Visualizable and Interpretable Regression Models With Good Prediction Power[1]

(Published in *IIE Transactions*, vol. 39, 565–579, 2007)

Hyunjoong Kim[2]
Department of Applied Statistics, Yonsei University,
134 Shinchon-dong, Seodaemun-gu, Seoul 120-749, Korea

Wei-Yin Loh[3]
Department of Statistics, University of Wisconsin,
1300 University Avenue, Madison, WI 53706, USA

Yu-Shan Shih[4]
Department of Mathematics, National Chung Cheng University,
Minghsiung Chiayi 621, Taiwan R.O.C.

Probal Chaudhuri
Division of Theoretical Statistics & Mathematics,
Indian Statistical Institute, Calcutta 700035, India

Abstract

Many methods can fit models with higher prediction accuracy, on average, than least squares linear regression. But the models, including linear regression, are typically impossible to interpret or visualize. We describe a tree-structured method that fits a simple but non-trivial model to each partition of the variable space. This ensures that each piece of the fitted regression function can be visualized with a graph or a contour plot. For maximum interpretability, our models are constructed with negligible variable selection bias and the tree structures are much more compact than piecewise-constant regression trees. We demonstrate, by means of a large empirical study involving twenty-seven methods, that the average prediction accuracy of our models is almost as high as that of the most accurate "black-box" methods from the statistics and machine learning literature.

Key words and phrases: Machine learning, piecewise linear, regression tree, selection bias.

[1]The authors are grateful to two referees for their comments.

[2]Kim's research was partially supported by Grant No. R01-2005-000-11057-0 from the Basic Research Program of the Korea Science and Engineering Foundation.

[3]Corresponding author, phone: 608-262-2598, fax: 608-262-0032, email: loh@stat.wisc.edu. Loh's research was partially supported by the National Science Foundation under grant DMS-0402470 and by the U.S. Army Research Laboratory and the U.S. Army Research Office under grant W911NF-05-1-0047.

[4]Shih's research was partially supported by a grant from the National Science Council of Taiwan.

1 Introduction

Box (1979) wrote, "All models are wrong but some are useful." This statement is unquestionably true, but it raises the question: Useful for what? There are two ways in which a model can be useful—it can improve our understanding of the system generating the data or it can make accurate predictions of future observations. For example, linear models for designed factorial experiments are useful because the terms they contain may be interpreted as main and interaction effects. On the other hand, accurate weather prediction models are useful even if they are hard to interpret.

There are many applications, however, where traditional statistical models are useless for prediction and for interpretation. One such is the study on house prices in the greater Boston area in 1970 reported in Harrison and Rubinfeld (1978) and made famous by Belsley, Kuh and Welsch (1980). There are 506 observations on a variety of variables, with each observation pertaining to one census tract. The goal of the study was to build a regression model for the median house price (MEDV) and to use it to estimate the "marginal-willingness-to-pay for clean air," namely, the effect of nitrogen oxide concentration (NOX). Table 1 lists the predictor variables. After transforming some variables to satisfy normal-theory assumptions, Harrison and Rubinfeld (1978) obtained the fitted model shown in Table 2. Note that because the whole population is represented in the data, there is nothing to predict. In particular, the t-statistics do not have their usual statistical meaning.

Table 1: Variables in the Boston housing data

Var	Definition	Var	Definition
MEDV	median value in $1000	AGE	% built before 1940
DIS	distance to employment centers	CRIM	per capita crime rate
RAD	accessibility to radial highways	ZN	% land zoned for lots
INDUS	% nonretail business	TAX	property tax/$10,000
CHAS	1 on Charles River, 0 else	PT	pupil/teacher ratio
NOX	nitrogen oxide conc. (p.p.10^9)	B	(% black - 63)^2/10
RM	average number of rooms	LSTAT	% lower-status pop.

Table 2: Least squares fit for $\log(\texttt{MEDV})$. The columns labeled by β, t, and ρ give the estimated regression coefficients, t-statistics, and correlation between $\log(\texttt{MEDV})$ and the corresponding X variable.

X	β	t	ρ	X	β	t	ρ
Intercept	4.6	29.5		RM^2	6.3E-3	4.8	0.6
log(LSTAT)	-3.7E-1	-14.8	-0.8	B	3.6E-4	3.5	0.4
CRIM	-1.2E-2	-9.5	-0.5	TAX	-4.2E-4	-3.4	-0.6
PT	-3.1E-2	-6.2	-0.5	CHAS	9.1E-2	2.8	0.2
log(DIS)	-1.9E-1	-5.7	0.4	AGE	9.1E-5	0.2	-0.5
NOX^2	-6.4E-1	-5.6	-0.5	ZN	8.0E-5	0.2	0.4
log(RAD)	9.6E-2	5.0	-0.4	INDUS	2.4E-4	0.1	-0.5

We may hope that the model can explain the effects of the predictor variables on the response. For example, the sign associated with the coefficient for NOX^2 suggests that it has a negative effect on MEDV. Similarly, the negative coefficient for $\log(\texttt{DIS})$ leads to the conclusion that MEDV is negatively associated with DIS. Table 2 shows, however, that the correlation between $\log(\texttt{DIS})$ and $\log(\texttt{MEDV})$ is positive! Another example is RAD, which has a positive regression coefficient but a negative correlation with MEDV. Of course, these apparent contradictions are easy to explain. First, a regression coefficient quantifies the residual effect

of the predictor after the linear effects of the other predictors in the model have been accounted for. Second, the correlation between a predictor and the response measures their linear association, ignoring the other predictors. Nevertheless, the contradictions in signs are not intuitive.

Can we construct models that are more interpretable and that also fit the data well? Since a model that involves a single predictor variable is easiest to interpret because the fitted function can be graphed, one solution is to employ the best single-predictor model. Unfortunately, because such a model does not incorporate the information contained in the other predictors, it may not fit the data as well as a model that uses more than one predictor. Further, a single-predictor model reveals nothing about the joint effect of all the predictors.

The goal of this paper is to study an alternative approach that (i) retains the clarity and ease of interpretation of relatively simple models, (ii) allows expression of the joint effect of several predictors, and (iii) yields models with higher average prediction accuracy than the traditional multiple linear regression model. We accomplish this by fitting simple models to partitions of the dataset and sample space. One such model for the Boston data is shown by the tree structure in Figure 1. The predictor space is split into three rectangular partitions. Within each partition, the best single predictor variable is selected to fit a linear model to `MEDV`. Notice that, unlike the Harrison-Rubinfeld model, we can directly model `MEDV` in terms of the original predictors without needing any transformations.

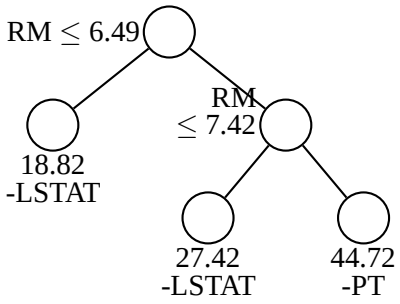

Figure 1: Piecewise simple linear regression tree for the Boston data. The sample mean `MEDV` value and the best linear predictor is printed beneath each leaf node, together with the sign of its coefficient. At each split, a case goes down the left branch if and only if the associated inequality is satisfied.

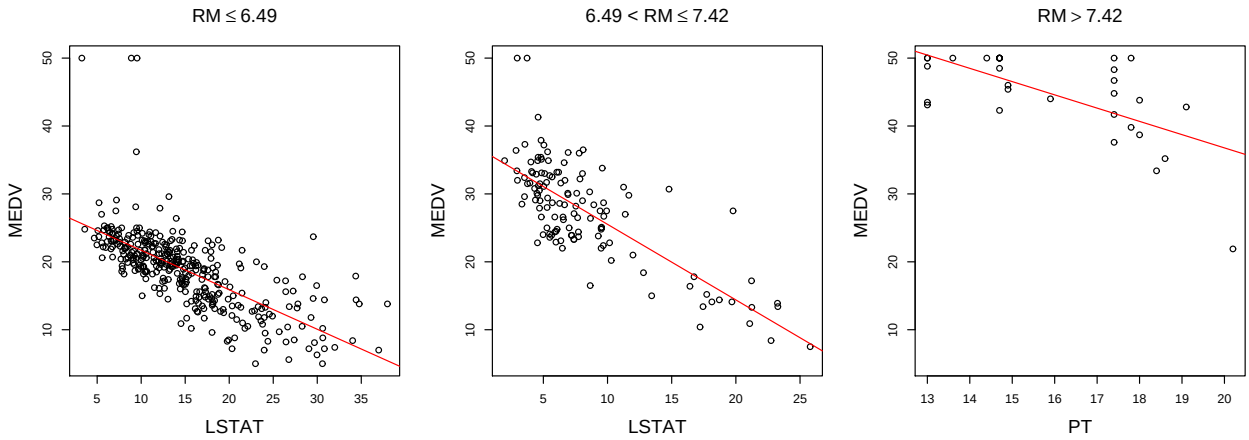

Figure 2: Data and fitted models in three leaf nodes of the tree in Figure 1

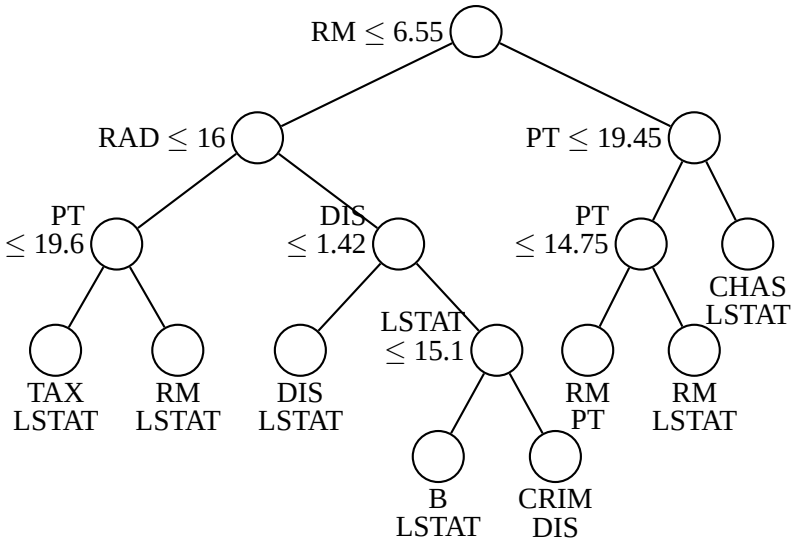

Figure 3: Piecewise two-regressor linear regression tree for MEDV (Boston data), with selected regressor variables beneath each leaf node

Figure 2 displays the data and fitted functions in the three partitions. The graphs indicate that LSTAT has a large negative effect on house price, except in census tracts with large houses (right panel) where PT is a stronger linear predictor. As expected, MEDV tends to increase with RM. These conclusions are consistent with the signs of the coefficients of $\log(\texttt{LSTAT})$ and $\texttt{RM}^2$ in the Harrison-Rubinfeld model.

