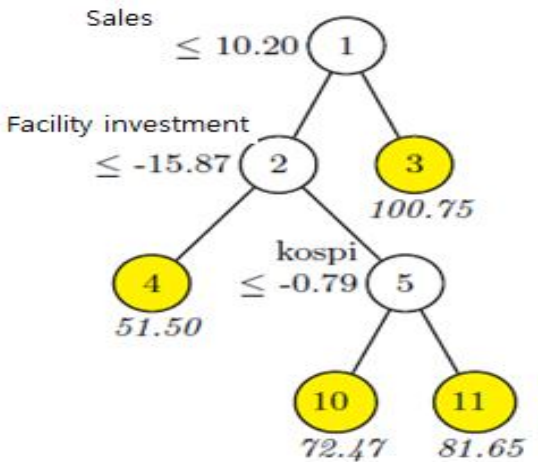


Figure 1. An example of regression tree

B. Piecewise regression tree

GUIDE, proposed by Loh (2002) is a flexible and semi-parametric model fitting algorithm, and very easy to understand. GUIDE stands for Generalized Unbiased Interaction Detection and Estimation. The performance of GUIDE, in terms of prediction accuracy, is also known to be better than any other

single tree approach in the data mining field such as CART proposed by Breiman et al (1984). GUIDE has nearly no selection bias because it has bias correction procedure based on bootstrap method during the model building steps. Chang and Kim (2011) summarized the algorithm as following.
<Algorithm of piecewise regression tree>

1. Let t be the current node. Fita stepwise regression model to the data in the current node. This procedure determines addition or deletion of predictors. The default values of F-to-enter and F-to-remove are 4.00 and 3.99.
2. Split a node if the R-square value of the current fitted model is less than 0.99 or the current node has more than 2n observations. n is a number which is previously specified.
3. Define the class variable Z by the sign of its residual after fitting a regression model for each observation. Z=1 if the observation is associated with a positive residual, Z=0 otherwise.
4. (Curvature test) For each predictor, we can construct a 2 × m cross-classification table. Note that the rows of the table are the values of Z, and the columns of the table are 4 intervals. The intervals are divided at the sample quartiles if the predictor is a numerical variable. If the predictor is a categorical variable, its m distinct categories are the columns of the table. Do the chi-squared test for each predictor and compute a p-value based on the contingency tables.
5. For numerical predictor pairs and categorical predictor do chi-squared tests to detect interactions. 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. The split variable X is determined by the comparison of p-values. Let tL and tR denote the left and right sub nodes of t. If X is a numerical predictor, the split point is the point which gives the lowest total of the sums of squared residuals in tL and tR. If X is a categorical predictor, the split combination is of the form $X \in C$, which gives the lowest weighted sum of the variances of Z in tL and tR, where C is a subset of the values taken by X. The weights are proportional to sample sizes.
7. Finally, prune the tree by cross-validation in general.
We can simply apply the piecewise regression tree algorithm. GUIDE algorithm described above is used in this paper to do a forecasting job.

C. Curvature test

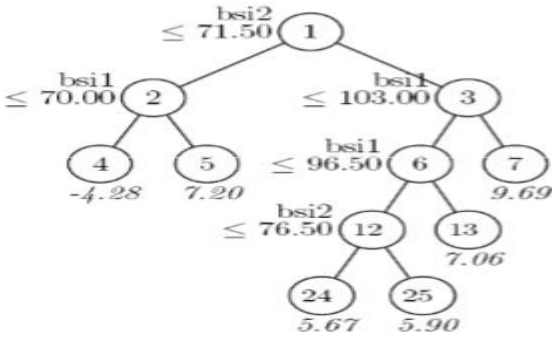
Chang and Kim (2011) also illustrated the flexible model fitting procedure based on curvature testing step in the piecewise regression tree algorithm. A linear regression model can be fitted at each node of the regression tree. If a predictor has nonlinear relationship with the response variable, it is a quite good candidate variable for the split. For example, a predictor x in the equation (2) has a "V" shape in the plane.

$$y_i = \beta_0 + \beta_1 |x_i - 2| + \epsilon_i \qquad (2)$$

The model can be regarded as the mixture of two linear lines of which border is x=2. We can interpret the model in the piecewise regression manner. We can think about the relationship between the predictor x and response y in terms of regression tree. In the piecewise regression tree, the split variable is x, where the split point is x=2. We can fit a negative slope regression line and a positive slope regression line separately in each subnodes of a regression tree after a node is split at x=2. Curvature test can show us which predictors have more nonlinear relationship with the response variable in the form of p-value. As Chang and Kim (2011) pointed out, curvature test procedure make us fit a more flexible nonlinear model through these steps.

