Regression Tree - Framingham Heart Study data

Sangsoo Park (D)

April 14, 2014

1 What is Regression Tree?

Regression tree analysis has been used to understand underlying structures of data to predict a continuous outcome variable. The analysis divides data into sub-groups using binary branches, which indicates partioning of the outcome variable based predictor variables. There are two benefits using the analysis. The first one is that results from the analysis are simple to understand. Another one is that we don't need to assume linearlity relationship between ouctcome and predictor variables that linear regression model has. In the analysis, splits are determined by minimum residual sum of squares.That is, splits can be selected by minimization of within group sum of squares during maximization of between group sum of squares (ANOVA). This process is repeated using meaningful predictor variables that can greatly improve accuracy of the tree (least squares) until a minimum size of the end sub-groups is reached or no longer improvements in the accuracy of adding splits are showen.Thus, we need to have decision criteria for not only stopping criteria of the analysis but also having proper results without overfitting data. More details will be explained in the validation of the result section.

2 Methods

2.1 Data filtering

Five categorical and six continuous variables were selected because of our groups initial decision on what variables we are going to use. The 11 variables were extracted from the original dataset and the new dataset (frm2) was used for regression tree analysis.

2.2 Regression tree analysis

First, 10 variables were used to predict systolic blood pressure (SYSBP). The numbers at the terminal are mean systolic blood pressures of the number of observations falling in each terminal.To start, the initial cp value was set to 0.005, which might result in overfitting of the data. There were missing values in the outcome and predictor variables. If the outcome variable has missing values, all observations from the predictors were removed. However, outcome variable values were kept when one or more predictor variables have missing values.

2.3 Validation of the result

Keep adding more splits results in increase in R-squared even though the added splits could just result from adding more splits. This is what we saw in the multiple regression model. To avoid this overfitting problem, we need decision criteria on the appropriate number of splits.

The cp indicates complexity parameter that allows us to decide the number of meaningful splits (cost of adding one additional split).To be specific, higher cp values lead to smaller number of splits. we can decide the number of splits where the overall R-squared value is not increased more by increase in cp values (See Figure 1.Bottom left).

The minimum number of splits was defined by changes in X-val relative error as a function of cp values. After the size of tree is larger than 6, there were no further improvements in the X-val relative error while the cp values increase (See Figure 1.Top and Bottom right) Thus, six splits (cp=0.01) were decided here.