Besides a piecewise single-regressor model, a piecewise two-regressor model can also be used to reveal more insight into the data. The tree structure for the latter is presented in Figure 3, with the selected regressors printed beneath the leaf nodes. By utilizing only two regressor variables in each node of the tree, we can employ shaded contour plots to display the fitted functions and the data points. These plots are shown in Figure 4, with lighter shades corresponding to higher values of MEDV. Note that some of the contour lines are not parallel; this is due to truncation of the predicted values, as explained by the algorithm in Section 2. We observe that the higher-priced census tracts tend to have high values of RM and low values of LSTAT. The lowest-priced tracts are mostly concentrated in one leaf node (bottom left panel in dark gray) with below average values of RM and DIS, and above average values of RAD, LSTAT, and CRIM. Although the regression coefficients in each leaf node model are afflicted by the problems of interpretation noted earlier, we do not need their values for a qualitative analysis. The contour plots convey all the essential information.

How well do the tree models fit the data compared to the Harrison-Rubinfeld model? Figure 5 plots the fitted versus observed values of MEDV. The piecewise two-regressor model clearly fits best of all. Notice the lines of points on the right edges of the graphs for the Harrison-Rubinfeld and the one-regressor tree models. They are due to the observed MEDV values being truncated at $50,000 (Gilley and Pace, 1996) and the inability of these two models to fit them satisfactorily. Our two-regressor model has no trouble with these points.

The rest of this article is organized as follows. Section 2 describes our regression tree algorithm. Section 3 analyzes another well-known dataset and compares the results with that of human experts. We take the opportunity there to highlight the important problem of selection bias. In Section 4 we compare the prediction accuracy of twenty-seven algorithms from the statistical and machine learning literature on fifty-two real datasets. The results show that some machine learning methods have very good accuracy and that our methods are quite competitive. We prove an asymptotic consistency result in Section 5 to lend theoretical

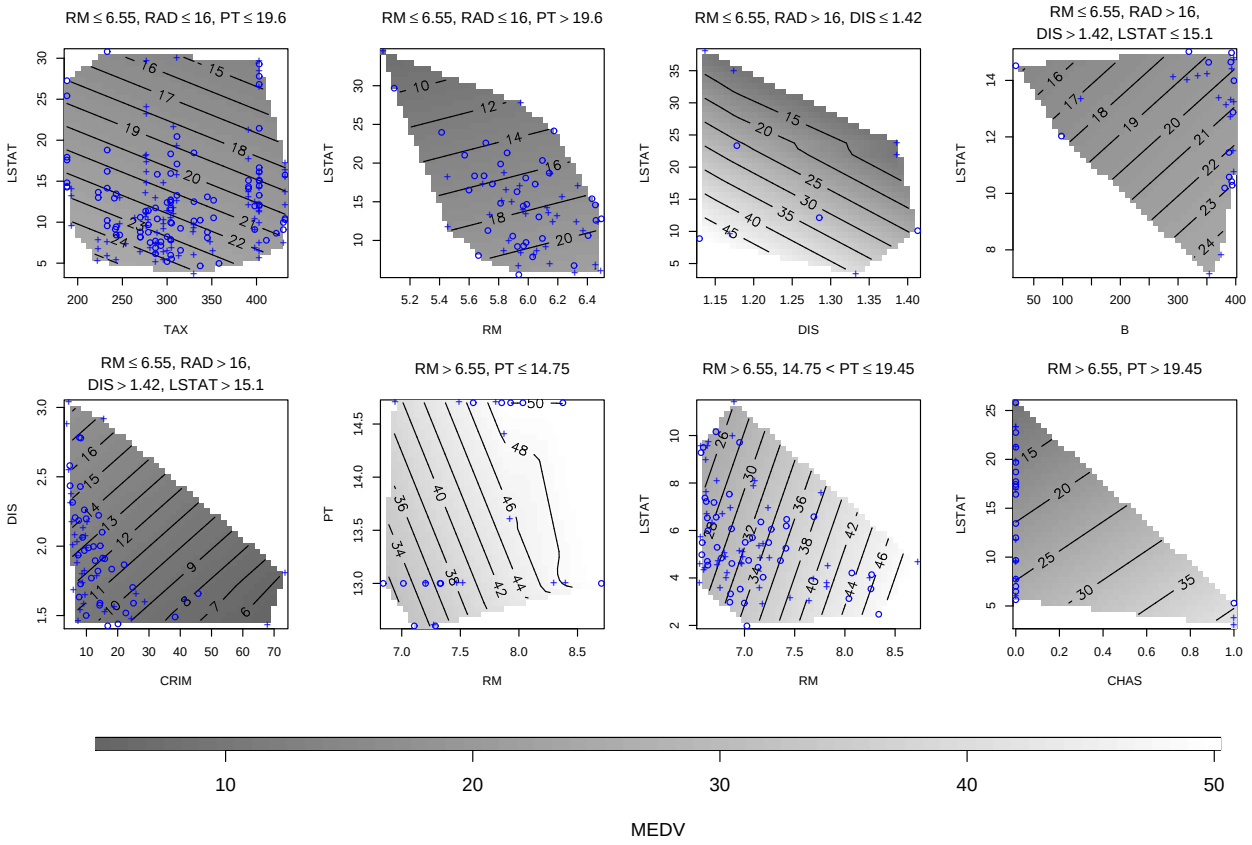

Figure 4: Contour plots of the fitted functions in the leaf nodes of the tree in Figure 3. Data points with positive and negative residuals are marked with + and o symbols, respectively.

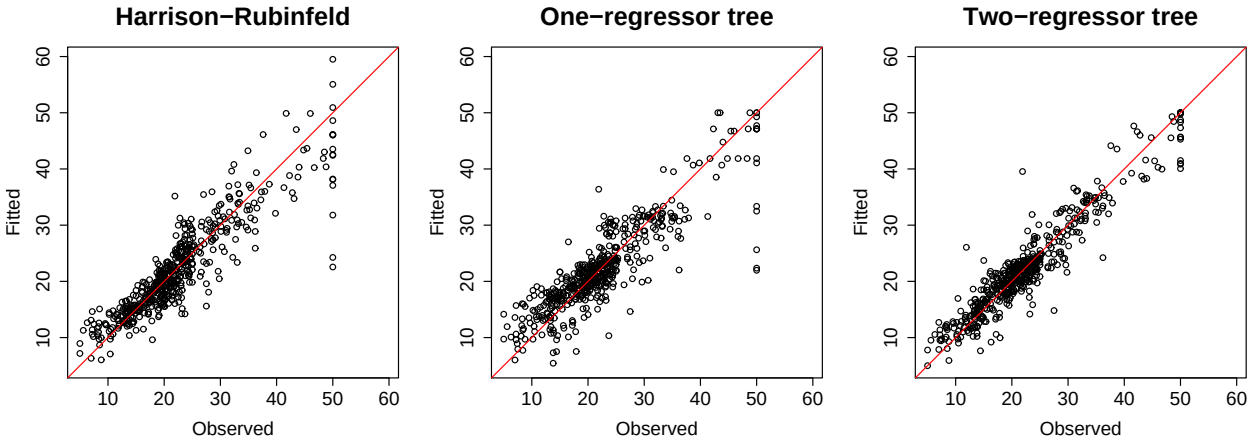

Figure 5: Fitted versus observed values for the Harrison-Rubinfeld and the piecewise one- and two-regressor models for the Boston data

support to the empirical findings and close with some remarks in Section 6.

2 Regression tree method

Our algorithm is an extension of the GUIDE algorithm (Loh, 2002), which fits a constant or a multiple linear model at each node of a tree. The only difference is that we now use stepwise linear regression instead. The number of linear predictors permitted at each node may be restricted or unrestricted, subject to the standard F-to-enter and F-to-remove thresholds of 4.0 (Miller, 2002). A one- or two-regressor tree model is obtained by restricting the number of linear predictors to one or two, respectively. We present here the recursive sequence of operations for a two-regressor tree model; the method for a one-regressor tree model is similar.

1. Let t denote the current node. Use stepwise regression to choose two quantitative predictor variables to fit a linear model to the data in t.

2. Do not split a node if its model $R^2 > 0.99$ or if the number of observations is less than $2n_0$, where n_0 is a small user-specified constant. Otherwise, go to the next step.

3. For each observation, define the class variable $Z = 1$ if it is associated with a positive residual. Otherwise, define $Z = 0$.

4. For each predictor variable X:

   (a) Construct a $2 \times m$ cross-classification table. The rows of the table are formed by the values of Z. If X is a categorical variable, its values define the columns, i.e., m is the number of distinct values of X. If X is quantitative, its values are grouped into four intervals at the sample quartiles and the groups constitute the columns, i.e., $m = 4$.

   (b) Compute the significance probability of the chi-squared test of association between the rows and columns of the table.

5. Select the X with the smallest significance probability to split t. Let t_L and t_R denote the left and right subnodes of t.

   (a) If X is quantitative, search for a split of the form $X \leq x$. For each x such that t_L and t_R each contains at least n_0 observations:

      i. Use stepwise regression to choose two quantitative predictor variables to fit a two-regressor model to each of the datasets in t_L and t_R.

      ii. Compute S, the total of the sums of squared residuals in t_L and t_R.

   Select the smallest value of x that minimizes S.

   (b) If X is categorical, search for a split of the form $X \in C$, where C is a subset of the values taken by X. For each C such that t_L and t_R each has at least n_0 observations, calculate the sample variances of Z in t_L and t_R. Select the set C for which the weighted sum of the variances is minimum, with weights proportional to sample sizes.

6. After splitting has stopped, prune the tree using the algorithm described in Breiman, Friedman, Olshen and Stone (1984, Sec. 8.5) with ten-fold cross-validation (CV). Let E_0 be the smallest CV estimate of prediction mean square error (MSE) and let α be a positive number. Select the smallest subtree whose CV estimate of MSE is within α times the standard error of E_0. We use the default value of $\alpha = 0.5$ here and call this the 0.5-*SE rule*. To avoid large prediction errors caused by extrapolation, truncate all predicted values so that they lie within the range of the training sample data values in their nodes. The non-parallel contour lines in some of the plots in Figure 4 are the result of this truncation.