D. For the forecasting

Due to flexible model fitting, regression tree can be used for forecasting. Figure 1 shows us the relationship between business sentiment index and several economic variables in regression tree at the same time point. In fact, the example is a kind of cross sectional study as we see. In addition to the usual regression model, we can easily fit the forecasting model in the form of regression tree. Once we get meaningful predictors which can be used in the prediction model, we can fit a quite simple piecewise regression tree model. For example, a simple time series model with a lagged variable can be fitted in the form of regression tree.
Figure 2 is an example of a piecewise regression tree which has two lagged predictors. The response variable is GDP growth rate at time t, and predictors in the tree are business sentiment index at the time of t-1 and t-2.



- NODE4 : GDP = −21.721 + 0.335*BSI1
- NODE5 : GDP = −0.411 + 0.0903*BSI1
- NODE24: GDP = −0.773 + 0.0868*BSI1
- NODE25: GDP = −8.771 + 0.170*BSI1 + 0.00705*BSI2
- NODE13: GDP = −36.250 + 0.436*BSI1
- NODE7: GDP = −10.303 + 0.0844*BSI1 + 0.0980*BSI2

Figure 2. An example of regression tree for forecasting

Regression models at each node are presented below the tree diagram. We can see that GDP growth rate is high if the business sentiment indexes one or two quarters earlier are generally high. Note that some models have both bsi1 and bsi2 variables, while others have only one variable. The variable selection result in the regression model at each node is owing to the stepwise selection procedure in the algorithm. Any leading indicator can be put into the regression tree as predictor variables.

E. Data

We fit a piecewise regression tree for the forecasting of GDP growth rate. The response variable in the model is GDP growth rate (year on year). Predictors are several preceding IP(industrial production) indexes which are consumer sales, facility investment, construction investment, etc. Since GDP growth rate is quarterly basis, we transform the monthly IP indexes into quarterly growth rates (year on year). Analysis period is from the 1st quarter of 2000 to 4th quarter of 2016. Three quarter lagged variables of every predictor variable are considered in the model initially. After we put the all the variables including the lagged variables, variable selection is done by the stepwise regression steps and curvature tests in the fitting procedure of regression tree. We set every option in the program as it is, say default values. For the computing program, we use GUIDE algorithm since it is known to be one of good algorithms which can implement the idea of the piecewise regression tree and give a good prediction accuracy. Data description is in Table 1. KOSIS stands for statistical database of Statistics Korea, and ECOS represents that of the Bank of Korea which is the central bank of Korea. The number at the end of variable name stands for the lagged quarters.

Name	Lagged variables	Source
Consumer sales	S1, S2, S3	KOSIS
Facility investment	F1, F2, F3	KOSIS
Construction investment	C1, C2, C3	KOSIS
Machinery orders	O1, O2, O3	KOSIS
Operation rate	Op1, Op2, Op3	KOSIS
Export	E1, E2, E3	ECOS
GDP	Response variable	ECOS

Table 1. Data description

Figure 3 shows the trends of Sales, Facility investment, and Construction indexes. We can see that both Facility investment and Construction investment indexes' movements are similar to each other, while Consumer sales index is somewhat different from others. The pattern of Sales index looks much more stable.

Figure 4 shows the series of Machinery orders, Operation rate, and Export indexes. We can see that the movements of all the indexes in this figure look quite similar to each other. This is a little different from the Figure 3. Moreover there seems to be a co-movement of several indexes, which could give a good background for proper forecasting.

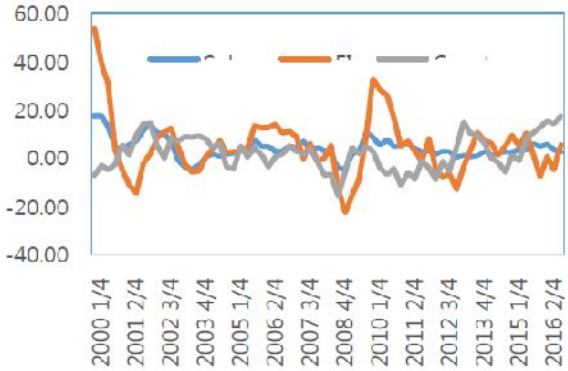


Figure 3. Sales, Facility investment, and Construction indexes

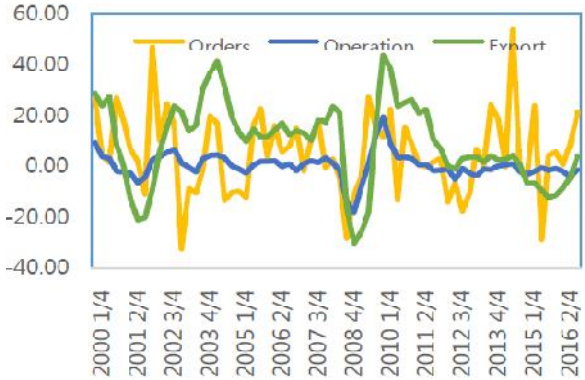


Figure 4. Machinery orders, Operation rate, and Export indexes

Figure 5 below is the year on year GDP growth rate of Korea. We can see that overall trend goes downward even if there is a big trough during the global financial crisis period between 2008 and 2009. There seems a level shift between the period before and after the global crisis. The average growth rate from the first quarter of 2000 to third quarter of 2008 is 5.39%, while that of the period between the first quarter of 2009 and the fourth quarter of 2016 is 3.56%.



