Extrapolation Errors in Linear Model Trees

WEI-YIN LOH, CHIEN-WEI CHEN, and WEI ZHENG
University of Wisconsin, Madison

Prediction errors from a linear model tend to be larger when extrapolation is involved, particularly when the model is wrong. This article considers the problem of extrapolation and interpolation errors when a linear model tree is used for prediction. It proposes several ways to curtail the size of the errors, and uses a large collection of real datasets to demonstrate that the solutions are effective in reducing the average mean squared prediction error. The article also provides a proof that, if a linear model is correct, the proposed solutions have no undesirable effects as the training sample size tends to infinity.

Categories and Subject Descriptors: I.2.6 [**Artificial Intelligence**]: Learning—*concept learning, induction, parameter learning*; I.5.1 [**Pattern Recognition**]: Models—*statistical*; I.5.2 [**Pattern Recognition**]: Design Methodology—*classifier design and evaluation*

General Terms: Algorithms, Performance

Additional Key Words and Phrases: Decision tree, prediction, regression, statistics

ACM Reference Format:

Loh, W.-Y., Chen, C.-W., and Zheng, W. 2007. Extrapolation errors in linear model trees. ACM Trans. Knowl. Discov. Data. 1, 2, Article 6 (August 2007), 17 pages. DOI = 10.1145/1267066.1267067 http://doi.acm.org/10.1145/1267066.1267067

1. INTRODUCTION

Given a training sample $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, the purpose of linear regression is to find a linear model approximation $\hat{f}(x)$ of a function $f(x) = E(y|x)$ that can be used to predict the value of a new y given x. When x lies outside the convex hull of the x_i-values, the prediction is called extrapolation. The latter usually produces larger prediction errors than interpolation, for which x lies inside the convex hull.

We consider the situation where $\hat{f}(x)$ is constructed using a regression tree algorithm, such as CART [Breiman et al. 1984], M5' [Wang and Witten 1997],

This 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.
Authors' address: University of Wisconsin, Madison, 1300 University Avenue, Madison, WI 53706, email: {loh, zheng}@stat.wisc.edu; chienweichen@wisc.edu.

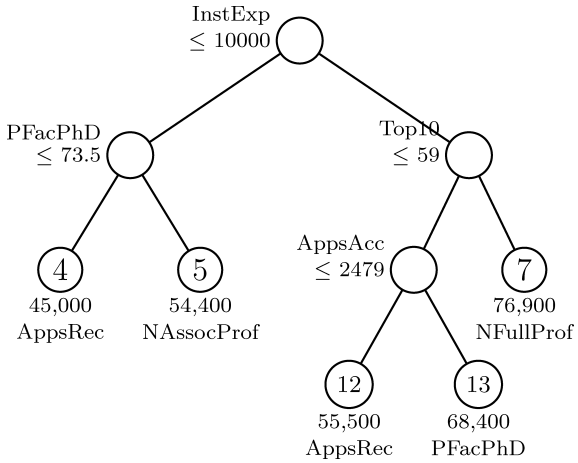

Fig. 1. GUIDE piecewise cubic model for full professor salary. At each intermediate node, a case goes to the left child node if and only if the condition is satisfied. Beneath each leaf node are the sample mean of full professor salary and the name of the selected regressor variable.

or GUIDE [Loh 2002]. These algorithms employ a divide-and-conquer strategy to find $\hat{f}(x)$: first partition the training sample into several pieces and then estimate the function within each partition (represented by a leaf node of the tree) with a linear model estimated from the training data in the partition. Thus, $\hat{f}(x)$ consists of one or more linear pieces, and extrapolation occurs whenever x lies within a partition but outside the convex hull of the data in that partition (note that x may lie within the convex hull of the whole training sample, in which case the problem may also be viewed as one of interpolation).

To illustrate the difficulties, let us examine some data on the 1995 salaries of full professors in U.S. colleges taken from StatLib (`http://lib.stat.cmu.edu`). We use a subset of 694 colleges with complete observations on twenty-five variables, including instructional expenditure per student (`InstExp`), number of applications received (`AppsRec`) and accepted (`AppsAcc`), number of associate (`NAssocProf`) and full professors (`NFullProf`), percent of faculty with PhDs (`PFacPhD`), and percent of new students from the top 10% of their high-school class (`Top10`).

Figure 1 shows a regression tree, constructed with version 4 of the GUIDE algorithm, where a cubic or lower-order polynomial in the best single predictor variable is fitted at each node. The order of the polynomial is decided by sequentially testing the statistical significance of the highest order term (starting with the cubic), and stopping when the test is significant at the five percent level. In the tree diagram, the mean full professor salary is printed beneath each leaf node, together with the name of the selected polynomial predictor variable. The fitted polynomials are shown in Figure 2 with the data points. An advantage of fitting a polynomial rather than a multiple linear model is that the latter cannot be presented graphically in this way.

One notable characteristic of the graphs is the nonuniform distributions of the data points. In node 4, for instance, the upswing of the cubic polynomial is

