**Illustrating Regression Tree Using Boston Housing Data**

Tahani Omer tao34c@mail.umkc.edu

**Abstract**

Regression tree is a commonly used method in data mining. Similar to regression analysis, the goal is creating a model to predict the value of a response from several predictors. In this poster, we describe how to generate a regression tree using the decision tree method. This includes how to first build a large tree, how to prune it to find the best sub tree, and how to identify the cost complexity parameter via the cross-validation approach. Then, Regression is employed for the Boston housing data set. I found that the tree indicates the higher percent of lower status of the population, corresponds to less expensive houses.

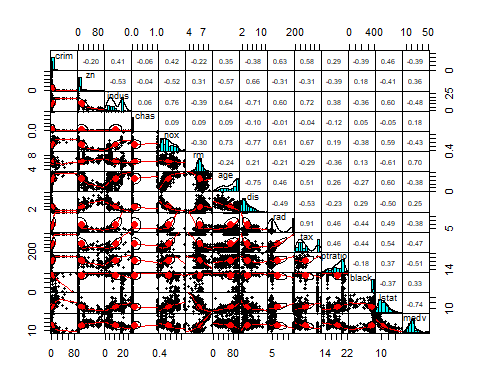
**Introduction**

To illustrate the decision tree, we used Boston dataset to predict the housing prices in Boston suburbs. This dataset is taken from the book “An Introduction to Statistical Learning with Application in R” by Gareth James, Daniela Witten, Trevor Hastie.

The dataset consists of 506 observations of 14 features. The median value of house price in $1000s, denoted by MEDV, is the outcome or the dependent variable in our model. Below is a brief description of each feature and the outcome in our dataset:

1. CRIM     – per capita crime rate by town
2. ZN     – proportion of residential land zoned for lots over 25,000 sq.ft
3. INDUS     – proportion of non-retail business acres per town
4. CHAS     – Charles River dummy variable (1 if tract bounds river; else 0)
5. NOX     – nitric oxides concentration (parts per 10 million)
6. RM     – average number of rooms per dwelling
7. AGE     – proportion of owner-occupied units built prior to 1940
8. DIS     – weighted distances to five Boston employment centers
9. RAD     – index of accessibility to radial highways
10. TAX     – full-value property-tax rate per $10,000
11. PTRATIO     – pupil-teacher ratio by town
12. Black    – 1000(Bk - 0.63) ^2 where Bk is the proportion of blacks by town
13. LSTAT     – % lower status of the population
14. MEDV     – Median value of owner-occupied homes in $1000’s

From the exploratory data analysis, we found that, out of all the variables, medv and lstat seem to have high negative correlation. In other words, we find that as the lower status of the population percentage increase the median value of owner-occupied homes decrease. And crim is strongly associated with variables rad and tax which implies as accessibility to radial highways increases, per capita crime rate increases. Also, indus has strong positive correlation with nox, which supports the notion that nitrogen oxides concentration is high in industrial areas. For more details the following figure show the correlation between all variables:



Since, the output variable” medv” is numerical, and the input variables are a mixture of continuous and categorical variables, thus the regression tree will be appropriate method for this dataset

**Method**

The method starts by searching for every distinct value of all its predictors, and splitting the value of a predictor that minimizes the following statistic:

Where are the average values of the dependent variable in groups and .

For groups and , the method will recursively split the predictor values within groups. In practice, the method stops when the sample size of the split group falls below certain threshold, e.g., 50.   
  
To prevent over-fitting, the constructed tree can be pruned by penalizing the SSE with tree size:

Where |T| is the number of terminal nodes of the tree T, and is the complexity parameter. Smaller will lead to larger trees, and vice versa. Of course, this parameter can also be tuned by cross-validation.

**Approach**

Methods of analysis includes the following: Summary statistics of the variables and finding correlation between variables, Exploratory data analysis using visualization. Random sampling of data set into 50/50 training and testing data set. Fitting various models such as Regression tree, Linear regression and K-Nearest Neighbors. Finally, comparing the models based on mean squared error (MSE).

**First**, we fit a regression tree to the Boston dataset. We built a large tree and stop when the number of the observations in the terminal node is the minimum number, then we prune the tree to find the best sub tree. The best idea is the cost-complexity parameter which is the best trad-of the best fit and the tree size. For that we used cross validation.

We divide the data in to k part, then we set side one part and fit the tree free size in the k – 1 part and evaluated the prediction error on the part we left out to choose alpha. Once we haven chosen alpha, we go back to the full tree and find sub tree which has the smallest error.

The model indicates that only three of the variables have been used in constructing the tree, the average number of rooms per dwelling” rm”, % lower status of the population “lstat” and the weighted distances to five Boston employment centers “dis”. The following figure show our model:

A screenshot of a cell phone

Description automatically generated

The variable lstat measures the percentage of individuals with lower socioeconomic status. The tree indicates that lower values of lstat correspond to more expensive houses. The tree predicts a median house price of $46,400 for a larger house in the suburbs in which residents have higher socioeconomic status (rm >= 7.437 and lstat < 9.715).