Figure 5. GDP growth rate of Korea (year on year)

III. RESULTS

We fit a piecewise regression trees using GUIDE. As described above, the response variable is the GDP growth rate, and predictors are lagged variables in the PI indexes. We fit two regression trees. One is a piecewise linear regression tree with a stepwise

regression model at the terminal nodes, and the other is a piecewise constant regression tree with a constant fitting at the terminal nodes instead of linear models. As expected, the size of piecewise linear regression tree is quite smaller than constant regression trees. In our analysis, the shape of piecewise linear regression tree turns out to be a single node after pruning procedure. Note that the tree, however, has a linear regression model at the single node which is presented in Table1.



Figure 6. Piecewise regression tree with a model at the terminal node

Regressor	Coeff	t-stat	p-vale
Constant	2.9367	12.97	0.000
S1	0.2341	5.40	0.000
C1	0.0668	3.17	0.002
Op1	0.1754	5.19	0.000

Table 1. Regression model at the terminal node

Figure 6 is a fitted piecewise linear regression tree of GDP growth rate. It is a GUIDE 0.50-SE piecewise linear least-squares regression tree with stepwise variable selection for predicting GDP. At each split, an observation goes to the left branch If and only if the condition is satisfied. Sample sizes (in italics) and means of GDP are printed below nodes. Second best split variable (based on curvature test) at root node is E3. We can see that Sales, Construction, Operation rate in the previous quarter are important variables to tell the current GDP growth rate as well as Export one quarter earlier.

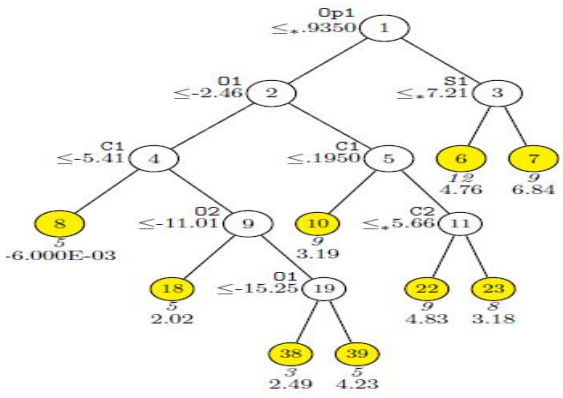


Figure 7. Piecewise constant regression tree

Figure 7 shows a GUIDE 0.50-SE piecewise constant least-squares regression tree for predicting GDP. At each split, an observation goes to the left branch if and only if the condition is satisfied. The symbol '≤*__' stands for '≤ or missing'. Sample sizes (in italics) and means of GDP are printed below nodes. Second best split variable (based on curvature test) at root node is S1.The reason we get such a complicated structure is that the models at the terminal nodes are only the constants. We can see that Operation rate, Machinery orders, Sales, Construction in the previous quarter as well as Construction and Machinery orders in the two quarters earlier are important variables. In general, the predictors which show up in the upper level of splits in a regression tree are regarded as more important variables.

We also compare the performance of the algorithms in terms of prediction accuracy. We set aside GDP growth rates for one year, namely four quarters in 2015 for a testing dataset. In fact, observations for last four quarters in 2016 are excluded for the testing purpose because of constantly lower growth rates during that period. We fit regression tree models with data until 2014 and predict growth rates based on the models for a year. Then we calculate a mean squared error, which is expressed as the equation (3) for the comparison. Surely, the lower value implies the better performance. We include an AR(2) model for a naïve model for comparison purpose. Out of sample forecasting errors are presented in Table 2.

$$MSE = \frac{\sum_{i=1}^{4}(y_i - \hat{y}_i)^2}{4} \qquad (3)$$

Algorithm (Model)	Piecewise linear regression tree	Piecewise constant regression tree	AR (2)
MSE	0.094	2.418	0.654

Table2. Comparison of prediction accuracy

We can see that the performance of piecewise linear regression tree is better than other algorithms.

IV. DISCUSSION

This paper suggests a forecasting method using a regression tree algorithm. The models are based on the piecewise linear regression tree and the piecewise constant regression tree model in detail. In this study, we can conclude that the prediction based on the regression tree gives us a reasonable and meaningful results. Both methods have been found to be more accurate than other general models. Overall, the use of data mining algorithms such as regression trees can be a good candidate model reflecting the relationship between economic predictors and GDP growth rate at the first level with minimum information. However, it is considered that the frequent change of the economic cycle in a relatively short period of time was limited to effectively predict the growth rate change in the future.

REFERENCES

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

[2] Chang, Y. and Kim, H. (2011). Tree-Structured Nonlinear Regression, The Korean Journal of Applied Statistics, 24(5), 759–768..

[3] Loh (2002). Regression trees with unbiased variable selection and interaction detection Statistica Sinica, 12, 361–386.