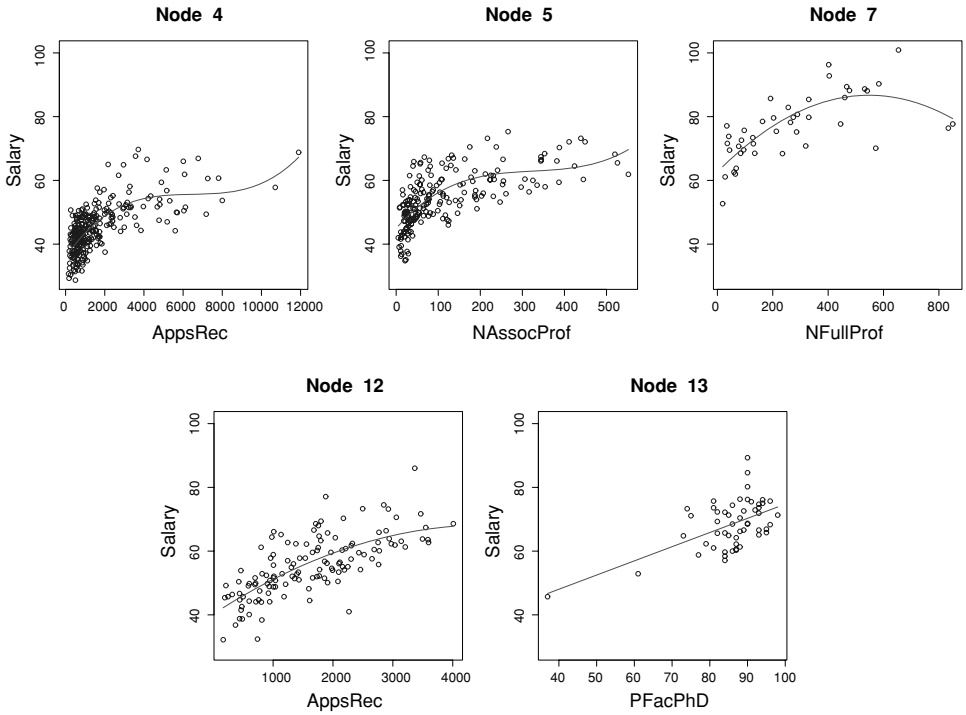

Fig. 2. Data and fitted functions in the leaf nodes of the GUIDE model in Figure 1.

essentially determined by two points on the right. Similarly, the downturn of the quadratic in node 7 is determined by two points on the right. Finally, in node 13, the slope of the fitted line is largely controlled by one point on the left. These high-leverage points, as they are called, have the potential to produce large extrapolation errors. Outlier deletion is not a solution here, because the outliers are not errors. For example, the two outliers with large values of `NFullProf` in node 7 are associated with large public schools that typically do not pay very high salaries.

The above problems are not rare, especially when the data partitions are generated by automated algorithms. One way to reduce the size of the extrapolation errors is to truncate the fitted functions. We study four simple methods for doing this. To be effective, they need to have two desirable properties: (i) reduced average prediction error when extrapolation is harmful and (ii) little or no increase in average prediction error when extrapolation is harmless. The methods are presented in Section 2, where they are applied to various types of GUIDE models and evaluated on a set of fifty-two real datasets. Truncation is shown to be generally helpful in reducing average prediction mean squared error (MSE), although the degree of improvement varies from one algorithm to another.

What happens when $f(x)$ is itself a linear model? Will truncation make matters worse? Fortunately, the answer to the latter question is "no", at least as long as the sample size is sufficiently large. This is established theoretically

in Section 3, where we prove that the effect of truncation on prediction MSE vanishes asymptotically as the training sample size tends to infinity. We compare the truncated GUIDE methods with M5', random forest, and two spline methods, with and without truncation, in Section 4, and conclude with some remarks in Section 5.

2. THE METHODS

In real applications, the value of the y variable is likely to be bounded above, below, or both. For example, college professor salaries cannot be arbitrarily high and there is a natural lower bound of zero. One way to avoid catastrophic extrapolation errors is to force $\hat{f}(x)$ to lie between the bounds of the training sample. There are three obvious ways to do this. Let $y_{(1)} = \min_{1\le i\le n} y_i$ and $y_{(n)} = \max_{1\le i\le n} y_i$ denote the smallest and largest y-values, respectively, in the training sample in the node, and let $r_n = y_{(n)} - y_{(1)}$ denote its range.

Type 1. Truncate $\hat{f}(x)$ outside the range of the training sample y-values in the node: $\max[\min\{\hat{f}(x), y_{(n)}\}, y_{(1)}]$. This method is used in Kim et al. [2007].

Type 2. Given a constant $c \ge 0$, truncate $\hat{f}(x)$ outside $(y_{(1)} - cr_n, y_{(n)} + cr_n)$:

$$\max[\min\{\hat{f}(x), y_{(n)} + cr_n\}, y_{(1)} - cr_n].$$

We use $c = 0.1$ in the empirical comparisons below.

Type 3. Truncate $\hat{f}(x)$ outside the range of the y-values of the entire training sample, that is, the range at the root node of the tree.

These methods would be effective in controlling extrapolation errors in three of the five nodes in Figure 2. In nodes 7 and 12, however, they are ineffective at the extreme right, because the value of $\hat{f}(x)$ stays within the range of the training data in each graph. An alternative solution is to make $\hat{f}(x)$ continuous but flat (i.e., constant) once x falls outside the range of its sample values in the node. We call this *Winsorization* because of its similarity to a technique in robust estimation where outliers are moved in closer to the bulk of the data; see, for example, Hampel et al. [1986, p. 179].

Type 4. If x is outside the range of the training sample in the node, replace it with x_0, where x_0 is the training sample value in the node nearest to x, and define $\hat{f}(x) = \hat{f}(x_0)$. For ease of computation when x is multidimensional, x_0 is defined as the point that is coordinate-wise nearest to x. That is, x_0 is the point, on the smallest hyper-rectangle containing the training sample in the node, that is nearest to x.

Before we evaluate the effectiveness of these methods, we need to distinguish between ordered variables (e.g., percent of faculty with PhDs) and unordered variables (e.g., type of college: I, IIA, or IIB). An ordered variable will be called a linear predictor and an unordered one a categorical predictor. If desired, each categorical predictor is converted into a set of 0-1 dummy variables when a model is fitted to the data in a node. These dummy variables are not used to

split the nodes; splits on categorical variables are on subsets of the categories. We consider the following seven models for fitting the data in the nodes of a GUIDE tree.

`Gm`: Multiple linear regression model using all the variables (including dummy variables) in each node.

