Bulletin of Mathematical Sciences and Applications
ISSN: 2278-9634, Vol. 16, pp 70-75
doi:10.18052/www.scipress.com/BMSA.16.70

Submitted: 2016-03-28
Accepted: 2016-07-06
Online: 2016-08-04

Transformed Tree-Structured Regression Method

Gloria Gheno[1, a]

[1]Department of Environmental Sciences, Informatics and Statistics,
Ca' Foscari University of Venice
Dorsoduro 2137, 30123 Venice (ITALY)
and ECLT,
Ca' Minich, San Marco 2940, 30124 Venice, (ITALY)

[a]gloriagheno@libero.it

Keywords: Birckel-Docksum transformation, nonlinearity, random forest, regression tree

Abstract. Many times the response variable is linked linearly to the function of the regressors and to the error term through its function $g(Y)$. For this reason the traditional tree-structured regression methods do not understand the real relationship between the regressors and the dependent variable. I derive a modified version of the most popular tree-structured regression methods to consider this situation of nonlinearity. My simulation results show that my method with regression tree is better than the tree-based regression methods proposed in literature because it understands the true relationship between the regressors and the dependent variable also when it is not possible to divide exactly the error part from the regressors part.

Introduction

Tree-structured regression is a popular choice for modeling and forecasting a continuous response variable as alternative approach to the traditional regressions. It is an appropriated nonparametric approach when the data are high dimensional, i.e. when the number of the regressors is much bigger than the number of the observations, and there is uncertainty about the form in which regressors ought to enter into the model. Many tree-based regression methods exist: those which consider an only tree (for example regression tree [4]) and those which consider more trees (for example bagging [2] and random forest [3]).

The regression tree [4] starts dividing the space of the values of the regressors into 2 subsets choosing as splitting variable the regressor which minimizes better the mean square error. In turn these obtained subsets are divided into other subsets so that a tree is built. The subsets are divided only if there is an improvement, which is measured by the difference between the deviance of the original subset minus the sum of deviances of the subsets obtained dividing it. The subsets, which are not divided, become final subsets. In each final subset, the regression tree fits a constant and this value is the fitted response variable.

To reduce the variance associated to the prediction, Breiman [2] modifies the regression tree proposing the bagging method which estimates and aggregates B trees. In bagging the dataset is divided in 2 subsets: in the first subset the tree is calculated using regression tree, in the second subset the error estimation is calculated. To reduce further the variance associated to the prediction lowering the correlation among the B trees, Breiman [3] modifies the bagging method proposing the random forest. The difference between the random forest and the bagging is the set where the splitting regressors are chosen. The bagging chooses the regressor which minimizes better the mean square error among all regressors, while the random forest chooses the regressor which minimizes better the mean square error among a random set of regressors.

In general, however, a tree method considers the response variable Y as a function $f(X_1,..., X_J)$ of J regressors. Then the error term is linked to the function $f(X_1,..., X_J)$ in additive form. Many times, instead, the response variable is linked linearly to the function of the regressors and to the error term through its function $g(Y)$. In economics, for example, the variable wage is analyzed in its logarithmic form, because it is log(wage) which is linearly linked to the regressors and not the variable wage. If the form of the function $g(Y)$ is known, there are not problems and the tree

methods can be applied substituting to the variable Y its function $g(Y)$. In this work I propose a method to find directly the best function $g(Y)$ which is linked linearly to $f(X_1,..., X_J)$ and to the error term when the form of the function $g(Y)$ is not known.

Transformed tree-structured regression method (TTSR)

Many times the response variable Y is not linked linearly to $f(X_1,..., X_J)$, but it is its function $g(Y)$ which is linked linearly. For this reason I propose a tree method where also the response variable is a function. To transform the response variable, I use Birckel-Doksum transformation [1]:

$$g(Y) = \frac{|Y|^{\tau} sign(Y) - 1}{\tau}$$

where τ is a positive parameter. If the random variable Y is always positive, Birckel-Doksum transformation is equal to Box-Cox modification [1]. Then in my method, which I call the transformed tree-structured regression method (TTSR), $g(Y)$ is equal to $f(X_1,..., X_J)+ \varepsilon$ and this model is estimated using the regression tree or the random forest. To estimate the values of parameter τ, I use a function which considers both the mean square error (MSE) and the R-squared (R^2), because both are often used in tree literature to evaluate the goodness of the tree. The mean square error measures the distance between the statistical model and the data while R-squared is a number between 0 and 1 which indicates how well a statistical model fits the data: if it is equal to 1, the model perfectly fits the data, while if it is equal to 0, the model does not fit the data at all. The best models have small MSE and big R-squared. Recalling that in a traditional tree method if MSE decreases, R^2 increases, I estimate the parameter τ maximizing the following function:

$$L(\lambda, \tau) = -MSE(\tau) + \lambda R^2(\tau)$$

where $\lambda \epsilon [0,1]$ is a parameter which determines the weight of the R-squared. In $L(\lambda,\tau)$ MSE and R-squared are function of parameter τ because they measure the performance of the estimated model $g(Y)= f(X_1,..., X_J)+ \varepsilon$. The best $L(\lambda,\tau)$ is given by small values of MSE and big values of R-squared, indeed the partial derivatives of $L(\lambda,\tau)$ are equal to:

$$\frac{\partial L(\lambda, \tau)}{\partial MSE} = -1 < 0$$
$$\frac{\partial L(\lambda, \tau)}{\partial R^2} = \lambda > 0$$

To maximize this function I use the genetic algorithm [6] [7]. Genetic Algorithm (GA) is a optimization algorithm based on the Darwinian evolutionary ideas of natural selection. GA starts with a randomly generated population of solutions (chromosomes) for an optimization problem [5]. In the second step, GA selects the chromosomes (parents) which it recombines creating new chromosomes (children). In the third step, GA measures the goodness of chromosomes parents and children using the fitness function and it selects the chromosomes which create the next generation [5]. The second and third steps are repeated until the stop criteria is met. I use as fitness function the function $L(\lambda,\tau)$.

Numerical studies

To check small sample performance of the transformed tree-structured regression method, I use 20 simulated datasets s where the response variable is not linked linearly to the regressors and to the error term. The response variable is simulated by the following model:

$$Y = [0.3(0.5X_1 + 0.6X_2 + \varepsilon) + 2]^{1/0.3}$$

where $X_1 \sim N(0,1.3^2)$, $X_2 \sim N(0,0.5^2)$ and the error term $\varepsilon \sim N(0,0.4^2)$. The sample size is equal to 50. Of course, the regressors are between them independent and the error term is independent of the regressors. This relation cannot be written in traditional form , i.e. $Y=f(X1,X2)+\varepsilon$ and for this reason I apply my method. Then the true linear model is:

$$g(Y) = f(X_1, X_2) + \varepsilon = 0.5X_1 + 0.6X_2 + \varepsilon$$

where $g(Y)$ is a normal variable with mean equal to 0 and variance equal to $0.25\text{Var}(X_1) + 0.36\text{Var}(X_2) + \text{Var}(\varepsilon)$. I compare my method, where the mean square error and the R-squared are obtained by the regression tree and by the random forest, with the most important tree-structured regression methods: regression tree [4] and random forest [3]. Of course my method estimates the model $g(Y) = f_{mod\ 2}(X_1,X_2) + \varepsilon_{mod\ 2}$, while the traditional methods estimate a non parametric approximation of the model $Y = f_{mod\ 1}(X_1,X_2) + \varepsilon_{mod\ 1}$. The traditional model estimate an approximation because it is not possible to write Y as a linear function of the function of the regressors and of the error term. To analyze the different tree-structured regression methods, I calculate the ratio, which I call H, between the explained variance and the overall variance. In the theoretical model, the true explained variance of the function $g(Y)$ is equal to

$$0.5^2 Var(X_1) + 0.6^2 Var(X_2) = 0.513$$

while its variance is equal to

$$0.5^2 Var(X_1) + 0.6^2 Var(X_2) + Var(\varepsilon) = 0.672$$

Then in the model which is not approximated, the ratio H is equal to

$$H_{true\ for\ mod\ 2} = \frac{0.5^2 Var(X_1) + 0.6^2 Var(X_2)}{Var\big(g(Y)\big)} = \frac{0.5^2 Var(X_1) + 0.6^2 Var(X_2)}{0.5^2 Var(X_1) + 0.6^2 Var(X_2) + Var(\varepsilon)} = 0.76$$

If I approximate the response variable of the simulated dataset so

$$Y \approx f_{mod\ approx}(X) + \varepsilon_{mod\ approx} = [0.3(0.5X_1 + 0.3X_2) + 2]^{1/0.3} + \varepsilon_{mod\ approx}$$

the true ratio H for the approximated model is equal to

$$H_{true\ for\ mod\ approx} = \frac{Var(f_{mod\ approx}(X_1, X_2))}{Var(Y)} = 0.75$$

To compare these true values with the values of the different methods, I use the absolute relative bias:

$$Absolute\ relative\ bias = \sum_{s=1}^{20} \left| \frac{\widehat{H}_s - H_{true}}{20 H_{true}} \right|$$

