Communications of the Korean Statistical Society
2012, Vol. 19, No. 2, 293–301

DOI: http://dx.doi.org/10.5351/CKSS.2012.19.2.293

Wage Determinants Analysis by Quantile Regression Tree

Youngjae Chang[1,a]

[a]Research Department, The Bank of Korea

Abstract

Quantile regression proposed by Koenker and Bassett (1978) is a statistical technique that estimates conditional quantiles. The advantage of using quantile regression is the robustness in response to large outliers compared to ordinary least squares(OLS) regression. A regression tree approach has been applied to OLS problems to fit flexible models. Loh (2002) proposed the GUIDE algorithm that has a negligible selection bias and relatively low computational cost. Quantile regression can be regarded as an analogue of OLS, therefore it can also be applied to GUIDE regression tree method. Chaudhuri and Loh (2002) proposed a nonparametric quantile regression method that blends key features of piecewise polynomial quantile regression and tree-structured regression based on adaptive recursive partitioning. Lee and Lee (2006) investigated wage determinants in the Korean labor market using the Korean Labor and Income Panel Study(KLIPS). Following Lee and Lee, we fit three kinds of quantile regression tree models to KLIPS data with respect to the quantiles, 0.05, 0.2, 0.5, 0.8, and 0.95. Among the three models, multiple linear piecewise quantile regression model forms the shortest tree structure, while the piecewise constant quantile regression model has a deeper tree structure with more terminal nodes in general. Age, gender, marriage status, and education seem to be the determinants of the wage level throughout the quantiles; in addition, education experience appears as the important determinant of the wage level in the highly paid group.

Keywords: Quantile regression, nonlinear quantile regression, tree-structured regression.

1. Introduction

Quantile regression originally proposed by Koenker and Bassett (1978) is a statistical technique that estimates conditional quantiles. It is originated from the linear l_1-regression problem by Barrodale and Roberts (1980), Bartels and Conn (1980) and others which is also based on Charnes *et al.* (1955) and Wagner (1959). Koenker and Bassett (1978) extended these algorithms to linear quantile regression. Koenker and D'Orey (1987) improved the linear quantile regression based on the simplex method by modifying the Barrodale-Roberts algorithm.

Koenker and Park (1994) proposed a new approach to the computation of nonlinear quantile regression estimators based on interior point methods for solving linear programs. They discussed the interior point methods to solve strictly linear programs (that include the linear quantile regression problem) and extended them to nonlinear problems. It turned out the interior point algorithm offered the natural extension to nonlinear problems unlike the simplex method. A regression tree approach has been applied to ordinary least squares(OLS) problems to fit flexible models. Loh (2002) proposed the GUIDE algorithm, which has a negligible selection bias and a relatively low computational cost. GUIDE is also known as a smart data mining tool with a flexible model fitting methods at each node.

Just as an OLS problem is solved by GUIDE, a quantile regression problem can be dealt with in a piecewise regression tree approach. Chaudhuri and Loh (2002) proposed a nonparametric quantile

[1] Economist, Research Department, The Bank of Korea, 39, Namdaemun-Ro, Jung-Gu, Seoul 110-794, Korea.
E-mail: yjchang@bok.or.kr

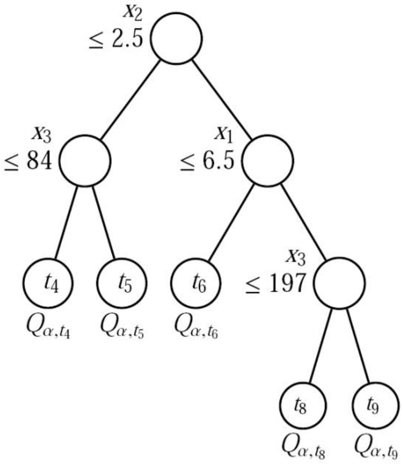

Figure 1: *Example of a quantile regression tree for α quantile: At each intermediate node, a case goes to the left child node if the condition is satisfied. Number beneath a leaf is sample quantile of the dependent variable.*

regression method that blends key features of piecewise polynomial quantile regression and tree-structured regression based on adaptive recursive partitioning. Unlike least squares regression trees that concentrate on modeling the relationship between the response and the covariates at the center of the response distribution, the method can provide insight into the nature of that relationship at the center as well as the tails of the response distribution.