`Gs`: Multiple linear regression model using forward-and-backward stepwise selection of variables (including dummy variables) in each node.

`Gp`: Two-predictor multiple linear regression model using the best pair of variables (including dummy variables) in each node.

`Ga`: Simple analysis of covariance (ANCOVA) model using in each node the best single linear predictor and stepwise selection of dummy variables from the categorical predictors.

`G1`: Simple polynomial regression model of highest order 1 using the best single linear predictor in each node.

`G2`: Simple polynomial regression model of highest order 2 (quadratic) using the best single predictor in each node.

`G3`: Simple polynomial regression model of highest order 3 (cubic) using the best single predictor in each node.

Figure 1 is obtained using `G3`. Model `Ga` is the same as `G1` if there are no categorical variables.

Combining the four truncation methods with the seven node model methods gives a total of twenty-eight variants of the GUIDE algorithm. We identify them by concatenating the name of the model with that of the truncation method. For example, the variant employing model `Gm` and truncation method 1 is named `Gm1`. To ensure that the cross-validation error estimates for pruning are valid, the truncation methods are built into the GUIDE tree construction algorithm—it is not applied *post-hoc*, after the pruned tree is found.

We evaluate the prediction accuracy of the variants by applying them to fifty-two real datasets. Their sources and characteristics are listed in Tables I and II. The datasets in the second table come with their own test sets, while those in the first table do not. We use two-fold cross-validation to estimate the prediction MSE of each method on the forty-six datasets in Table I. Specifically, each dataset is randomly divided into two roughly equal-sized subsets. Each algorithm is trained on one subset, and the other subset is used as a test sample to give an estimate of the MSE. The process is repeated with the roles of the two subsets reversed to obtain another estimate of MSE. The average of the two estimates gives the cross-validation estimate of the MSE of the algorithm. For the six datasets in Table II, the MSE is estimated directly from the test sets.

To measure the effect of the truncation methods, we first divide the estimated MSE of each truncation-model variant by that of the same method without truncation, for each dataset. This gives a measure of the effectiveness of truncation for each truncation-model variant on the dataset. The geometric mean of these measures over the datasets gives a sense of the average effectiveness of the variant. The geometric mean is a better summary than the arithmetic

Table I. Forty-Six Datasets without Separate Test Samples. N Denotes the Number of Training Cases, L the Number of Linear Predictors, C the Number of Categorical Predictors, and $F = L +$ Number of Dummy Variables

Name	N	L	C	F	Reference
Abalone	4177	7	1	9	[Blake and Merz 1998]
Ais	202	11	1	19	[Cook and Weisberg 1994]
Alcohol	2467	12	6	29	[Kenkel and Terza 2001]
Amenity	3044	19	2	24	[Chattopadhyay 2003]
Attend	838	7	2	37	[Cochran 2002]
Baseball	263	18	2	63	Statlib
Baskball	96	4	0	4	[Simonoff 1996]
Boston	506	13	0	13	[Belsey et al. 1980]
Boston2	506	13	1	104	[Belsey et al. 1980]
Budget	1729	10	0	10	[Bollino et al. 2000]
Cane	3775	6	3	56	[Denman and Gregory 1998]
Cardio	375	6	3	26	[Bryant and Smith 1996]
College	694	23	1	25	Statlib
County	3114	12	1	57	[Harrell 2001]
Cps	534	7	3	16	[Berndt 1991]
Cpu	209	6	1	35	[Blake and Merz 1998]
Deer	654	10	3	22	[Onoyama et al. 1998]
Diabetes	375	14	1	16	[Harrell 2001]
Diamond	308	1	3	12	[Chu 2001]
Edu	1400	5	0	5	[Martins 2001]
Enroll	258	6	0	6	[Liu and Stengos 1999]
Fame	1318	21	1	27	[Cochran 2000]
Fat	252	14	0	14	[Penrose et al. 1985]
Fishery	6806	11	3	22	[Fernandez et al. 2002]
Hatco	100	12	1	14	[Hair et al. 1998]
Insur	2182	4	2	18	[Hallin and Ingenbleek 1983]
Labor	2953	18	0	18	[Aaberge et al. 1999]
Laheart	200	13	3	23	[Afifi and Azen 1979]
Medicare	4406	21	0	21	[Deb and Trivedi 1997]
Mpg	392	6	1	8	[Blake and Merz 1998]
Mpg2001	849	5	5	63	www.fueleconomy.gov
Mumps	1523	3	0	3	Statlib
Mussels	201	3	1	7	[Cook 1998]
Ozone	330	8	0	8	[Breiman and Friedman 1988]
Price	159	15	0	15	[Blake and Merz 1998]
Rate	144	9	0	9	[Lutkepohl et al. 1999]
Rice	171	13	2	17	[Horrace and Schmidt 2000]
Scenic	113	9	1	12	[Neter et al. 1996]
Servo	167	2	2	10	[Blake and Merz 1998]
Smsa	141	9	1	12	[Neter et al. 1996]
Strike	625	4	1	21	Statlib
Ta	324	3	3	73	Authors
Tecator	215	10	0	10	Statlib
Tree	100	8	0	8	[Rawlings 1988]
Triazine	186	28	0	28	[Torgo 1999]
Wage	3380	13	0	13	[Schafgans 1998]

Table II. Six Datasets with Separate Test Samples. N Denotes the Number of Training Cases, N' the Number of Test Cases, L the Number of Linear Predictors, C the Number of Categorical Predictors, and $F = L +$ Number of Dummy Variables

Name	N	N'	L	C	F	Reference
Cps95	21252	42504	8	6	32	ftp.stat.berkeley.edu/pub/datasets/fam95.zip
Engel	11986	11986	5	0	5	[Delgado and Mora 1998]
Houses	6880	13760	8	0	8	[Pace and Barry 1997]
Labor2	5443	5443	17	0	17	[Laroque and Salanie 2002]
Pole	5000	10000	26	0	26	[Weiss and Indurkhya 1995]
Spouse	11136	11136	21	0	21	[Olson 1998]