Table 1: Comparison in the first simulated dataset

	TTRS (regression tree)	TTRS (random forest)	Random forest	Regression tree
Absolute relative bias	0.084	0.122	0.119	0.087

where $\widehat{H}_s$ is the ratio in sample s between the estimated variance explained by the model and the estimated variance of the function $g(Y)$. Of course for the random forest and for the regression tree $g(Y)$ is equal to Y and H_{true} is equal to $H_{true\ for\ mod\ approx}$, while for TTRS H_{true} is equal to $H_{true\ for\ mod2}$. The results are shown in Table 1. The transformed tree-structured regression method with regression tree is the best method, because its absolute relative bias is the smallest. If I calculate the absolute relative bias of the regression tree and of the random forest using $H_{true\ for\ mod\ 2}$, I find that the absolute relative bias for the regression tree is equal to 0.097 while that for the random forest is equal to 0.131. Then TTRS with regression method is yet the best method because it has the smallest absolute relative bias (0.084<0.097<0.131). Considering, instead, the relative bias I find that all methods produce negative relative biases, i.e. each method underestimates the ratio H giving more importance to the error term.

To analyze further the role of the error term, I calculate the difference between the average ratio H obtained by the TTRS with regression tree and that obtained by the regression tree. This difference is positive. If I repeat the same analysis considering the TTRS with random forest and the traditional random forest, I obtain newly a positive difference. Then in these cases, the traditional tree-based methods underestimate the ratio H most respect to the TTRS methods giving more importance to the error term. The results obtained by the analysis of the average ratio H coincide with those which would be obtained by the linear parametric models. To explain this, I use a simple linear parametric example. I consider this relation

$$Y = (\beta X + \varepsilon)^2 = \beta^2 X^2 + 2\beta X\varepsilon + \varepsilon^2 = f_{mod1}(X) + 2\beta X\varepsilon + \varepsilon^2$$

which can be so rewritten:

$$\sqrt{Y} = \beta X + \varepsilon = f_{mod2}(X) + \varepsilon.$$

I suppose that X and ε are independent normal variables with zero mean and variance equal respectively to σ^2_X and to σ^2_ε. Then in the first model H^{-1} is equal to:

$$H_{mod1}{}^{-1} = 1 + \frac{4\beta^2\sigma_X^2\sigma_\varepsilon^2 + 3\sigma_\varepsilon^4}{3\beta^4\sigma_X^4},$$

while in the second model it is equal to

$$H_{mod2}{}^{-1} = 1 + \frac{\sigma_\varepsilon^2}{\beta^2\sigma_X^2}\ .$$

Then if I estimate the first model, I give more importance to the error term: $H_{mod\ 2} > H_{mod\ 1}$. This can be so proofed:

$$\frac{4\beta^2\sigma_X^2\sigma_\varepsilon^2 + 3\sigma_\varepsilon^4}{3\beta^4\sigma_X^4} > \frac{\sigma_\varepsilon^2}{\beta^2\sigma_X^2} \Rightarrow \frac{4\beta^2\sigma_X^2 + 3\sigma_\varepsilon^2}{3\beta^2\sigma_X^2} > 1 \Rightarrow \frac{4}{3} + \frac{\sigma_\varepsilon^2}{\beta^2\sigma_X^2} > 1\,.$$

Table 2: Comparison in the second simulated dataset

Absolute relative bias using:	TTRS (regression tree)	TTRS (random forest)	Random forest	Regression tree
$H_{true\ for\ mod\ 1}$			0.138	0.093
$H_{true\ for\ mod\ 2}$	0.070	0.128	0.126	0.084

Now I consider another dataset where the division between the error part and the part linked to regressors is more problematic. I use the same X_1, X_2 and ε to generate other 20 datasets, where the response variable is so calculated:

$$Y = 2(0.5X_1 + 0.6X_2 + \varepsilon)^{-1}$$

The true ratio H for the relation $g(Y)=0.5X_1+0.6X_2+\varepsilon$ remain equal to 0.76 as in the previous simulation. If I approximate the response variable of the simulated dataset so

$$Y \approx f_{mod\ approx}(X_1, X_2) + \varepsilon_{mod\ approx} = 2 * (0.5X_1 + 0.3X_2)^{-1} + \varepsilon_{mod\ approx},$$

I find that the true ratio H for this approximation is equal to

$$H_{true\ for\ mod\ approx} = \frac{Var(f_{mod\ approx}(X_1, X_2))}{Var(Y)} = 0.77$$

I calculate the absolute relative bias as in the previous simulation: I use $H_{\text{true for mod approx}}$ and $H_{\text{true for mod 2}}$ for the traditional tree-based methods, while I use only $H_{\text{for mod 2}}$ for the TTRS. The results for the 4 analyzed methods are shown in Table 2. The transformed tree-structured regression method with regression tree is the best method, because its absolute relative bias is the smallest. As for the previous simulation, I calculate the difference between the average ratio H obtained by the TTRS with regression tree and that obtained by the regression tree. This difference is positive. If I repeat the same analysis considering TTRS with random forest and the traditional random forest, this way I obtain a small negative difference.