To test the accuracy of the model we calculate the mean squared error, which is given by

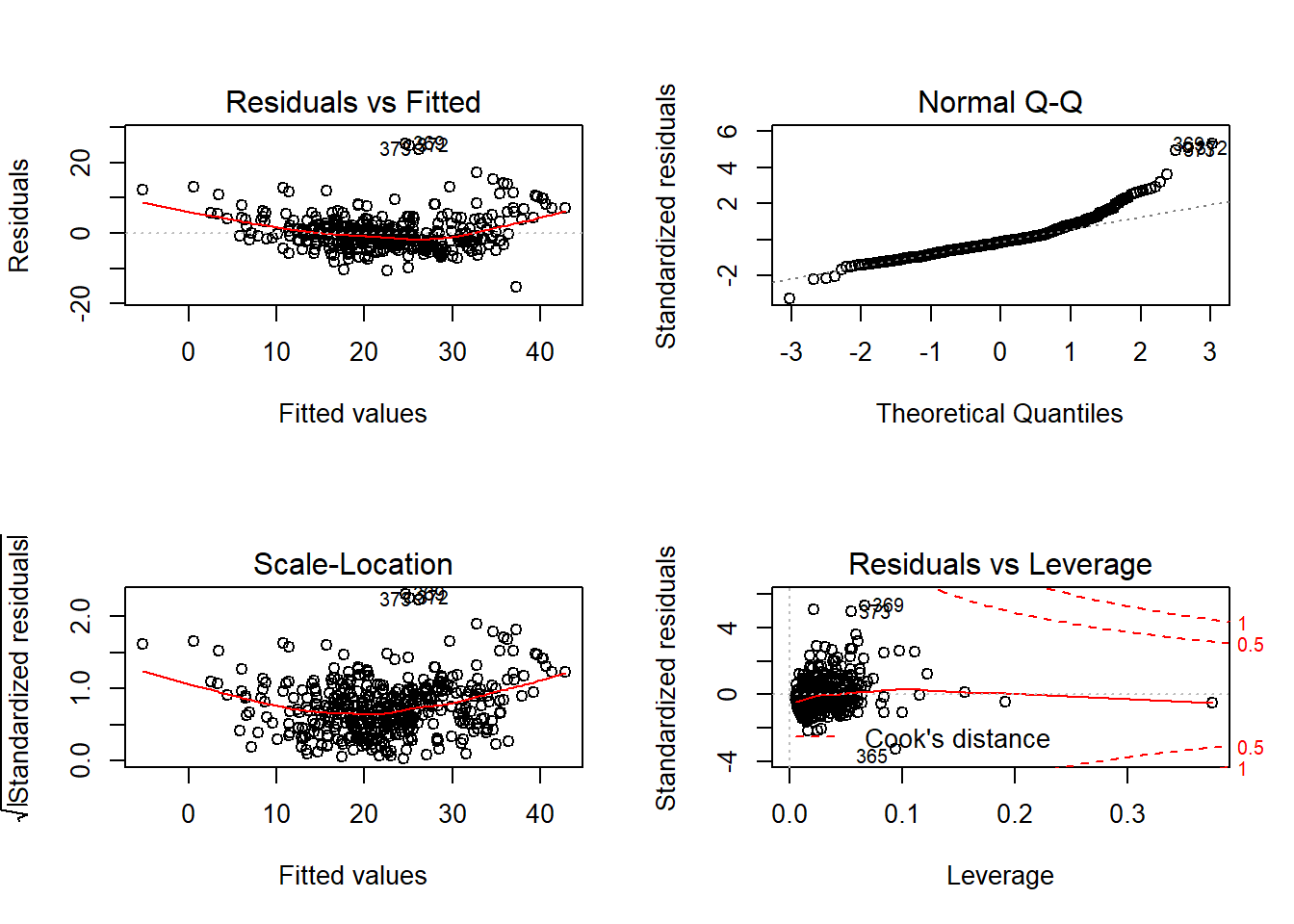
MSE

The test set MSE associated with regression tree is 25.05. The square root of the MSE is therefore around 5.005, indicating that this model leads to test predictions that are within around $5005 of the true median home value for the suburbs.

**Second**, we try generalized linear regression model with medv as the dependent variable and all the remaining variables as independent variables

We start off by fitting our model using all the data, and we found that all predictors are significant except the indus and the age of the house, based on p-value. We removed the insignificant variable, based on the best subset regression and Cross-validation, resulting in a model with adjusted R-squared 0.7097 and test prediction error of 15.7751.

Then we plot the chosen model to check the model assumption. Residuals vs Fitted plot shows that the relationship between” medv” and predictors is not completely linear. Also, normal qq plot is skewed implying that residuals are not normally distributed. A different functional form may be required. The following figure is the model diagnostics for the linear regression



**Third**, we fit K-Nearest Neighbors model. We started by fitting the model using different values of k and a different subset. We found that when k=10 and the subset of all variable except indus and age will give the best test prediction error. When we fit the model with k=10 the model seems to be able to capture non-linear relations in the data set. but still doesn’t give too much improvement compared to tree model.

**Finally**, a comprehensive model comparison was done between the three methods we used. Lower mean squared error, better is the model. And It was observed Regression Tree will perform the best among the models if we will use methods like bagging, random forests, and boosting, to improve the predictive performance.

**Conclusion**

The model has learned what the best questions to ask about the input data are, and can respond with a prediction for the median value of house price. Boston data collected from 1978 is not of much value in today’s world. But this data gives us a good example to apply the decision tree technique. So, throughout the Boston dataset we illustrated the decision tree from end-to-end and we learned and obtained several insights about regression tree model and how it is developed.

**Advantages**

Among other data mining methods, decision trees have various advantages:

* Simple to understand and interpret
* Able to handle both numerical and [categorical](https://en.wikipedia.org/wiki/Categorical_variable) data
* Requires little data preparation
* Built-in [feature selection](https://en.wikipedia.org/wiki/Feature_selection)
* Interpretation of a complex Decision Tree model can be simplified by its visualizations

#### **Disadvantages**

* There is a high probability of overfitting in Decision Tree.
* Generally, it gives low prediction accuracy for a dataset as compared to other machine learning algorithms.
* Trees can be very non-robust. A small change in the [training data](https://en.wikipedia.org/wiki/Training,_test,_and_validation_sets) can result in a large change in the tree.