Our split selection approach is uniquely different from that of CART (Breiman *et al.*, 1984) and M5 (Quinlan, 1992), two other regression tree algorithms. CART constructs piecewise *constant* trees only and it searches for the best variable to split and the best split point simultaneously at each node. This requires the evaluation of all possible splits on every predictor variable. Thus, if there are K quantitative predictor variables each taking M distinct values at a node, $K(M-1)$ splits have to be evaluated. To extend the CART approach to piecewise linear regression, two linear models must be fitted for each candidate split. This means that $2K(M-1)$ regression models must be computed before a split is found. The corresponding number of regression models for K categorical predictors each having M distinct values is $2K(2^{M-1}-1)$. Clearly, the computational cost grows rapidly with K and M.

Our approach avoids the computational problem by separating split variable selection from split point selection. To select a variable for splitting, only one regression model is fitted (step 1 of the algorithm). If the selected variable is quantitative with M distinct values, split set selection requires only $2(M-1)$ models to be fitted (step 5a). On the other hand, if the selected variable is categorical, no regression fitting is needed to find the set of split values (step 5b).

M5 uses a hybrid strategy to build a piecewise linear model. First it constructs a large piecewise constant tree using exhaustive search to minimize a weighted sum of standard deviations. Then it prunes the tree using a heuristic argument instead of cross-validation. A single linear model is fitted to each node during pruning.

M5 also treats categorical variables differently. Our piecewise one and two-regressor models use categorical variables for split selection only; they do not use them as regressors in the linear models. M5, on the other hand, first converts each categorical variable into a vector of zeros and ones and then treats the elements of the vector as quantitative variables for split selection and for regression modeling.

A consequence of these differences in approach is that our method possesses an important property that CART and M5 do not, namely, conditional unbiasedness in split variable selection. We say that a selection method is *unbiased* if, under the assumption that the predictor variables are statistically independent of the response variable, each predictor has the same chance of being selected. Unbiasedness is desirable because even a small amount of selection bias in a tree can lead to erroneous or inexplicable conclusions. The reason our method is unbiased can be traced to step 4 of the algorithm where selection of a variable to split a node is based on contingency table analyses of the residual distribution versus the distributions of the predictor variables. Suppose X_1 and X_2 are the regressor variables in a two-regressor model. If the other predictor variables are independent of the response variable, they will also be independent of the residuals. Hence conditionally on X_1 and X_2 being the selected regressors, all the other variables have the same chance of being selected to split the node. In contrast, since CART and M5 are based on exhaustive search, their split selection methods are biased toward variables that allow more splits, particularly categorical variables with many distinct values. We demonstrate this with an example in the next section.

3 Baseball data

This example utilizes a well-known baseball dataset provided by the American Statistical Association Section on Statistical Graphics for its 1988 data exposition. The data consist of the 1987 opening day salaries and various career and 1986 performance statistics of 263 major league baseball hitters (see Table 3). The purpose of the exposition was to invite statisticians to analyze the data and answer the question, "Are players paid according to their performance?" Fifteen teams took up the challenge and their analyses were published in the conference proceedings.

Hoaglin and Velleman (1995) give a critique of the published results. Defining as "best" the models that are "most parsimonious, most interpretable, and best fitting," they conclude that a log transformation of `Salary` is most successful, and that the best predictor variables are `Yrs`, `Runcr`/`Yrs`, and a 1986 performance measure. They also find seven outliers in the data. Six are due to errors in the data and are

Table 3: Predictor variables and their definitions for the baseball data

Bat86	#times at bat, 1986	Rbcr	#runs batted in, career
Hit86	#hits, 1986	Wlkcr	#walks, career
Hr86	#home runs, 1986	Leag86	league, end 1986 (2 values)
Run86	#runs, 1986	Div86	division, end 1986 (2 values)
Rb86	#runs batted in, 1986	Team86	team, end 1986 (24 values)
Wlk86	#walks, 1986	Pos86	position, 1986 (23 values)
Yrs	#years in major leagues	Puto86	#put outs, 1986
Batcr	#times at bat, career	Asst86	#assists, 1986
Hitcr	#hits, career	Err86	#errors, 1986
Hrcr	#home runs, career	Leag87	league, start 1987 (2 values)
Runcr	#runs, career	Team87	team, start 1987 (24 values)

Table 4: Hoaglin-Velleman model for $\log(\texttt{Salary})$

X	β	t
Intercept	3.530	31.1
Runcr/Yrs	0.016	9.4
$\sqrt{\texttt{Run86}}$	0.083	4.1
min[$(\texttt{Yrs}-2)_+$,5]	0.346	22.8
$(\texttt{Yrs}-7)_+$	-0.041	-4.4

omitted from their analysis. The seventh outlier (Pete Rose) is retained because it is not erroneous. The model fitted to the 257 cases is given in Table 4 (Hoaglin and Velleman use base-10 log whereas we use natural log here).

Although the t-statistics in Table 4 are all highly significant, they again do not have their usual statistical meaning because the dataset is essentially the whole population of major league hitters for 1987. Besides, even if the data were a random sample, the t-statistics are expected to be inflated by the process of variable selection and transformation. Despite this, the variables in the model make sense: Yrs accounts for experience, Runcr/Yrs for productivity rate, and Run86 for 1986 performance. What is difficult to explain is the negative coefficient for $(\texttt{Yrs}-7)_+$. It appears to suggests that the players were penalized for experience beyond seven years.

Using the data from all 263 hitters, but with the errors replaced by the correct values listed in Hoaglin and Velleman (1995), our piecewise one-regressor model tree has three leaf nodes. It splits twice on Yrs, as shown on the left side of Figure 6. If $\texttt{Yrs} \leq 3$, the best predictor is Runcr; if $3 < \texttt{Yrs} \leq 6$, the

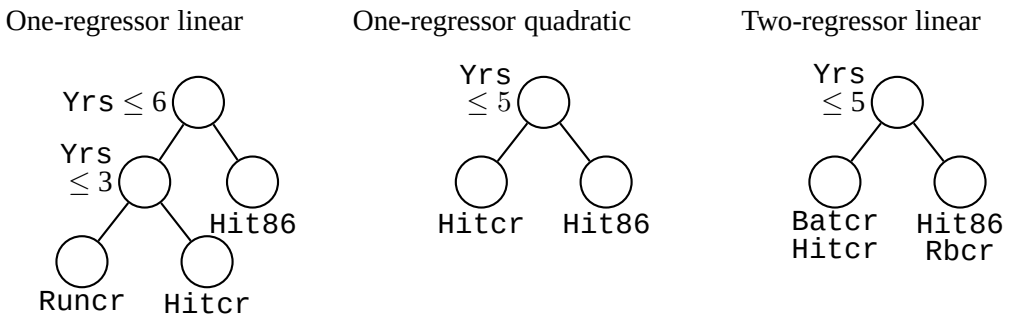

Figure 6: Three regression tree models for the baseball data

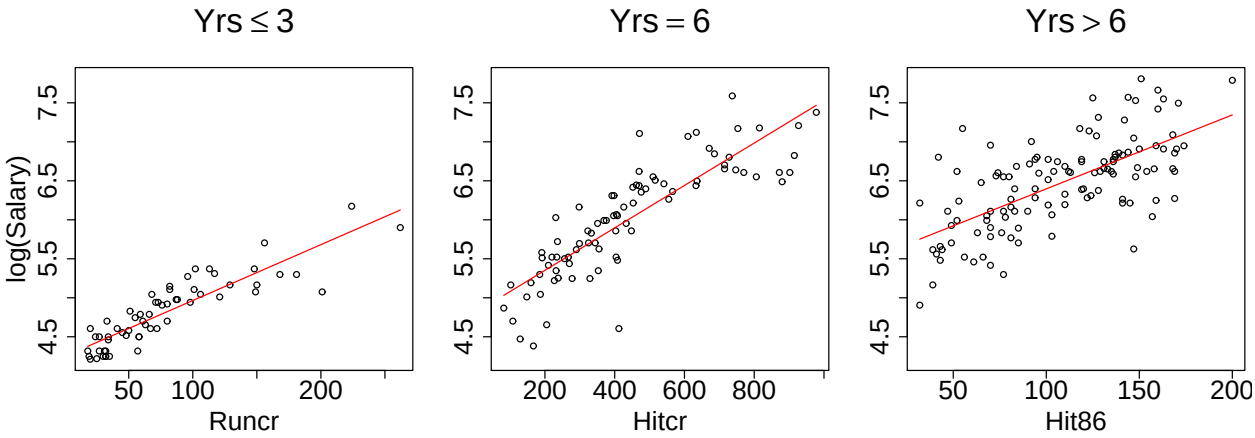

Figure 7: Fitted functions and data for the one-regressor linear tree model for the baseball data

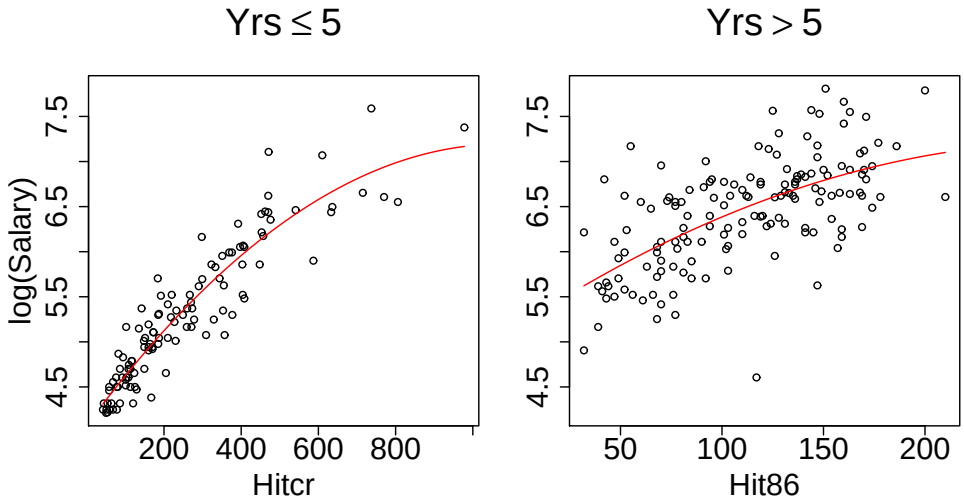

Figure 8: Fitted functions and data for the one-regressor quadratic tree model for the baseball data

best predictor is `Hitcr`; otherwise the best predictor is `Hit86`. Our model is thus quite similar to that of Hoaglin and Velleman. While the latter is a three-piece model divided along `Yrs` at 2 and 7, ours divides `Yrs` at 3 and 6. Our model similarly suggests that salary depends on career performance for `Yrs` ≤ 6; beyond 6 years, 1986 performance is more important. Figure 7 shows the data and fitted functions in the three leaf nodes of the tree.