Figure 1 shows an example of a quantile 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. At the terminal nodes, each quantile regression model is fitted based on data partitioned.

Tree-structured quantile regression algorithm has the advantage of fitting flexible models since they can capture non-linearity by piecewise linear models fitted at the terminal nodes.

Lee and Lee (2006) investigated wage determinants in the Korean labor market using the Korean Labor and Income Panel Study(KLIPS). They used the quantile regression method for each conditional quantile wage group. Quantile regressions in the paper examined more comprehensive pictures for different quantile wage groups while most other previous labor market analyses use (mean) regression analysis, that focused only on average statistics.

They discovered that education does not seem always appear to provide the necessary job skills, so that the return on education is fairly low compared to the US labor market. Age; however, it was shown to be one of the most important factors for wage determination especially for the higher wages groups. Likewise, they did several analyses to find the respective relationship between wage and independent variables in the data. We are interested in the analysis of KLIPS data using a quantile regression tree approach. We may find more interesting results compared to those of the simple quantile regression method used in Lee and Lee (2006). The characteristics of a quantile regression tree may give us a more helpful and meaningful result. This paper is organized as follows. In Section 2, we introduce the concept of quantile regression followed by a tree-structured quantile regression algorithm (GUIDE). Section 3 covers real data analysis and Section 4 concludes with a summary of the real data analysis using a quantile regression tree.

2. Quantile Regression and GUIDE

2.1. Quantile regression

Quantile regression analysis focuses on the conditional α^{th} quantile of the response variable given the predictor variables. Unlike usual regression analysis that focuses on the conditional mean of the response given the predictors, quantile regression gives insight into the center as well as the lower and upper tails of the conditional distribution of the response with varying choices of α. Chaudhuri and Loh (2002) pointed out that quantile regression is quite effective as a tool to explore and model the nature of dependence of a response on the predictors when the predictors have different effects on different parts of the conditional distribution of the response that occurs in many econometric problems. For example, in marketing studies, where predictor variables may have different effects on high, medium and low consumption groups, quantile regression can be useful to understand the nature of the dependence between the response and the predictors. Besides the effective modeling, the advantage of using quantile regression is the robustness in response to large outliers compared to ordinary least squares(OLS) regression that can be easily understood.

Quantile regression can be described as following according to Koenker (2005); Let Y be a dependent variable, X a (d-dimensional) predictor variable,

$$Q_\alpha(X = x) = \inf\{y : F(y|X = x) \geq \alpha\}.$$

The conditional distribution function $F(y|X = x)$ is,

$$F(y|X = x) = P(Y \leq y|X = x),$$

where $F(\cdot)$ is cdf of Y and consequently α^{th} quantile of Y is $F^{-1}(\alpha) = \inf\{y : F(y|X = x) \geq \alpha\}$. Let the check function be ρ_α is $\rho_\alpha(u) = u(\alpha - I(u < 0))$. Then looking for the $\hat{y}$ that minimizes

$$E_{\rho_\alpha}(Y - \hat{y}) = (\alpha - 1)\int_{-\infty}^{\hat{y}} (Y - \hat{y})dF(y) + \alpha \int_{\hat{y}}^{\infty} (Y - \hat{y})dF(y).$$

Leads to the first order condition,

$$0 = (1 - \alpha)\int_{-\infty}^{\hat{y}} dF(y) - \alpha \int_{\hat{y}}^{\infty} dF(y) = F(\hat{y}) - \alpha.$$

While least-squares regression focuses only on the conditional mean $E(Y|X = x)$ that minimizes the expected squared error loss, the objective of quantile regression is to find the conditional quantile that minimizes the expected loss $E(\rho_\alpha)$,

$$Q_\alpha(X = x) = \arg\min_{\beta \in R^d} E(\rho_\alpha(Y - x'\beta)).$$

2.2. GUIDE quantile regression

The aim of regression analysis is to discover the relationships between the response variable and the predictor variables, and eventually to use the relationships to make predictions based on the information. 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. Chaudhuri and Loh (2002) proposed a nonparametric quantile regression method using a regression tree. Quantile regression trees have a piecewise constant, piecewise polynomial, or piecewise multiple linear option, where each piece is obtained by fitting respective corresponding models to the data in the terminal node of a binary tree. The tree is constructed by recursively partitioning the data based on repeated analyses of the residuals obtained after model fitting with quantile regression. This idea is implemented in GUIDE(Generalized, Unbiased, Interaction Detection and Estimation (Loh, 2002)) software, of which a multiple linear procedure is briefly sketched as follows:

1. Fit a quantile regression model to the data in the node using the algorithm in Koenker and D'Orey (1987) and compute the residuals.