As last study case, I consider a dataset where the division between the error part and the part linked to the regressors is minus problematic than in the previous simulations. I use the same X_1, X_2 and ε to generate other 20 datasets, where the response variable is so calculated:

$$Y = [exp(0.5X_1 + 0.6X_2 + \varepsilon)]^2 = exp[2(0.5X_1 + 0.6X_2)]exp(2\varepsilon) = exp(X_1 + 1.2X_2)exp(2\varepsilon)$$

Recalling that the approximation of exp(w) around to 0 is equal to (1+w) and that $E[2(0.5X_1 + 0.6X_2)]=0$ and $E(2\varepsilon)=0$, the approximation of the variable Y around 0 becomes equal to

$$Y \approx 1 + 2(0.5X_1 + 1.6X_2) + \varepsilon_{math\ approx} = f_{math\ approx}(X_1, X_2) + \varepsilon_{math\ approx}.$$

Using also an approximation similar to that proposed in the previous simulations, i.e

$$Y \approx exp[2(0.5X_1 + 0.6X_2)] + \varepsilon_{mod\ approx} = f_{mod\ approx}(X_1, X_2) + \varepsilon_{mod\ approx},$$

I find that the true ratios H for these 2 approximations are equal to

Table 3: Comparison in the third simulated dataset

Absolute relative bias using:	TTRS (regression tree)	TTRS (random forest)	Random forest	Regression tree
$H_{true\ for\ mod\ 1}$			0.904	0.720
$H_{true\ for\ math\ approx}$			45.73	40.59
$H_{true\ for\ mod\ 2}$	0.068	0.131	0.324	0.398

$$H_{true\ for\ mod\ approx} = \frac{Var\left(f_{mod\ approx}(X_1, X_2)\right)}{Var(Y)} = \frac{(e^{4*0.513} - 1)e^{4*0.513}}{(e^{4*0.672} - 1)e^{4*0.672}} = 0.27$$

$$H_{true\ for\ math\ approx} = \frac{Var\left(f_{math\ approx}(X_1, X_2)\right)}{Var(Y)} = \frac{4 * 0.513}{(e^{4*0.672} - 1)e^{4*0.672}} = 0.011$$

To calculate the absolute relative bias I use for the traditional tree-based methods $H_{\text{true for mod approx}}$, $H_{\text{true for mod 2}}$ and $H_{\text{true for math approx}}$, while I use only $H_{\text{true for mod 2}}$ for the TTRS. The results for the 4 analyzed methods are shown in Table 3. The TTRS with regression tree is the best method, because its absolute relative bias is the smallest.

By these analyses, I conclude that TTRS with regression method is the best method because it can understand better the real relationship between the regressors and the dependent variable also when it is not possible to divide exactly the error part from the regressors part.

Conclusion

Many times a function of the response variable is linked linearly to the function of the regressors and to the error term. In this case I propose the use of transformed tree-structured regression methods (TTSR) which consider as response variable a Birckel-Docksum transformation of the original response variable. Using simulated datasets, I show that TTSR method with regression tree works better than other tree-structured regression methods when it is not possible to divide the error part from the regressors part.

References

[1] P.J. Bickel, K.A. Doksum, An analysis of transformations revisited. Journal of the American Statistical Association, 76 (1981) 296-231.

[2] L. Breiman, Bagging predictors. Machine learning, 24 (1996) 123-140

[3] L. Breiman, Random Forest. Machine Learning , 45 (2001) 5-32

[4] L. Breiman, J.H. Freidman, R.A. Olshen, and C.J. Stone, Classification and regression trees. Wadsworth, Belmont CA, 1984

[5] J. McCall, Genetic algorithms for modelling and optimisation. Journal of Computational and Applied Mathematics,184 (2005) 205-222

[6] J.H. Holland JH, Adaptation in Natural and Artificial Systems. The University of Michigan Press, Ann Arbor, 1975

[7] S. N. Sivanandam, S.N. Deepa, Introduction to Genetic Algorithms. Springer-Verlag, Berlin, 2007