The split of `Yrs` into three intervals suggests a curvature effect. We can better model this with a piecewise one-regressor quadratic model. This tree is shown in the middle of Figure 6. It is simpler, with only one split—again on `Yrs`. The best predictor variable is `Hitcr` if `Yrs` ≤ 5 and `Hit86` otherwise. Again, salary tends to increase with career performance for the junior players and with 1986 performance for the senior players. Plots of the fitted curves are shown in Figure 8.

The tree for our piecewise two-regressor linear model is displayed on the right side of Figure 6. It has the same structure as that for the one-regressor quadratic model, except that two variables are selected as linear regressors in each leaf node. We observe from the contour plots of the fitted functions in Figure 9 that `Batcr` and `Hitcr` are the most important predictor variables for players with five or fewer years of

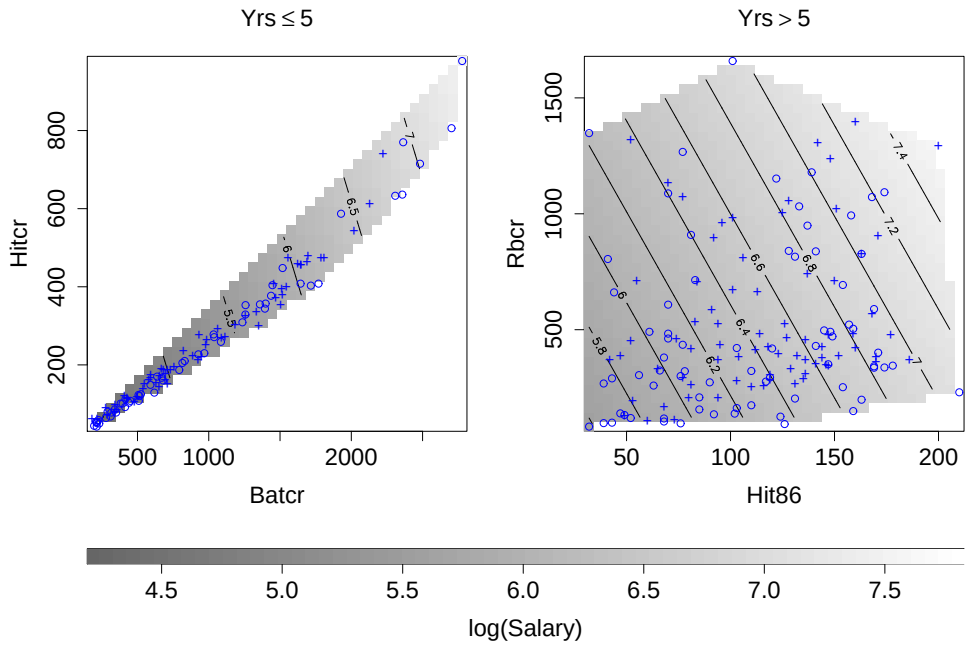

Figure 9: Data and contour plots of the piecewise two-regressor model for the baseball data. Points associated with positive and negative residuals are denoted by `+` and `o`, respectively.

experience. As may be expected, the plot shows that these two variables are highly correlated. Further, except for a couple of players, most of the junior ones have lower salaries. For players with more than five years of experience, the most important predictors are `Hit86` and `Rbcr`. It is noteworthy that the sample correlation between these two variables here is quite low at 0.18. Many of the highest-paid players belong to this group. Figure 10 shows that the fit of this model is as good as that of the Hoaglin-Velleman model.

Figure 11 shows the CART and M5 models for these data. The M5 tree is obtained using the WEKA (Witten and Frank, 2000) implementation. The similarity in the splits is due to CART and M5 both constructing piecewise constant regression trees prior to pruning. CART minimizes residual sum of squares but M5 minimizes a weighted sum of standard deviations. This is the reason for the different split values at the root node.

Given the prominent role of `Yrs` in our models and that of Hoaglin and Velleman, it may seem odd that this variable is completely absent from the CART tree. (`Yrs` is also not used to split the M5 tree, but it is used as linear predictor in the leaf node models.) CART and M5 split first on `Batcr` instead. Since the latter is positively correlated with `Yrs`, we conjecture that `Batcr` is acting as a surrogate to `Yrs`. The reason it is chosen is most likely due to selection bias—`Batcr` has 256 permissible splits while `Yrs` has only 20. Another possible manifestation of the bias appears in the CART split on the categorical variable `Team86`, which neither the Hoaglin-Velleman nor our tree models find important. It turns out that `Team86` has 22 distinct categories at that node. Therefore it yields $2^{21} - 1 = 2{,}097{,}151$ permissible splits. As a result, the residual sum of squares can be reduced much more by splitting on `Team86` than on an ordered quantitative variable. M5 is similarly affected by this selection bias. It uses `Team86` and `Team87` as split variables in the lower branches of the tree, but these branches are removed by pruning. Nonetheless, some tell-tale evidence involving these variables can be found in the piecewise linear regression functions L_1–L_5 in the leaf nodes. For example,

$$10^4 L_1 \;=\; 41300 + 97.2\,\texttt{Yrs} - 29.2\,\texttt{Bat86} + 18.8\,\texttt{Batcr} + 112\,\texttt{Hit86} - 44.7\,\texttt{Hitcr}$$

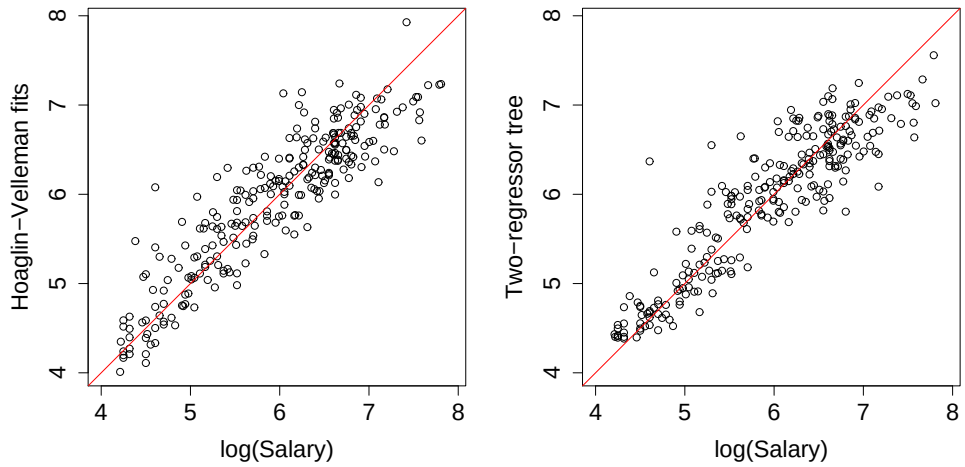

Figure 10: Fitted versus observed $\log(\texttt{Salary})$ for the baseball data

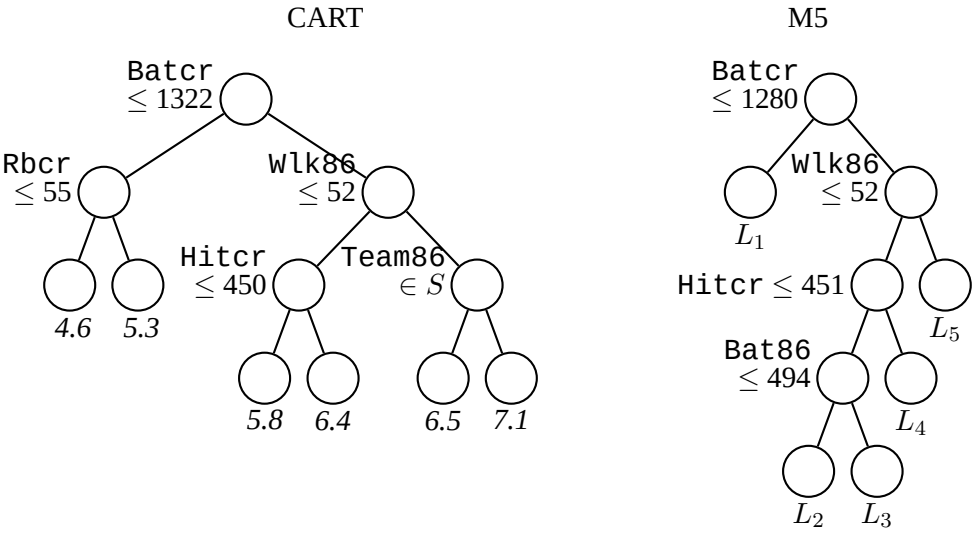

Figure 11: CART piecewise constant and M5 piecewise multiple linear models for the baseball data. The fitted values for the CART model are given in italics beneath the leaf nodes. S consists of the teams: Atlanta, California, Cincinnati, Cleveland, Detroit, Houston, Los Angeles, Oakland, Pittsburgh, San Diego, San Francisco, and Seattle. L_1–L_5 are linear functions.