2. 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 $Z = 0$.

3. 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.

4. 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.

5. After splitting has stopped, prune the tree with a test sample or by cross-validation.

For details, see Loh (2002).

3. Real Data Analysis

3.1. Data description

We use data from the Korean Labor and Income Panel Study(KLIPS) due to Lee and Lee (2006). The Korea Labor Institute began to collect detailed data for households and individuals starting in 1998. The data collection is modeled after the Panel Study of Income Dynamics(PSID) from the University of Michigan. We use data of 2007 and currently employed ones at the year of 2007 are selected from among the 13,738 individual observations in the dataset. That is, self-employed or unemployed observation are excluded for the analysis. The wage variable is an average monthly wage in Korean won in 10,000 won units. Independent variables are generated as described in Lee and Lee (2006). Education is measured as the total number of years in school. The original education variable in the dataset is a categorical variable that was converted to a numerical variable based on duration of schooling. For example, graduation from elementary school gives six, from middle school gives nine, and so on. Occupational types are categorized as highly skilled white-collar, lower-skilled

Table 1: List of variables

	Variables
Dependent variable	Wage
Independent variables	Education, Age, Job experience Total jobs, High White, Low White, High Blue, Low Blue Origin 1 through 4, Region 1 through 3, Gender, Married, Union

white-collar, highly skilled blue-collar and lower skilled blue-collar jobs. Origin variables are dummy variables of birthplace. Origin 1 is for Kyungsang-do, Origin 2 for Seoul, Incheon, and Kyunggi-do, Origin 3 for Chola-do and Jeju-do, and Origin 4 for Chungcheng-do and Kangwon-do and the rest of Korea. Regional variables are also dummy variables with Region 1 for Seoul, Region 2 for other large metropolitan areas, and Region 3 for all other areas. The list of variables is in Table 1.

3.2. Results

We fit three quantile regression models to the KLIPS data using GUIDE. Piecewise constant, piecewise multiple linear, and piecewise simple linear quantile regression models are fitted. Each tree of quantiles $\alpha = 0.05, 0.20, 0.50, 0.80$ and 0.95 is presented. These three kinds of trees give quite similar results; however, some trees may have more detailed information based on the models fitted at the terminal nodes than others.

3.2.1. Piecewise constant quantile regression tree

We can see that AGE is the first split variable in the lower quantiles ($\alpha = 0.05, 0.20$). GENDER and MARRIED are also the common variables that show up in the lower quantile trees. Especially, AGE appears three times in the lowest quantile, which means AGE divides the wage level into many pieces compared to other quantiles. We can see that the married males between 24.5 and 43.5 years old get paid most in the 5 percentile wage level. The 20 percentile tree looks a little different from the lowest quantile tree. The education duration more than 14.5 years gives the highest wage level among people over 43.5 years old. The next highest level belongs to married males not more than 43.5 years old. The older people get paid more provided that they are relatively well educated. The median tree gives quite different result from previous two trees with lower quantiles. The education experience determines the wage level in the root node. It is regarded as the most important variable in the tree that tells the wage level. The highest level comes from the group with more than 15.5 years of education experience, with ages older than 36.5. We can see that the split point of the AGE variable is quite low that means there exist well-paid young people in the median level compared to the lower quantiles. Concerning GENDER, males get paid more like the previous two trees. The only split variable is EDUCATION in the 80 and 95 percentile trees. Education duration more than 15.5 years get paid more in the higher quantiles. The education duration of 15.5 years could be regarded as nearly university-graduate level. Other particular things we can see from these high wage level trees are the disappearing of the two variables AGE and GENDER. It seems that education mainly determines the wage level in the well-paid groups.