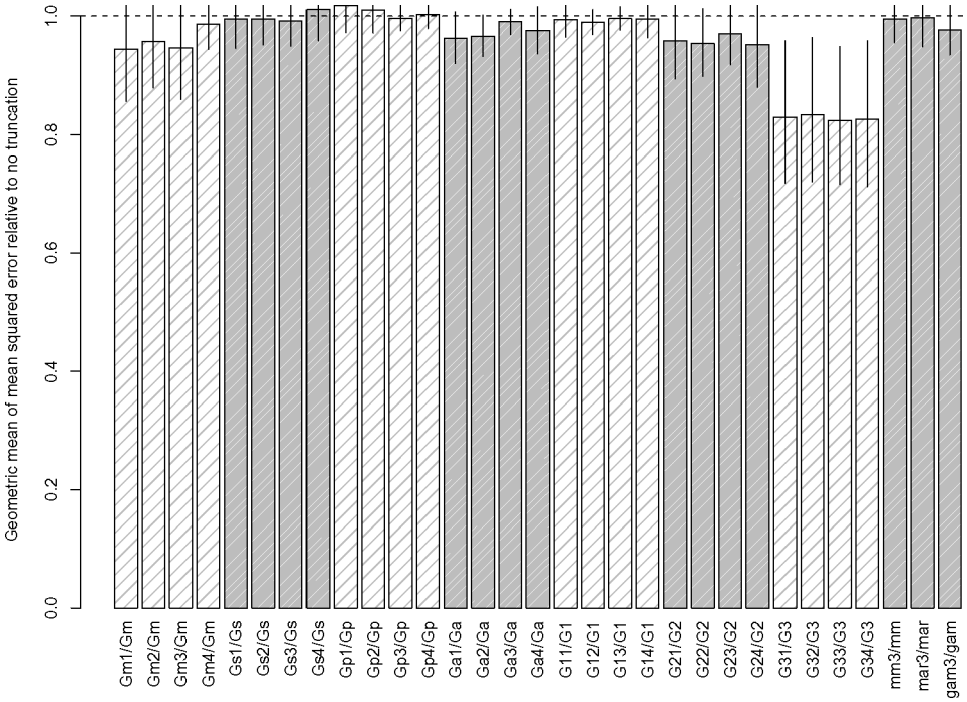

Fig. 3. Barchart of geometric means of mean squared prediction error relative to no truncation for the forty-six datasets in Table I. The vertical line at the end of each bar is an approximate 95% confidence interval for the geometric mean.

mean because the measures are bounded between 0 and infinity, with unity representing no effect.

A barchart of the geometric means for the GUIDE and three other methods (introduced in Section 4 below) based on the forty-six datasets in Table I is shown in Figure 3. The corresponding barchart for the six datasets in Table II is shown in Figure 4. A 95% confidence interval is drawn as a vertical line at the end of each bar. We see that truncation seldom increases prediction MSE. On the other hand, it is quite beneficial for Gm, G2, and particularly G3. The effects are more pronounced in Figure 4, where the test sample sizes are larger, and hence provide more opportunities for extrapolation errors. The confidence intervals in this barchart show that truncation is almost never harmful. Overall,

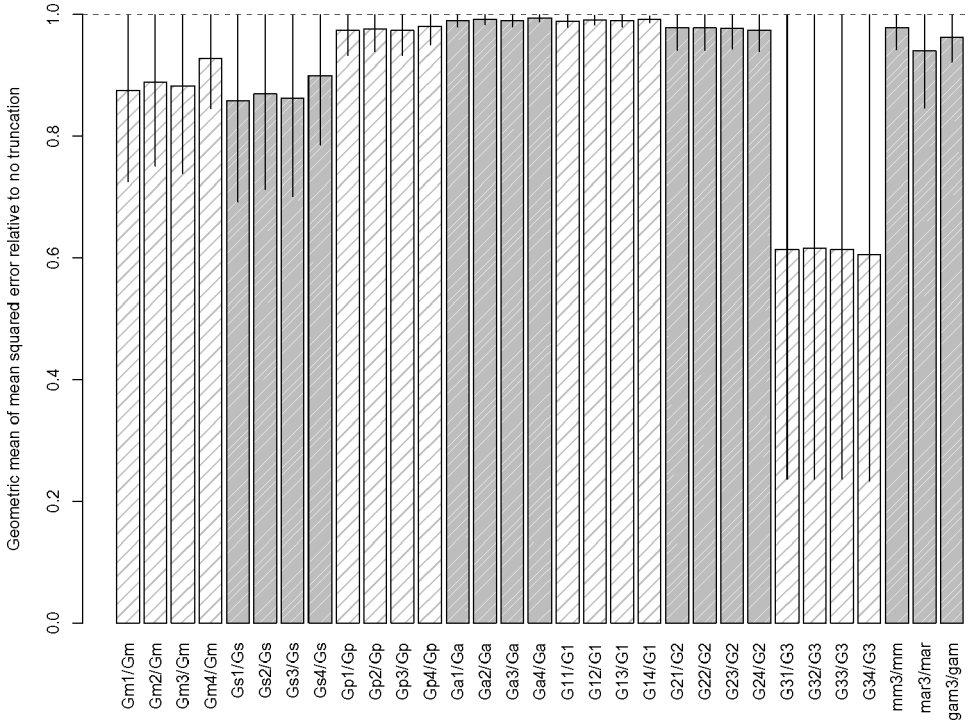

Fig. 4. Barchart of geometric means of mean squared prediction error relative to no truncation for the six datasets in Table II. The vertical line at the end of each bar is an approximate 95% confidence interval for the geometric mean.

truncation type 3 appears to be best for `Gm`, `Gs`, `Gp`, and `Ga`, all of which employ two or more linear predictors in the nodes. For `G1`, `G2`, and `G3`, which employ a simple polynomial in each node, the best truncation method is type 2 by a small margin.