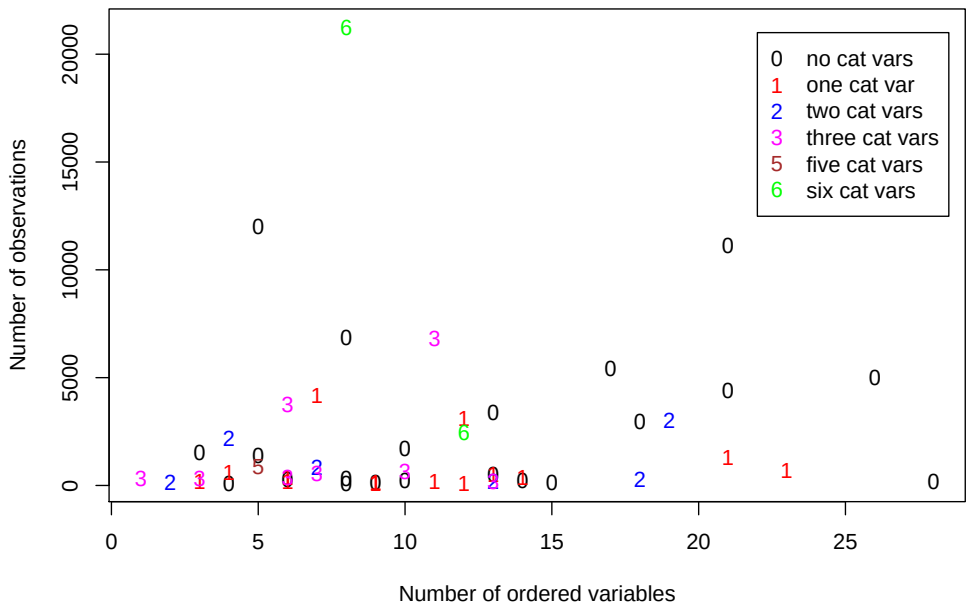

Figure 12: Sample size versus number of predictor variables for 52 datasets. Plot symbol indicates the number of categorical variables.

$$
\begin{aligned}
&+ 17.1\,\texttt{Wlk86} + 2.2\,\texttt{Hrcr} + 0.456\,\texttt{Runcr} + 35.6\,\texttt{Rbcr} - 1.68\,\texttt{Wlkcr} \\
&+ 63\,I(\texttt{Team86} \in \{\texttt{Atl}, \texttt{Bal}, \texttt{Bos}, \texttt{Cal}, \texttt{Chi}, \texttt{Cin}, \texttt{Cle}, \texttt{Det}, \texttt{Hou}, \texttt{LA}, \texttt{NY}, \texttt{Oak}, \texttt{Tor}\}) \\
&+ 418\,I(\texttt{Team86} = \texttt{StL}) + 875\,I(\texttt{Team87} \in \{\texttt{Atl}, \texttt{Bal}, \texttt{Bos}, \texttt{Chi}, \texttt{LA}, \texttt{NY}, \texttt{Oak}, \texttt{StL}, \texttt{Tor}\}) \\
&+ 534\,I(\texttt{Team87} \in \{\texttt{Cal}, \texttt{Cin}, \texttt{Cle}, \texttt{Det}, \texttt{Hou}, \texttt{KC}, \texttt{Mil}, \texttt{Min}, \texttt{Mon}, \texttt{Phi}, \texttt{SD}, \texttt{SF}, \texttt{Tex}\}).
\end{aligned}
$$

Notice the counter-intuitive change in signs between the coefficients of `Bat86` and `Batcr` and between those of `Hit86` and `Hitcr`. Since the M5 tree contains five linear regression functions, it is five times as hard to interpret as a traditional multiple linear regression model.

4 Prediction accuracy

We now put aside the issue of model interpretation and consider how our methods compare against other methods in terms of prediction accuracy when applied to real data. Since there are few published empirical studies comparing statistical with machine learning regression methods, we include some of the most well-known algorithms from each discipline. The results reported here are obtained using twenty-seven algorithms and fifty-two datasets.

4.1 Datasets

The datasets are listed in Table 5 together with information on sample sizes, numbers of quantitative and categorical predictor variables, and their sources. Sample size ranges from 96 to 21252, number of quantitative predictor variables from 1 to 28, and number of categorical variables from 0 to 6. All binary predictor variables are treated as quantitative. Figure 12 summarizes the information in a graph. The datasets mostly come from books and journal articles, although several are from Statlib (`http://lib.stat.cmu.edu/`) and the UCI data repository (Blake and Merz, 1998). Observations with missing or incorrect values are removed.

Table 5: Datasets. N denotes the number of training cases, N^* the number of test cases (if any), Q the number of quantitative predictors, and C the number of categorical predictors (category sizes in parentheses).

Name	N	N^*	Q	C	Source
Abalone	4177		7	1 (3)	UCI
Ais	202		11	1 (9)	Cook and Weisberg (1994)
Alcohol	2467		12	6 (3,3,3,4,4,6)	Kenkel and Terza (2001)
Amenity	3044		19	2 (3,4)	Chattopadhyay (2003)
Attend	838		7	2 (3,29)	Cochran (2002)
Baseball	263		18	2 (23,24)	Statlib
Baskball	96		4	0	Simonoff (1996)
Boston	506		13	0	Belsley, Kuh and Welsch (1980)
Boston2	506		13	1 (92)	Belsley, Kuh and Welsch (1980)
Budget	1729		10	0	Bollino, Perali and Rossi (2000)
Cane	3775		6	3 (14,15,24)	Denman and Gregory (1998)
Cardio	375		6	3 (7,8,12)	Bryant and Smith (1996)
College	694		23	1 (3)	Statlib
County	3114		12	1 (46)	Harrell (2001)
Cps	534		7	3 (3,3,6)	Berndt (1991)
Cps95	21252	42504	8	6 (2,3,4,5,7,9)	ftp.stat.berkeley.edu/pub/datasets/fam95.zip
Cpu	209		6	1 (30)	UCI
Deer	654		10	3 (2,6,7)	Onoyama, Ohsumi, Mitsumochi and Kishihara (1998)
Diabetes	375		14	1 (3)	Harrell (2001)
Diamond	308		1	3 (3,5,6)	Chu (2001)
Edu	1400		5	0	Martins (2001)
Engel	11986	11986	5	0	Delgado and Mora (1998)
Enroll	258		6	0	Liu and Stengos (1999)
Fame	1318		21	1 (7)	Cochran (2000)
Fat	252		14	0	Penrose, Nelson and Fisher (1985)
Fishery	6806		11	3 (3,5,6)	Fernandez, Ley and Steel (2002)
Hatco	100		12	1 (3)	Hair, Anderson, Tatham and Black (1998)
Houses	6880	13760	8	0	Pace and Barry (1997)
Insur	2182		4	2 (7,9)	Hallin and Ingenbleek (1983)
Labor	2953		18	0	Aaberge, Colombino and Strom (1999)
Labor2	5443	5443	17	0	Laroque and Salanie (2002)
Laheart	200		13	3 (4,4,5)	Afifi and Azen (1979)
Medicare	4406		21	0	Deb and Trivedi (1997)
Mpg	392		6	1 (3)	UCI
Mpg2001	852		5	5 (3,3,5,12,42)	www.fueleconomy.gov
Mumps	1523		3	0	Statlib
Mussels	201		3	1 (5)	Cook (1998)
Ozone	330		8	0	Breiman and Friedman (1988)
Pole	5000	10000	26	0	Weiss and Indurkhya (1995)
Price	159		15	0	UCI
Rate	144		9	0	Lutkepohl, Terasvirta and Wolters (1999)
Rice	171		13	2 (3,3)	Horrace and Schmidt (2000)
Scenic	113		9	1 (4)	Neter, Kutner, Nachtsheim and Wasserman (1996)
Servo	167		2	2 (5,5)	UCI
Smsa	141		9	1 (4)	Neter, Kutner, Nachtsheim and Wasserman (1996)
Spouse	11136	11136	21	0	Olson (1998)
Strike	625		4	1 (18)	Statlib
Ta	324		3	3 (3,30,40)	Authors
Tecator	215		10	0	Statlib
Tree	100		8	0	Rawlings (1988)
Triazine	186		28	0	Torgo (1999)
Wage	3380		13	0	Schafgans (1998)

Table 6: Plot symbols of the twenty-seven algorithms

`Cart`	CART piecewise constant	`Mc`	M5 piecewise constant
`Cr`	CUBIST rule-based model	`Mcb`	Bagged `Mc`
`Ci`	`Cr` and nearest-neighbor (composite)	`Mm`	M5 piecewise multiple linear
`Crb`	Boosted `Cr` (committee model)	`Mmb`	Bagged `Mm`
`Gc`	GUIDE piecewise constant	`mars`	MARS
`Gl`	GUIDE piecewise simple linear	`mart`	MART
`Gq`	GUIDE piecewise simple quadratic	`nnet`	Neural network
`Gm`	GUIDE piecewise multiple linear	`pol`	POLYMARS
`Gs`	GUIDE piecewise stepwise linear	`ppr`	Projection pursuit regression
`Gs2`	GUIDE two-regressor stepwise linear	`Rc`	RT piecewise constant
`Gf2`	GUIDE two-regressor forward linear	`Rm`	RT piecewise multiple linear
`gam`	Generalized additive model	`Rp`	RT piecewise partial linear
`lad`	Least absolute deviations regression	`rreg`	Huber's robust regression
`lr`	Least squares linear regression		

For the six datasets accompanied by test sets, we apply each algorithm to the training set and compute the prediction MSE of the fitted model on the test set. For the other datasets, we use ten-fold cross-validation to estimate the prediction MSE. That is, we first randomly divide each dataset into ten roughly equal-sized subsets. Then we set aside one subset in turn, pool the observations in the other nine subsets, apply each algorithm to the combined data, and compute the prediction MSE of the fitted model on the set-aside subset. The average of the ten results yields the cross-validation estimate.

4.2 Algorithms

Table 6 lists the twenty-seven algorithms. Unless stated otherwise, each algorithm is used with its default parameter values. For those algorithms that cannot directly deal with categorical predictor variables, we follow the standard practice of converting them to 0-1 dummy vectors. Each vector component is then treated as a quantitative predictor variable.

CART. Piecewise constant regression tree (Breiman *et al.*, 1984). CART is a registered trademark of California Statistical Software, Inc. We use version 4 of the Windows implementation (Steinberg and Colla, 1995), with 10-fold cross-validation and the default 0-SE rule. The minimal node size is 10 except for the `Cps95` dataset where the value is changed to 100 because of the program's memory limitations.

CUBIST. A rule-based algorithm due to R. Quinlan (`www.rulequest.com/cubist-info.html`). We use Release 1.10. Three type of models are studied: rule-based only (`Cr`), composite (`Ci`), and committee (`Crb`) with five members. The `Ci` model combines `Cr` with a nearest-neighbor method. `Crb` is a boosted version of `Cr`.

GAM. Generalized additive model (Hastie and Tibshirani, 1990). We use the `S-Plus` function `gam` with the Gaussian family and nonparametric smoothing splines (option `s`).

GUIDE. Generalized regression tree (Loh, 2002). `Gc` and `Gm` denote piecewise constant and piecewise multiple linear models. Categorical variables are used for splitting and for regression modeling (via dummy vectors) in `Gm`. Our proposed piecewise simple linear and simple quadratic models are denoted by `Gl` and `Gq`, respectively. `Gs` denotes the method where forward and backward stepwise

regression is used in each node. If the number of regressors is limited to two, the method is denoted by `Gs2`. Finally, `Gf2` denotes the method using two-regressor forward-only stepwise regression at each node. The trees are pruned with the default 0.5-SE rule.