3.2.2. Multiple linear quantile regression tree

For the multiple linear quantile regression trees, we do not see any variable for the split at the 5, 20, and 95 percentiles. The model fitted at the terminal node is in the form of multiple regression, so the split structure rarely appears in comparison to the constant quantile regression trees. The multiple regression model may contain the curvature structures that are presented by the split variables in

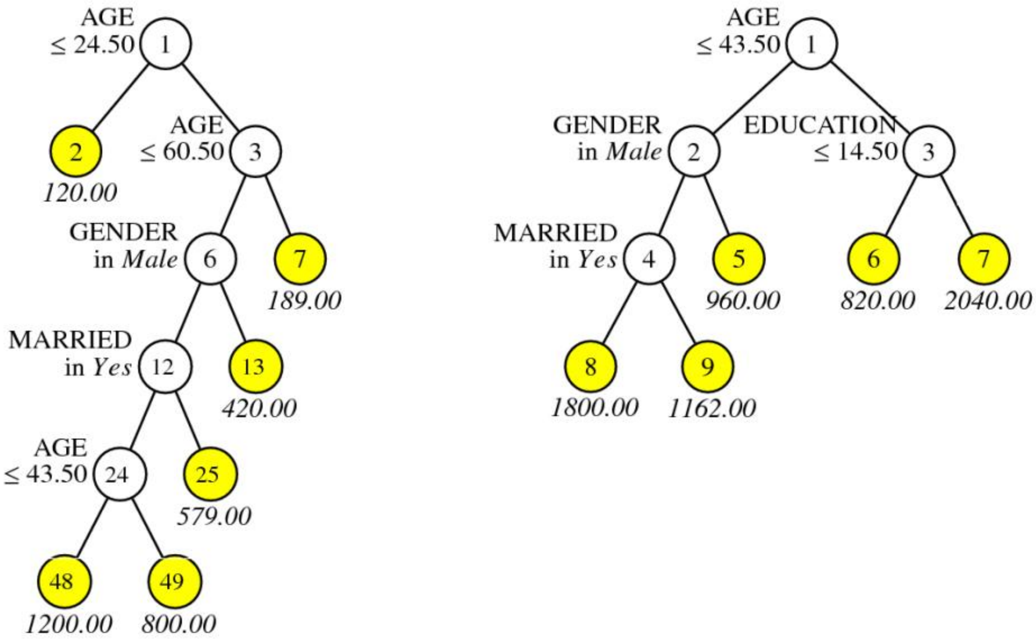

Figure 2: *GUIDE piecewise constant quantile regression tree (The left tree is for the 5 percentile and the right one is for the 20 percentile).*

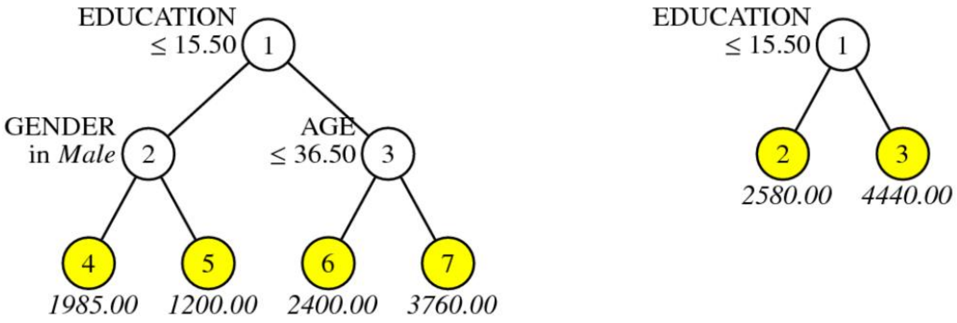

Figure 3: *GUIDE piecewise constant quantile regression tree (The left tree is for the 50 percentile and the right one is for the 80 percentile).*

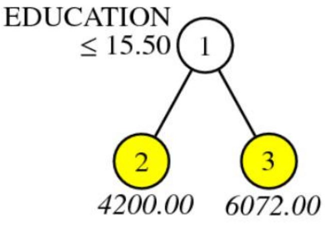

Figure 4: *GUIDE piecewise constant quantile regression tree (This tree is for the 95 percentile).*