3. THEORETICAL PROPERTIES

The empirical results suggest that when truncation is not beneficial, it is quite harmless. At worst, the prediction MSE is increased by one or two percent. This behavior can be explained by an asymptotic analysis. We prove here that, under weak regularity conditions when the true $f(x)$ is linear, any increase in prediction MSE due to truncation vanishes in the limit as the training sample size tends to infinity.

Let β_0 be a constant scalar and β_1 be a fixed p-dimensional vector (all vectors are column vectors here). For any p-dimensional vector x, define $f(x) = \beta_0 + x^t\beta_1$, where the superscript t denotes matrix transposition. Let $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ be a training sample such that $\{x_1, x_2, \ldots, x_n\}$ is a random sample of p-dimensional vectors from a distribution F_X and $y_i = f(x_i) + \varepsilon_i$, $i = 1, 2, \ldots, n$, where $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^t$ is a vector of n independent variables with mean 0 and variance σ^2. Define $\beta = (\beta_0, \beta_1^t)^t$ and let $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1^t)^t$ denote

the least squares estimate of β. The design matrix of the training sample is

$$Z = \begin{pmatrix} 1 & x_1^t \\ 1 & x_2^t \\ \vdots & \vdots \\ 1 & x_n^t \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix},$$

where $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})^t$. If $Z^t Z$ is invertible,

$$\hat{\beta} = (Z^t Z)^{-1} Z^t y = \beta + (Z^t Z)^{-1} Z^t \varepsilon. \tag{1}$$

Let $x_* = (x_{*1}, x_{*2}, \dots, x_{*p})^t$ be another independent observation from F_X and $y_* = f(x_*) + \varepsilon_*$ for some independent ε_* with mean 0 and variance σ^2. The least squares prediction for y_* is $\hat{y} = \hat{\beta}_0 + x_*^t \hat{\beta}_1$ and the expected squared prediction error is $\mathrm{MSE}(\hat{y}) = E_*(y_* - \hat{y})^2 = E_*\{f(x_*) - \hat{y}\}^2 + \sigma^2$. Here E_* denotes expectation taken over x_* and ε_*, with the training sample held fixed. For any vector x, let $|x|$ denote the Euclidean norm of x. Define the $(p+1) \times (p+1)$ random matrix

$$C = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ x_{11} & x_{11}^2 & x_{11}x_{12} & \dots & x_{11}x_{1p} \\ x_{12} & x_{12}x_{11} & x_{12}x_{12} & \dots & x_{12}x_{1p} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ x_{1p} & x_{1p}x_{11} & x_{1p}x_{12} & \dots & x_{1p}^2 \end{pmatrix}.$$

We assume the following two conditions throughout.

Condition 3.1. The expected matrix $E(C)$ is nonsingular.

Condition 3.2. $E_*|x_*|^2 < \infty$.

The first condition ensures that $E(C)$ is invertible. The second condition is necessary for the existence of certain expectations.

For $j = 1, 2, \dots, p$, define $a_j(n) = \min_{1 \le i \le n} x_{ij}$ and $b_j(n) = \max_{1 \le i \le n} x_{ij}$. Let $R_n = \cap_{j=1}^p [a_j(n), b_j(n)]$ be the smallest p-dimensional hyper-rectangle containing the set of points represented by the vectors $\{x_1, x_2, \dots, x_n\}$. The Winsorized predicted value $\dot{y}$ is given by $\dot{y} = \hat{\beta}_0 + x_w^t \hat{\beta}_1$, with the jth element of x_w being

$$x_{wj} = \begin{cases} x_{*j}, & a_j(n) \le x_{*j} \le b_j(n) \\ a_j(n), & x_{*j} < a_j(n) \\ b_j(n), & x_{*j} > b_j(n) \end{cases}$$

for $j = 1, 2, \dots, p$. Note that $x_w = x_*$ if $x_* \in R_n$. The expected squared prediction error of $\dot{y}$ is $\mathrm{MSE}(\dot{y}) = E_*\{f(x_*) - \dot{y}\}^2 + \sigma^2$.

We first prove two important lemmas. The result in the first lemma was derived by Lai et al. [1977] under weaker assumptions, but its proof is much harder.

LEMMA 3.3. *Under Conditions* 3.1 *and* 3.2, $\hat{\beta} \to \beta$ *with probability one as the training sample size tends to infinity.*

PROOF. Since

$$Z^tZ = \begin{pmatrix} 1 & \sum_i x_{i1} & \sum_i x_{i2} & \dots & \sum_i x_{ip} \\ \sum_i x_{i1} & \sum_i x_{i1}^2 & \sum_i x_{i1}x_{i2} & \dots & \sum_i x_{i1}x_{ip} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ \sum_i x_{ip} & \sum_i x_{ip}x_{i1} & \sum_i x_{ip}x_{i2} & \dots & \sum_i x_{ip}^2 \end{pmatrix},$$

we see that $(n^{-1}Z^tZ)^{-1} \to [E(C)]^{-1}$ with probability one. Further,

$$n^{-1}Z^t\epsilon = \begin{pmatrix} n^{-1}\sum_i \epsilon_i \\ n^{-1}\sum_i x_{i1}\epsilon_i \\ n^{-1}\sum_i x_{i2}\epsilon_i \\ \vdots \\ n^{-1}\sum_i x_{ip}\epsilon_i \end{pmatrix} \to \mathbf{0} \quad \text{almost surely.}$$