Least absolute deviations regression. We use the `S-Plus` function `l1fit`.

Least-squares linear regression. We use the `R` function `lm`.

NNET. Neural network using the `R` function `nnet` with `size=3, decay=0.001, linout=TRUE, skip=TRUE`, and `maxit=200`.

M5. Piecewise constant and linear regression tree. We use the implementation in version 3.2 of WEKA (Witten and Frank, 2000). `Mc` denotes piecewise constant and `Mm` piecewise multiple linear. If bagging is employed, we use ten iterations. The resulting methods are denoted by `Mcb` and `Mmb`, respectively.

MARS. Multivariate adaptive regression splines (Friedman, 1991). We use the `R` function `mars` in the `mda` library with parameter values `degree=1, penalty=2`, and `thresh=0.001`.

MART. Multivariate adaptive regression tree (Friedman, 2001). This is a stochastic gradient boosting algorithm applied to regression trees. We use the software from `www-stat.stanford.edu/~jhf/mart.html` with 10-fold cross-validation and 200 boosting iterations.

POLYMARS. An adaptive regression procedure using piecewise linear splines (Kooperberg, Bose and Stone, 1997). We use the `R` function `polymars` in the `polspline` library. The `gcv` option is used for model selection. The maximum number of basis functions is $\min(6n^{1/3}, n/4, 100)$ and the maximum number of knots per predictor is $\min(20, \texttt{round}(n/4))$, where n is the sample size.

Projection pursuit regression. (Friedman and Stuetzle, 1981). We use the `R` function `ppr` with `optlevel=2` in the `modreg` library.

Robust regression with M-estimate. (Huber, 1981, p. 194). We use the `R` function `rlm` with `init=ls, k2=1.345, maxit=20`, and `acc=1e-4` in the `MASS` library (Venables and Ripley, 1999).

RT. A regression tree algorithm due to (Torgo, 1999). Like M5, it first grows a piecewise constant tree and then fits various linear models to the leaf nodes during pruning. We use version 4 of the software. `Rc` denotes piecewise constant, `Rm` piecewise multiple linear, and `Rp` piecewise partial linear with bandwidth size 10.

4.3 Results

Because the measurement scale of the response variable varies from one dataset to another, it is necessary to standardize the prediction MSEs. Let the square root of the prediction MSE be denoted by RMSE. We divide the RMSE of each algorithm by that of least squares linear regression (`lr`) and then take its logarithm. We call the result the log relative root mean squared error (LRMSE). A positive LRMSE indicates that the algorithm is less accurate than `lr`.

Table 7 gives the geometric means of the RMSE relative to linear regression and Figure 13 displays them in a barplot. The two algorithms with the lowest geometric means are the ensemble methods `Crb` and `Mmb`, which have geometric means of 0.78 and 0.80, respectively. They are followed by `Mm`, `Ci`, `Cr` (all due to Quinlan), `Gs`, `Gm`, and `gam`. After these come our proposed two-regressor tree methods `Gf2` and `Gs2`, which employ forward-only and forward-backward stepwise regression, respectively.

It would be erroneous to conclude that if one algorithm has a smaller geometric mean RMSE than another, then the former always has smaller prediction error than the latter, because there is substantial

Table 7: Geometric means of RMSE relative to linear regression, in increasing order. Algorithms in the first row are not significantly different from `Crb`.

`Crb`	`Mmb`	`Mm`	`Ci`	`Cr`	`Gs`	`Gm`	`gam`	`Gf2`	`Gs2`	`mars`	`Mcb`	`ppr`
0.78	0.80	0.82	0.82	0.84	0.86	0.86	0.90	0.90	0.90	0.91	0.91	0.92
`nnet`	`mart`	`Gl`	`Rp`	`Cart`	`Rm`	`Gq`	`Mc`	`rreg`	`Gc`	`Rc`	`lad`	`pol`
0.94	0.96	0.96	0.97	0.97	0.98	0.98	0.99	1.0	1.0	1.0	1.1	1.2

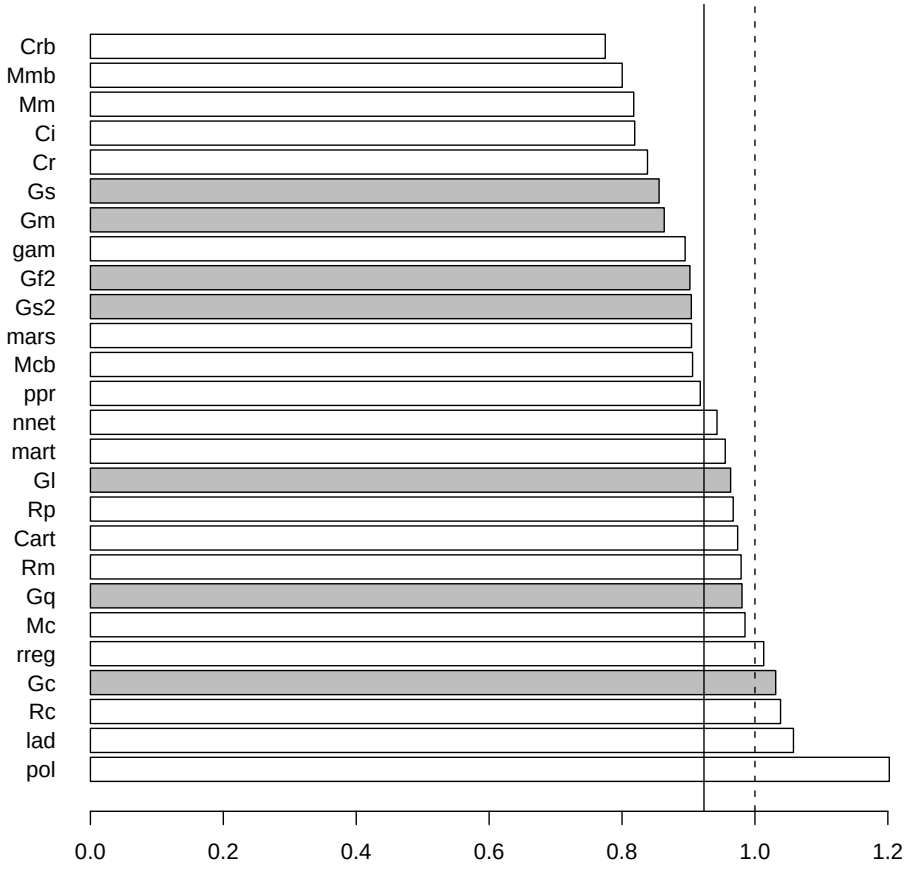

Figure 13: Barplot of geometric mean prediction RMSE relative to that of linear regression. The RMSE of an algorithm is not statistically significantly different from that of `Crb` if its associated bar ends to the left of the solid vertical line. The GUIDE algorithms are in gray.

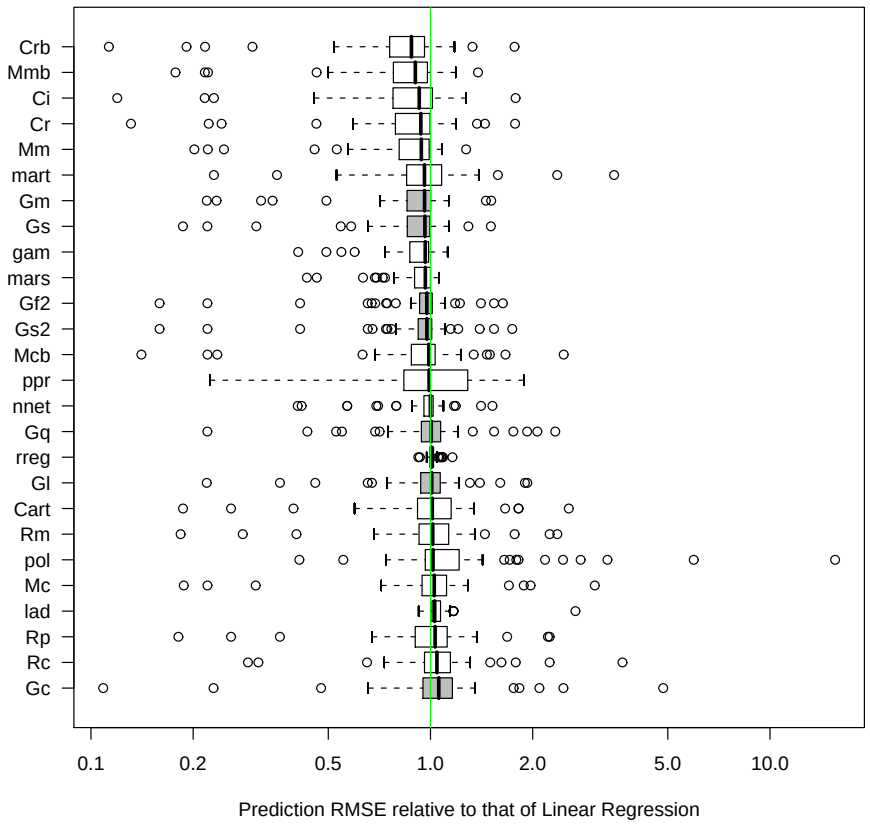

Figure 14: Boxplots of RMSE relative to that of linear regression, ordered by their medians. GUIDE algorithms are in gray.