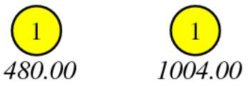

Figure 5: *GUIDE piecewise multiple linear quantile regression tree (The left tree is for the 5 percentile and the right one is for the 20 percentile).*

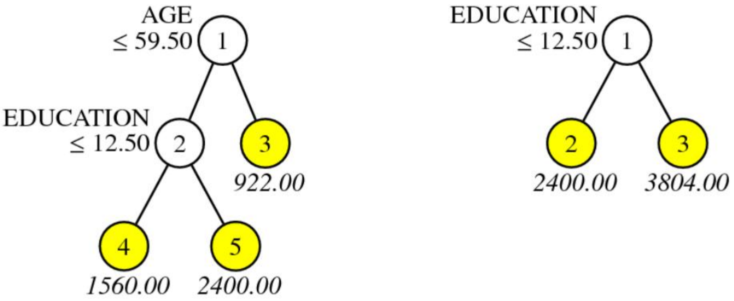

Figure 6: *GUIDE piecewise multiple linear quantile regression tree (The left tree is for the 50 percentile and the right one is for the 80 percentile).*

Figure 7: *GUIDE piecewise multiple linear quantile regression tree (This tree is for the 95 percentile).*

the constant regression trees. Only the 50 and 80 percentile trees have the splits with AGE and EDUCATION variables The 50 percentile multiple linear regression tree has the same split variables as those in the 50 percentile constant tree, but the split points are different because of the model fitted at the terminal nodes. The 80 percentile multiple linear tree also has the same split variable, EDUCATION, but the split point is 12.50 and is a little smaller than that in the constant tree. The difference is explained by the model at the terminal node. The multiple linear quantile regression trees comply with constant quantile regression trees.

3.2.3. Simple linear quantile regression tree

A simple linear tree is the tree that has a simple regression model at the terminal nodes. The difference between the multiple linear regression tree and simple linear regression tree is the number of predictors of the models at the terminal nodes. So, we fit a model with only one predictor at the terminal node. The best simple linear model fitted at the terminal nodes is best in the sense that the model gives the lowest mean squared error. Such a simple linear regression tree is useful when a piecewise constant tree has too many nodes but a piecewise linear one has too few. MARRIED, AGE, and GENDER variables come out as split variables at the 5 percentile tree that also appear as split variables in the 5 percentile constant tree; however, the structure is quite different from each other. Note that EDUCATION appears almost everywhere as the predictor variable at one of the terminal nodes.

4. Conclusion and Future Work

Following Lee and Lee, we fit three kinds of quantile regression tree model to KLIPS data with respect to the quantiles, 0.05, 0.2, 0.5, 0.8, and 0.95. Among the three models, multiple linear piecewise

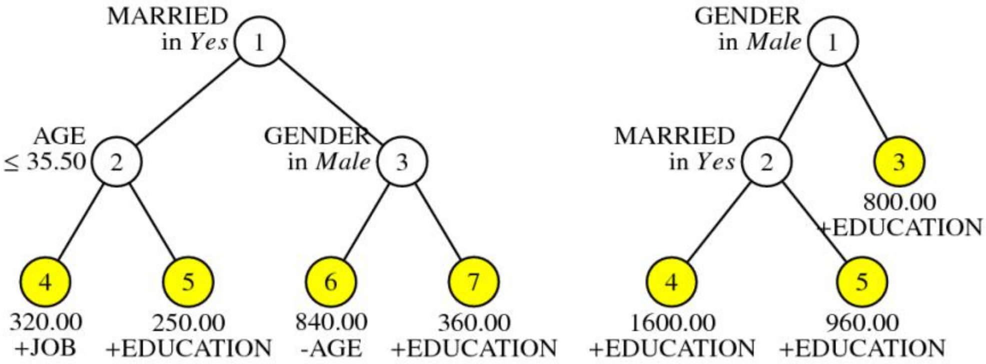

Figure 8: *GUIDE piecewise simple linear quantile regression tree (The left tree is for the 5 percentile and the right one is for the 20 percentile).*