Hence, it follows from Eq. (1) that $\hat{\beta} = \beta + (n^{-1}Z^tZ)^{-1}(n^{-1}Z^t\varepsilon) \to \beta$ with probability one. □

LEMMA 3.4. *As $n \to \infty$, $x_* - x_w \to 0$ with probability one.*

PROOF. For each $j = 1, 2, \dots, p$, let $a_j(0) < b_j(0)$ be the endpoints of the support of the distribution of x_{*j}. Thus, $a_j(0) \le x_{*j} \le b_j(0)$ with probability one. On the other hand, $a_j(n) \to a_j(0)$ and $b_j(n) \to b_j(0)$ with probability one, as $n \to \infty$. Since $x_{*j} = x_{wj}$ if $a_j(n) \le x_{*j} \le b_j(n)$, it follows that $x_{*j} - x_{wj} \to 0$ with probability one, for every $j = 1, 2, \dots, p$. □

Let $\bar{R}_n$ denote the complement of the set R_n and let $I(A)$ denote the indicator function for the event A, that is, $I(A) = 1$ if the event A occurs, otherwise $I(A) = 0$.

THEOREM 3.5 (TRUNCATION TYPE 4). *Under Conditions* 3.1 *and* 3.2, $MSE(\hat{y}) - MSE(\dot{y}) \to 0$ *as $n \to \infty$ for almost every training sample sequence.*

PROOF. Observe that

$$\begin{aligned} & MSE(\hat{y}) - MSE(\dot{y}) \\ &= E_*(\hat{y} - \dot{y})\{\hat{y} + \dot{y} - 2f(x_*)\}I(x_* \in \bar{R}_n) \\ &= E_*(x_* - x_w)^t\hat{\beta}_1\{2(\hat{\beta}_0 - \beta_0) + 2x_*^t(\hat{\beta}_1 - \beta_1) + (x_w - x_*)^t\hat{\beta}_1\}I(x_* \in \bar{R}_n) \\ &= 2E_*(x_* - x_w)^t\hat{\beta}_1(\hat{\beta}_0 - \beta_0)I(x_* \in \bar{R}_n) \\ &\quad + 2E_*x_*^t(\hat{\beta}_1 - \beta_1)(x_* - x_w)^t\hat{\beta}_1 I(x_* \in \bar{R}_n) \\ &\quad + E_*\{(x_w - x_*)^t\hat{\beta}_1\}^2 I(x_* \in \bar{R}_n) \\ &= 2(\hat{\beta}_0 - \beta_0)\hat{\beta}_1^t E_*(x_* - x_w)I(x_* \in \bar{R}_n) \\ &\quad + (\hat{\beta}_1 - \beta_1)^t E_*x_*(x_* - x_w)^t I(x_* \in \bar{R}_n)\hat{\beta}_1 \\ &\quad + \hat{\beta}_1^t E_*(x_* - x_w)(x_* - x_w)^t I(x_* \in \bar{R}_n)\hat{\beta}_1. \end{aligned}$$

The terms on the right side of the last equation converge to 0 by Lemmas 3.3 and 3.4, the inequality $|x_* - x_w| \le |x_* - x_1| \le |x_*| + |x_1|$, and the dominated convergence theorem. □

A similar result holds for truncation types 1, 2, and 3. Given a constant $c \ge 0$, define

$$\tilde{y}_c = \begin{cases} y_{(n)} + cr_n, & \hat{y} > y_{(n)} + cr_n \\ \hat{y}, & y_{(1)} - cr_n \le \hat{y} \le y_{(n)} + cr_n \\ y_{(1)} - cr_n, & \hat{y} < y_{(1)} - cr_n. \end{cases}$$

THEOREM 3.6 (TRUNCATION TYPES 1, 2, AND 3). *Under Conditions* 3.1 *and* 3.2, $MSE(\hat{y}) - MSE(\tilde{y}_c) \to 0$ *as* $n \to \infty$ *for each* $c \ge 0$ *and almost every training sample sequence.*

PROOF. As in the previous proof, write

$$\begin{aligned} \mathrm{MSE}(\hat{y}) - \mathrm{MSE}(\tilde{y}_c) \;=\; & E_*[(\hat{y} - \tilde{y}_c)\{2f(x_*) - \hat{y} - \tilde{y}_c\}\{I(\hat{y} > y_{(n)} + cr_n) \\ & + I(\hat{y} < y_{(1)} - cr_n)\}]. \end{aligned}$$

Let $z_* = (1, x_{*1}, x_{*2}, \ldots, x_{*p})^t$ and A_n be the event that $\hat{y} > y_{(n)} + cr_n$. Then

$$\begin{aligned} & |E_*(\hat{y} - \tilde{y}_c)\{2f(x_*) - \hat{y} - \tilde{y}_c\}I(A_n)| \\ &= |E_*(\hat{y} - y_{(n)} - cr_n)(2f(x_*) - \hat{y} - y_{(n)} - cr_n)I(A_n)| \\ &= |E_*(\hat{y} - y_{(n)} - cr_n)\{2(f(x_*) - \hat{y}) + (\hat{y} - y_{(n)} - cr_n)\}I(A_n)| \\ &\le 2E_*|\hat{y} - y_{(n)} - cr_n|\,|f(x_*) - \hat{y}|I(A_n) + E_*(\hat{y} - y_{(n)} - cr_n)^2 I(A_n) \\ &\le 2E_*|\hat{y} - y_1|\,|f(x_*) - \hat{y}|I(A_n) + E_*(\hat{y} - y_1)^2 I(A_n) \\ &= 2E_*|z_*^t\hat{\beta} - y_1|\,|z_*^t(\beta - \hat{\beta})|I(A_n) + E_*(z_*^t\hat{\beta} - y_1)^2 I(A_n) \\ &\le 2E_*(|z_*|\,|\hat{\beta} - \beta| + |z_*|\,|\beta| + |y_1|)|z_*|\,|\hat{\beta} - \beta| + E_*(|z_*|\,|\beta| + |y_1|)^2 I(A_n) \\ &= 2|\hat{\beta} - \beta|\{(|\hat{\beta} - \beta| + |\beta|)E_*|z_*|^2 + |y_1|E_*|z_*|\} + E_*(|z_*|\,|\beta| + |y_1|)^2 I(A_n). \end{aligned}$$