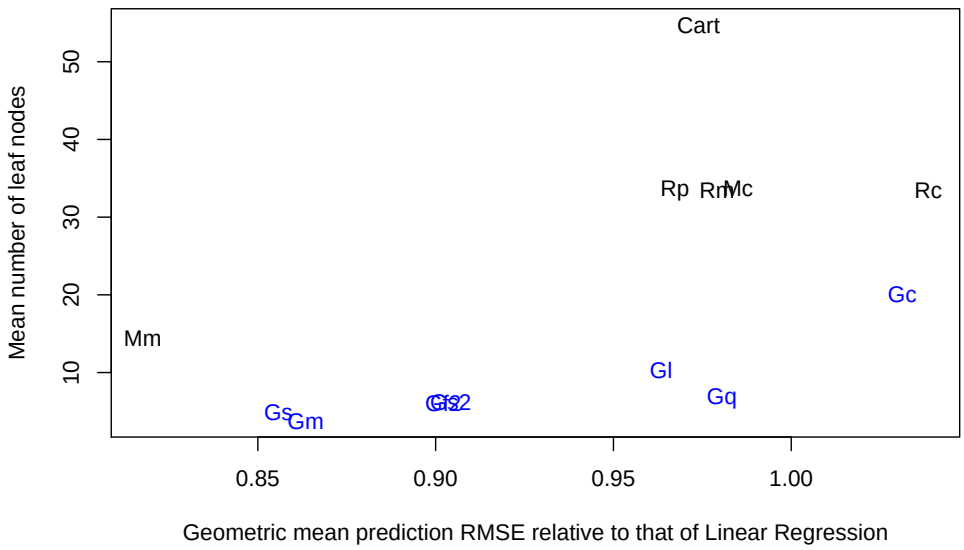

Figure 15: Mean number of leaf nodes versus geometric mean of prediction root mean square error relative to that of linear regression. The plot symbols for `Gf2` and `Gs2` almost coincide.

variability within each algorithm across datasets. This can be seen in Figure 14, which shows boxplots of the relative RMSE values by algorithm, ordered by their medians. We see there are datasets for which some algorithms (e.g., `Crb` and `Gc`) have RMSE values as low as one-tenth that of linear regression. On the other hand, there are also datasets for which algorithms such as `mart` and `Mcb` have relative RMSEs that are two or more.

To find out whether the differences in RMSEs are statistically significant, we fit a mixed-effects model to the LRMSE values, using algorithm as a fixed effect, dataset as a random effect, and their interaction as another independent random effect in place of the usual "error term". Calculations based on Tukey 95% simultaneous confidence intervals for pairwise differences show that algorithms with geometric mean RMSEs less than 0.923 are not significantly different from `Crb`. Thus differences among the prediction RMSEs of the top thirteen algorithms (listed in the first row of Table 7) are not statistically significant. Our piecewise two-regressor trees `Gf2` and `Gs2` belong to this group but not `Gl` and `Gq`. Also belonging to this group are `gam`, `mars`, `Mcb`, and `ppr`. Within this top group of thirteen, only `Gf2` and `Gs2` yield interpretable and visualizable models.

Although regression trees are often thought to be more interpretable than other methods, it should not be forgotten that interpretability depends on the complexity of a tree. All other things being equal, a tree with many leaf nodes takes more effort to interpret than one with few nodes. Figure 15 shows how the thirteen regression tree algorithms compare in terms of mean number of leaf nodes. The piecewise constant tree methods (`Cart`, `Gc`, `Rtc`, and `Mc`) tend to produce trees with many leaf nodes—20 for `Gc`, 34 for all versions of `Rt`, and 55 for `Cart`. This makes them quite hard to interpret in practice. `Gm` has the lowest average of 3.7. But its trees are not necessarily interpretable because each node is fitted with a multiple linear model. The class of tree methods that lie in between, namely the piecewise one- and two-regressor models, strikes a compromise with relatively simple node models and relatively compact trees: `Gl`, `Gq`, `Gf2`, and `Gs2` have on average 10.3, 6.7, 6.1, and 6.1 leaf nodes, respectively.

5 Asymptotic behavior of regression estimates

We state and prove here an asymptotic consistency theorem for piecewise two-regressor models. The theorem provides hope that the good empirical performance of the method will scale up to arbitrarily large sample sizes. First, we establish some notation. Let $(Y_1, \mathbf{X}_1)$, $(Y_2, \mathbf{X}_2)$, ..., $(Y_n, \mathbf{X}_n)$ be n independent observations forming the training sample. The Y's are real valued and the $\mathbf{X}$'s take values in a d-dimensional compact hyper-rectangle C in the Euclidean space R^d. Suppose that T_n is a partition of C consisting of sets that are also hyper-rectangles in R^d. For any $t \in T_n$, we will denote by $\delta(t)$ the diameter of t defined as $\delta(t) = \sup\{\|\mathbf{x} - \mathbf{z}\| : \mathbf{x}, \mathbf{z} \in t\}$, where $\|.\|$ is the usual Euclidean norm of a vector, and we define the norm of the partition T_n as $|T_n| = \sup\{\delta(t) : t \in T_n\}$.

For $t \in T_n$, let N_t be the number of $\mathbf{X}$'s in t, $N_n = \min\{N_t : t \in T_n\}$, and $\bar{\mathbf{X}}_t = N_t^{-1} \sum_{\mathbf{X}_i \in t} \mathbf{X}_i$. Given a non-negative integer m, let $\mathbf{U}$ be a set of pairs $\mathbf{u} = (u_1, u_2)$ of non-negative integers such that $[\mathbf{u}] \leq m$, where $[\mathbf{u}] = u_1 + u_2$. Let $s(\mathbf{U})$ denote the size of $\mathbf{U}$. For $\mathbf{x}, \mathbf{z} \in t$, define the $s(\mathbf{U})$-dimensional vector $\Gamma(\mathbf{x}, \mathbf{z}; p, q) = [\{\delta(t)\}^{-[\mathbf{u}]}(x_p - z_p)^{u_1}(x_q - z_q)^{u_2}]_{\mathbf{u} \in \mathbf{U}}$, where $1 \leq p, q \leq d$, $\mathbf{x} = (x_1, x_2, \ldots, x_d)$, and $\mathbf{z} = (z_1, z_2, \ldots, z_d)$. Also define $D(p, q, t) = \sum_{\mathbf{X}_i \in t} \Gamma^T(\mathbf{X}_i, \bar{\mathbf{X}}_t; p, q)\, \Gamma(\mathbf{X}_i, \bar{\mathbf{X}}_t; p, q)$.

Consider the least-squares fit of a two-regressor polynomial model of order m in partition t and let p_t and q_t be the indices of the two selected variables. Then the estimate of the regression function $g(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x})$ for $\mathbf{x} \in t$ is given by the expression

$$\hat{g}(\mathbf{x}) = \Gamma(\mathbf{x}, \bar{\mathbf{X}}_t; p_t, q_t) D^{-1}(p_t, q_t, t) \sum_{\mathbf{X}_i \in t} \Gamma^T(\mathbf{X}_i, \bar{\mathbf{X}}_t; p_t, q_t)\, Y_i.$$

Let λ_t denote the smallest eigenvalue of $D(p_t, q_t, t)$ and write $\lambda_n = \min\{\lambda_t : t \in T_n\}$. Further, let $\psi(a|\mathbf{x}) = E[\exp\{a|Y - g(\mathbf{x})|\}|\mathbf{X} = \mathbf{x}]$ for any $a > 0$ such that the expectation is finite.

Theorem 1 *Assume that the regression function is continuous in C. Suppose that $|T_n|$ and $\log n/N_n$ tend to zero and that λ_n remains bounded away from zero in probability as $n \to \infty$. If there exists $a > 0$ such that $\psi(a|\mathbf{x})$ is bounded in C, then $\sup\{|\hat{g}(\mathbf{x}) - g(\mathbf{x})| : \mathbf{x} \in C\} \to 0$ in probability as $n \to \infty$.*

Proof: First observe that

$$\begin{aligned} \hat{g}(\mathbf{x}) &= \Gamma(\mathbf{x}, \bar{\mathbf{X}}_t; p_t, q_t) D^{-1}(p_t, q_t, t) \sum_{\mathbf{X}_i \in t} \Gamma^T(\mathbf{X}_i, \bar{\mathbf{X}}_t; p_t, q_t) g(\mathbf{X}_i) \\ &\quad + \Gamma(\mathbf{x}, \bar{\mathbf{X}}_t; p_t, q_t) D^{-1}(p_t, q_t, t) \sum_{\mathbf{X}_i \in t} \Gamma^T(\mathbf{X}_i, \bar{\mathbf{X}}_t; p_t, q_t)\, \epsilon_i \end{aligned}$$

where $\epsilon_i = Y_i - g(\mathbf{X}_i)$. Replacing $g(\mathbf{X}_i)$ with $g(\mathbf{x}) + \{g(\mathbf{X}_i) - g(\mathbf{x})\}$ in the first term on the right hand side above, we obtain after some straightforward algebraic simplification

$$\begin{aligned} \hat{g}(\mathbf{x}) - g(\mathbf{x}) &= \Gamma(\mathbf{x}, \bar{\mathbf{X}}_t; p_t, q_t) D^{-1}(p_t, q_t, t) \sum_{\mathbf{X}_i \in t} \Gamma^T(\mathbf{X}_i, \bar{\mathbf{X}}_t; p_t, q_t) \{g(\mathbf{X}_i) - g(\mathbf{x})\} \\ &\quad + \Gamma(\mathbf{x}, \bar{\mathbf{X}}_t; p_t, q_t) D^{-1}(p_t, q_t, t) \sum_{\mathbf{X}_i \in t} \Gamma^T(\mathbf{X}_i, \bar{\mathbf{X}}_t; p_t, q_t)\, \epsilon_i. \end{aligned}$$

This is a consequence of $D^{-1}(p_t, q_t, t) \sum_{\mathbf{X}_i \in t} \Gamma^T(\mathbf{X}_i, \bar{\mathbf{X}}_t; p_t, q_t)$ being an $s(\mathbf{U})$-dimensional vector with 1 as the first coordinate and the other coordinates all equal to zero.

Now, following the ideas in the proof of Theorem 1 in Chaudhuri, Huang, Loh and Yao (1994), the first term on the right hand side in the expression for $\hat{g}(\mathbf{x}) - g(\mathbf{x})$ can be viewed as a *bias term* while the second term can be thought of as a *variance term* that occurs in the decomposition of the error in a nonparametric regression estimate. Since the function $g(\mathbf{x})$ is a uniformly continuous function in C, it follows immediately

that the bias term tends to zero in probability uniformly in $\mathbf{x}$ as the sample size grows to infinity. Further, since the partition sets are hyper-rectangles and the moment generating function of $Y - g(\mathbf{X})$ is bounded, the arguments in the proof of Chaudhuri, Huang, Loh and Yao (1994, Theorem 1) imply that the variance term tends to zero in probability uniformly in $\mathbf{x}$.

6 Conclusion

We set out seeking an algorithm that can automatically generate interpretable and visualizable models with good prediction accuracy. We gave as motivation the difficulty of interpreting the coefficients in a multiple linear model. Our solution embraces rather than discards the linear model, but we limit it to at most two regressors and apply it to partitions of the data. As the Boston and baseball examples demonstrate, this approach can yield models that fit the data at least as well as those built by human experts. Our models do not require special training or equipment for visualization; all that is needed are tree diagrams, graphs, and contour maps. Our trees are also substantially more compact than piecewise constant trees. And they can be interpreted without worrying about selection bias.