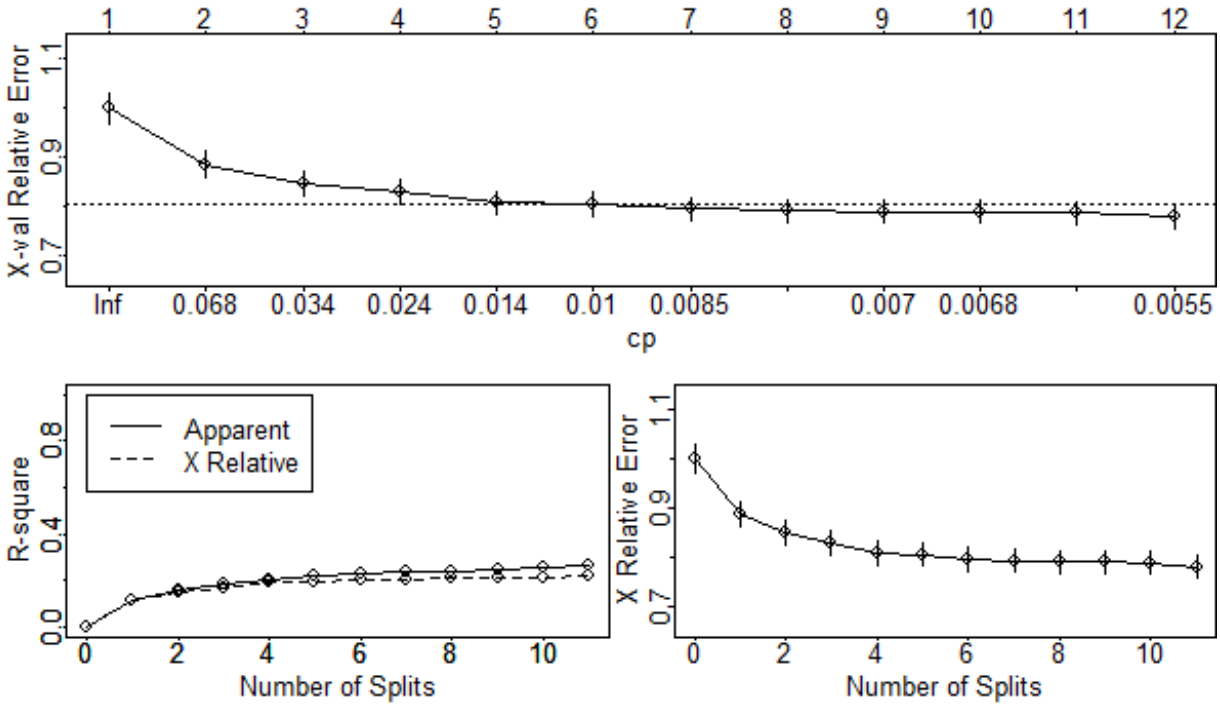

Figure 1: Decision on the number of splits (Top: X-Val Relative error change as a function of cp, Bottom Right: R-squared value, Bottom left: X Relative Error)

2.4 Comparison

The original tree was pruned based on the previous decision.The original tress used four variables to predict systolic blood pressure. The pruned tres also used the same four variables to predict the outcome variable but it used smaller number of splits (6 vs 11).The pruned and the original tree were visualized to see difference between the two trees (See Figure 2).

2.5 Multiple linear regression

Multiple linear regression analysis was performed with the same variables that the regression tree used. Results from the multiple regression gives us how an single outcome variable is changed with one unit increase in one of predictor variables while other predictor variables remain constant. In this context, regression tree encourages us to understand data set easier by binary branches or splits.

In terms of accuracy, the multiple linear regression analysis showed higher R-squared value (0.3 vs 0.2). However, it is really dependent on how many splits can be included in the regreesion tree model even though this report decided 6 splits. Thus, we might need other criteria to compare the two models.

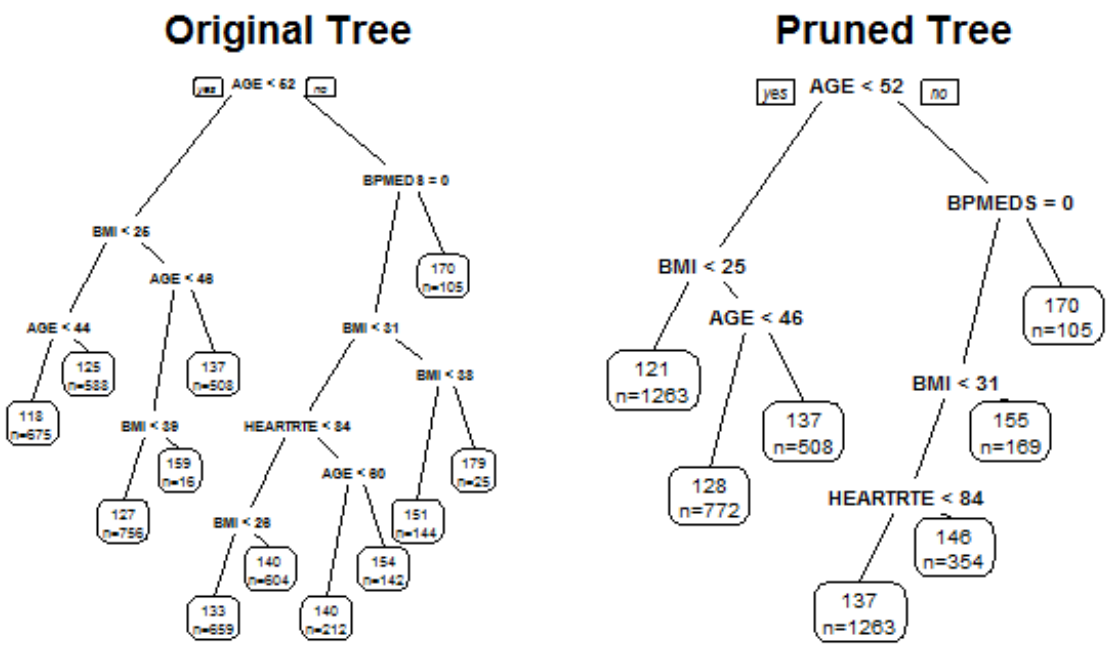

Figure 2: Original Tree vs Pruned Tree. Numbers at the terminal nodes indicate mean systolic blood pressure of the number of subjects that is falling into the terminal node, n = number of subjects
. If answer to condition in the circle is yes, you can go the left branch.