Since $\hat{\beta} \to \beta$ and $I(A_n) \to 0$ with probability one for each x_*, the first term on the right side of the last inequality converges to 0. The second term also converges to 0 by the dominated convergence theorem, because $E_*(|z_*|\,|\beta| + |y_1|)^2 < \infty$. Therefore, $E_*(\hat{y} - \tilde{y}_c)\{2f(x_*) - \hat{y} - \tilde{y}_c\}I(\hat{y} > y_{(n)} + cr_n) \to 0$. It follows similarly that $E_*(\hat{y} - \tilde{y}_c)\{2f(x_*) - \hat{y} - \tilde{y}_c\}I(\hat{y} < y_{(1)} - cr_n) \to 0$. This completes the proof. □

4. COMPARISON WITH OTHER METHODS

So far, we have been studying the effectiveness of truncation versus no truncation for GUIDE. We now use the same datasets to compare the performance of the methods with other methods. Based on the conclusions in Section 2, we focus our attention on the GUIDE variants `Gm3`, `Gs3`, `Gp3`, `Ga3`, `G12`, `G22`, and `G32`. The other methods are

`rpart`: R [R Development Core Team 2005] implementation of the CART algorithm [Breiman et al. 1984].

`rF`: R implementation of the random forest algorithm [Breiman 2001] with the default of 500 trees.

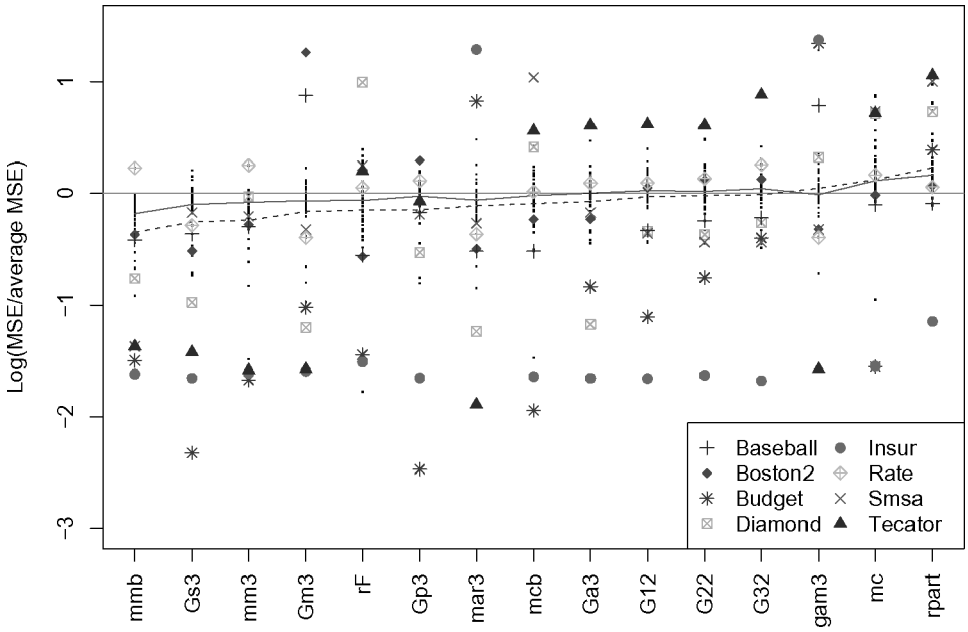

Fig. 5. Dot plots of log(MSE/average MSE), where the average MSE is over the fifteen methods. Methods are ordered according to their means, which are joined by a dashed broken line. The medians are joined by a solid broken line. Each dot or symbol refers to one dataset.

gam: R implementation of generalized additive model [Hastie and Tibshirani 1990].

mar: R implementation of multivariate adaptive regression splines [Friedman 1991].

mc: M5' [Witten and Frank 2005; Wang and Witten 1997] with constant node models. This is a modified version of the M5 algorithm [Quinlan 1992].

mcb: mc with bagging using the default of 10 trees.

mm: M5' with multiple linear node models.

mmb: mm with bagging using the default of 10 trees.

Also included are gam3, mar3, and mm3, the type-3 truncated versions of gam, mar, and mm, respectively. The barcharts in Figures 3 and 4 show that type-3 truncation is beneficial for these three methods too, although not as much as for the GUIDE methods.

For each dataset, we first normalize the estimated MSE of each method by dividing it by the average value over the fifteen methods: Gm3, Gs3, Gp3, Ga3, G12, G22, G32, rpart, rF, gam3, mar3, mc, mcb, mm3, and mmb. We call each of these values a "relative MSE." Then we take the (natural) log of the relative MSE. Besides being easier to visualize, the log scale renders differences between any two methods independent of the normalization. The results are shown graphically in Figure 5. The bagged version of M5' (mmb) has the smallest mean log relative MSE, followed by Gs3, mm3, and Gm3. The piecewise-constant methods, mc and rpart, have the highest means. Note that there is substantial variability in