In terms of prediction accuracy, our piecewise two-regressor model `Gf2` yields on average about 80% of the prediction MSE of least squares linear regression. Although `Gf2` does not have the lowest average, its prediction MSE does not differ significantly from the lowest at the 0.05 simultaneous level of significance. We note that the lowest average value over the twenty-seven algorithms is 60%. If the datasets used in our study are representative of all real datasets, this result suggests that it is hard to beat the prediction MSE of least squares linear regression by a very large amount in real applications. Given this, it is quite reasonable to demand more from all algorithms, including interpretability and visualizability of their models.

References

[1] Aaberge, R., Colombino, U. and Strom, S. (1999) Labor supply in Italy: An empirical analysis of joint household decisions, with taxes and quantity constraints. *Journal of Applied Econometrics*, 14, 403–422.

[2] Afifi, A. and Azen, S. (1979) *Statistical Analysis: A Computer Oriented Approach*, 2nd edn, Academic Press, New York.

[3] Belsley, D. A., Kuh, E. and Welsch, R. E. (1980) *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, Wiley, New York.

[4] Berndt, E. R. (1991) *The Practice of Econometrics*, Addison-Wesley, New York.

[5] Blake, C. and Merz, C. (1998) *UCI repository of machine learning databases*. `http://www.ics.uci.edu/~mlearn/MLRepository.html`.

[6] Bollino, C. A., Perali, F. and Rossi, N. (2000) Linear household technologies. *Journal of Applied Econometrics*, 15, 253–274.

[7] Box, G. E. P. (1979) Robustness in the strategy of scientific model building, in R. L. Launer and G. N. Wilkinson (eds), *Robustness in Statistics*, Academic Press, New York, pp. 201–236.

[8] Breiman, L. and Friedman, J. (1988) Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 83, 580–597.

[9] Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984) *Classification and Regression Trees*, Wadsworth, Belmont.

[10] Bryant, P. G. and Smith, M. A. (1996) *Practical Data Analysis: Case Studies in Business Statistics*, Vol. 3, Irwin/McGraw Hill.

[11] Chattopadhyay, S. (2003) Divergence in alternative Hicksian welfare measures: The case of revealed preference for public amenities. *Journal of Applied Econometrics*, 17, 641–666.

[12] Chaudhuri, P., Huang, M.-C., Loh, W.-Y. and Yao, R. (1994) Piecewise-polynomial regression trees. *Statistica Sinica*, 4, 143–167.

[13] Chu, S. (2001) Pricing the C's of diamond stones. *Journal of Statistics Education*. `http://www.amstat.org/publications/jse`.

[14] Cochran, J. J. (2000) Career records for all modern position players eligible for the major league baseball hall of fame. *Journal of Statistics Education*. `http://www.amstat.org/publications/jse`.

[15] Cochran, J. J. (2002) Data management, exploratory data analysis, and regression analysis with 1969–2000 major league baseball attendance. *Journal of Statistics Education*. `http://www.amstat.org/publications/jse`.

[16] Cook, D. (1998) *Regression Graphics: Ideas for Studying Regression Through Graphics*, Wiley, New York.

[17] Cook, D. and Weisberg, S. (1994) *An Introduction to Regression Graphics*, Wiley, New York.

[18] Deb, P. and Trivedi, P. K. (1997) Demand for medical care by the elderly: A finite mixture approach. *Journal of Applied Econometrics*, 12, 313–336.

[19] Delgado, M. A. and Mora, J. (1998) Testing non-nested semiparametric models: An application to Engel curves specification. *Journal of Applied Econometrics*, 13, 145–162.

[20] Denman, N. and Gregory, D. (1998) Analysis of sugar cane yields in the Mulgrave area, for the 1997 sugar cane season. Technical report, MS305 Data Analysis Project, Department of Mathematics, University of Queensland.

[21] Fernandez, C., Ley, E. and Steel, M. F. J. (2002) Bayesian modelling of catch in a north-west Atlantic fishery. *Applied Statistics*, 51, 257–280.

[22] Friedman, J. (1991) Multivariate adaptive regression splines (with discussion). *Annals of Statistics*, 19, 1–141.

[23] Friedman, J. (2001) Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29, 1189–1232.

[24] Friedman, J. H. and Stuetzle, W. (1981) Projection pursuit regression. *Journal of the American Statistical Association*, 76, 817–823.

[25] Gilley, O. W. and Pace, R. K. (1996) On the Harrison and Rubinfeld data. *Journal of Environmental Economics and Management*, 31, 403–405.

[26] Hair, J. F., Anderson, R. E., Tatham, R. L. and Black, W. C. (1998) *Multivariate Data Analysis*, Prentice Hall, New Jersey.

[27] Hallin, M. and Ingenbleek, J.-F. (1983) The Swedish automobile portfolio in 1977: A statistical study. *Scandinavian Actuarial Journal*, 83, 49–64.

[28] Harrell, Jr., F. E. (2001) *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*, Springer-Verlag, New York.

[29] Harrison, D. and Rubinfeld, D. L. (1978) Hedonic prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5, 81–102.

[30] Hastie, T. and Tibshirani, R. (1990) *Generalized Additive Models*, CRC Press.

[31] Hoaglin, D. C. and Velleman, P. F. (1995) A critical look at some analyses of major league baseball salaries. *American Statistician*, 49, 277–285.

[32] Horrace, W. C. and Schmidt, P. (2000) Multiple comparisons with the best, with economic applications. *Journal of Applied Econometrics*, 15, 1–26.

[33] Huber, P. J. (1981) *Robust Statistics*, Wiley.

[34] Kenkel, D. S. and Terza, J. V. (2001) The effect of physician advice on alcohol consumption: Count regression with an endogenous treatment effect. *Journal of Applied Econometrics*, 16, 165–184.

[35] Kooperberg, C., Bose, S. and Stone, C. (1997) Polychotomous regression. *Journal of the American Statistical Association*, 92, 117–127.

[36] Laroque, G. and Salanie, B. (2002) Labor market institutions and employment in France. *Journal of Applied Econometrics*, 17, 25–28.

[37] Liu, Z. and Stengos, T. (1999) Non-linearities in cross country growth regressions: A semiparametric approach. *Journal of Applied Econometrics*, 14, 527–538.

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

[39] Lutkepohl, H., Terasvirta, T. and Wolters, J. (1999) Investigating stability and linearity of a German M1 money demand function. *Journal of Applied Econometrics*, 14, 511–525.

[40] Martins, M. F. O. (2001) Parametric and semiparametric estimation of sample selection models: An empirical application to the female labour force in Portugal. *Journal of Applied Econometrics*, 16, 23–40.

[41] Miller, A. (2002) *Subset Selection in Regression*, 2nd edn, Chapman & Hall.

[42] Neter, J., Kutner, M. H., Nachtsheim, C. J. and Wasserman, W. (1996) *Applied Linear Statistical Models*, 4th edn, Irwin.

[43] Olson, C. A. (1998) A comparison of parametric and semiparametric estimates of the effect of spousal health insurance coverage on weekly hours worked by wives. *Journal of Applied Econometrics*, 13, 543–565.

[44] Onoyama, K., Ohsumi, N., Mitsumochi, N. and Kishihara, T. (1998) Data analysis of deer-train collisions in eastern Hokkaido, Japan, in C. Hayashi, N. Ohsumi, K. Yajima, Y. Tanaka, H.-H. Bock and Y. Baba (eds), *Data Science, Classification, and Related Methods*, Springer-Verlag, Tokyo, pp. 746–751.

[45] Pace, R. K. and Barry, R. (1997) Sparse spatial autoregressions. *Statistics and Probability Letters*, 33, 291–297.

[46] Penrose, K., Nelson, A. and Fisher, A. (1985) Generalized body composition prediction equation for men using simple measurement techniques. *Medicine and Science in Sports and Exercise*, 17, 189.

[47] Quinlan, J. R. (1992) Learning with continuous classes, in Proceedings of the 5th Australian Joint Conference on Artificial Intelligence, pp. 343–348.

[48] Rawlings, J. O. (1988) *Applied Regression Analysis: A Research Tool*, Wadsworth & Brooks/Cole.

[49] Schafgans, M. M. (1998) Ethnic wage differences in Malaysia: Parametric and semiparametric estimation of the Chinese-Malay wage gap. *Journal of Applied Econometrics*, 13, 481–504.

[50] Simonoff, J. (1996) *Smoothing Methods in Statistics*, Springer-Verlag, New York.

[51] Steinberg, D. and Colla, P. (1995) *CART: Tree-Structured Non-Parametric Data Analysis*, Salford Systems, San Diego, CA.

[52] Torgo, L. (1999) *Inductive Learning of Tree-based Regression Models*, PhD thesis, Department of Computer Science, Faculty of Sciences, University of Porto.

[53] Venables, W. N. and Ripley, B. D. (1999) *Modern Applied Statistics with S-Plus*, Springer-Verlag, New York, NY.

[54] Weiss, S. and Indurkhya, N. (1995) Rule-based machine learning methods for functional prediction. *Journal of Artificial Intelligence Research*, 3, 383–403.

[55] Witten, I. and Frank, E. (2000) *Data Mining: Practical Machine Learning Tools and Techniques with JAVA Implementations*, Morgan Kaufmann, San Fransico, CA.

Biographical sketches

Hyunjoong Kim is Assistant Professor, Department of Applied Statistics, Yonsei University, Korea. He has a PhD in Statistics from the University of Wisconsin, Madison. His research interests are in data mining, tree-based statistical modeling, and statistical computing.

Wei-Yin Loh is Professor, Department of Statistics, University of Wisconsin, Madison. He has a PhD in Statistics from the University of California, Berkeley. His research interests are in statistical theory and methodology, including data mining and machine learning techniques.

Yu-Shan Shih is Professor, Department of Mathematics, National Chung Cheng University, Taiwan, ROC. He has a PhD in Statistics from the University of Wisconsin, Madison. His research interests are in multivariate methods, computational statistics and tree-structured methods.

Probal Chaudhuri is Professor, Theoretical Statistics and Mathematics Unit, Indian Statistical Institute, Calcutta, India. He has a PhD degree in Statistics from the University of California, Berkeley. His research interests include nonparametric statistics, robustness and applications of statistics in engineering and biological sciences. He is a Fellow of the Indian Academy of Sciences.