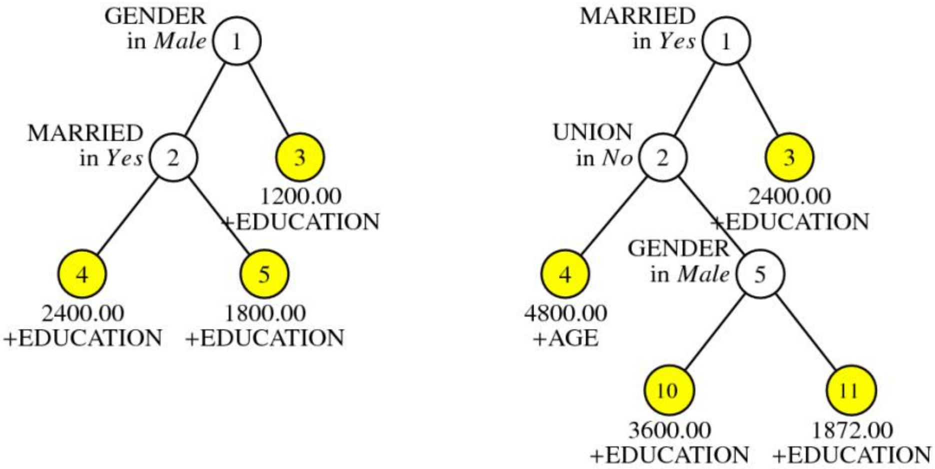

Figure 9: *GUIDE piecewise simple linear quantile regression tree (The left tree is for the 50 percentile and the right one is for the 80 percentile).*

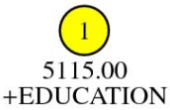

Figure 10: *GUIDE piecewise simple linear quantile regression tree (This tree is for the 95 percentile).*

quantile regression model forms the shortest tree structure, while the piecewise constant quantile regression model has a deeper tree structure with more terminal nodes in general. This implies that we can simplify the tree structure by fitting a linear model instead of a constant at each node. This result corresponds to that of a usual regression tree approach. Similar to Chang's analysis (2010) of the impact factors for the Business Survey Index(BSI) using regression trees, we can easily detect the important factors that impact wage levels from several quantile regression trees in this paper. AGE appears as a very important determinant of wage in the lowest paid group. It seems that EDUCATION mainly determines the wage level in the well-paid groups. Concerning GENDER, males get paid more than females in general. There is also some room for research improvement in the near future. We can think about an extension of the cross-sectional quantile regression tree analysis to compare the changes in determinants as time goes on. We may consider the application of the panel data analysis method to the regression trees to make it possible.

References

Barrodale, I. and Roberts, F. D. K. (1980). Solution of the constrained l_1 linear approximation problem, *ACM Transactions on Mathematical Software*, **6**, 231–235.

Bartels, R. and Conn, A. (1980). Linearly constrained discrete 1 problems, *ACM Transactions on Mathematical Software*, **6**, 594–608.

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

Charnes, A., Cooper, W. W. and Ferguson, R. O. (1955). Optimal estimation of executive compensation by linear programming, *Management Science*, **Jan**, 138–151.

Chaudhuri, P. and Loh, W.-Y. (2002). Nonparametric estimation of conditional quantiles using quantile regression trees, *Bernoulli*, **8**, 561–576.

Koenker, R. (2005). *Quantile Regression*, Econometric Society Monograph Series, Cambridge University Press.

Koenker, R. and Bassett, G. W. (1978). Regression quantiles, *Econometrica*, **46**, 33–50.

Koenker, R. and D'Orey, V. (1987). Algorithm AS229: Computing regression quantiles, *Applied Statistics*, **36**, 383–393.

Koenker, R. and Park, B. J. (1994). An interior point algorithm for nonlinear quantile regression, *Journal of Econometrics*, **71**, 265–283.

Lee, B.-J. and Lee, M. J. (2006). Quantile regression analysis of wage determinants in the Korean labor market, *The Journal of the Korean Economy*, **7**, 1–31.

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

Wagner, H. M. (1959). An integer linear-programming model for machine scheduling, *Naval Research Logistics Quarterly*, **6**, 131–140.

Received January 24, 2012; Revised February 21, 2012; Accepted March 6, 2012