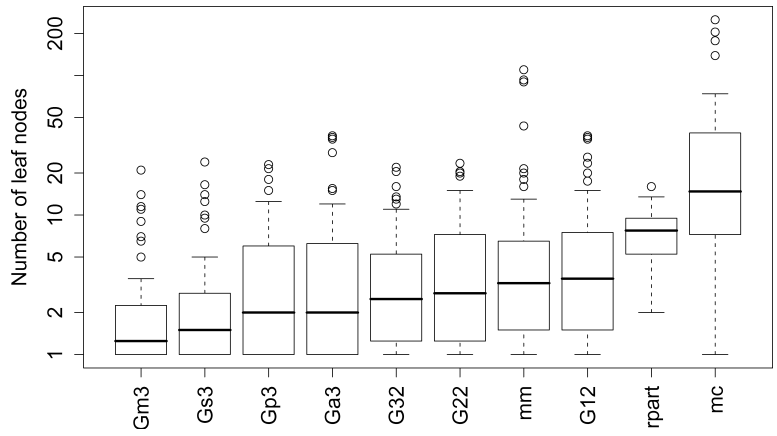

Fig. 6. Boxplots of number of leaf nodes ordered by medians

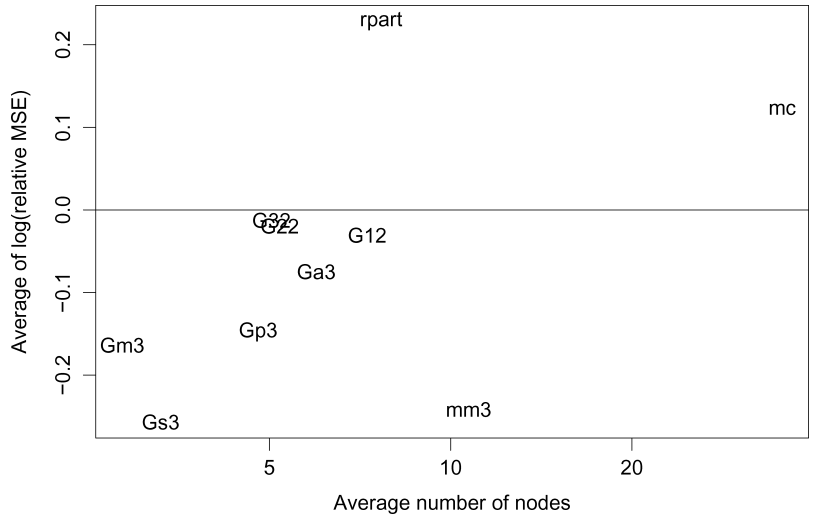

Fig. 7. Average of log(relative MSE) versus mean number of nodes. The mean number of nodes for `Gm3` and `Gs3` are 2.8 and 3.3, respectively.

each method across datasets. For example, although `mmb` is best in terms of this measure of performance, there is one dataset (`Rate`) for which its MSE is higher than the average MSE for all the methods. No method is best for all datasets.

Another important aspect of the methods that we have not considered is the size of the tree structures. Obviously, if two methods have the same prediction accuracy, the one yielding trees with fewer leaf nodes is preferred. Figure 6 shows boxplots of the number of leaf nodes for the regression tree methods and Figure 7 shows a plot of the average of the log relative MSE versus the average number of leaf nodes. The `mc` method has the largest average of 35.5 leaf nodes, followed by `mm3` and `rpart` with 10.7 and 7.7 leaf nodes, respectively. The `Gm3` and `Gs3` methods have the lowest averages of 2.8 and 3.3 leaf nodes, respectively.

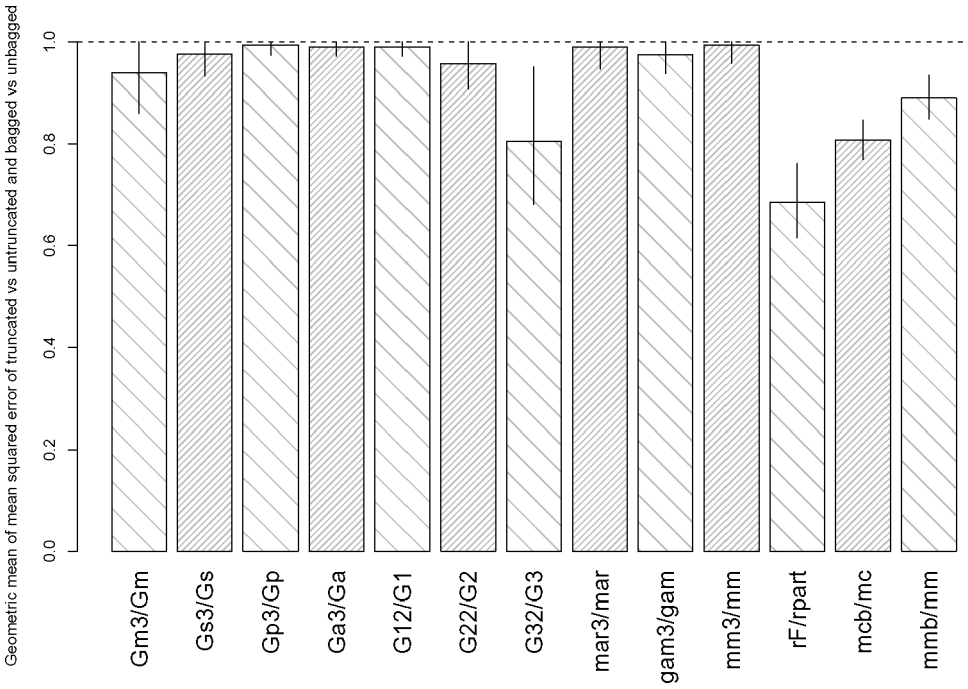

Fig. 8. Barchart of geometric means of mean squared prediction error of truncated vs untruncated and bagged vs unbagged methods for all fifty-two datasets. The vertical line at the end of each bar is an approximate 95% confidence interval for the geometric mean.

5. CONCLUSION

We have demonstrated empirically that there is usually some reduction in MSE from truncating or Winsorizing the predicted values from a model. Our results indicate that the amount of reduction tends to increase with the size of the test sample. In particular, the reduction is less if we had used ten-fold, instead of two-fold, cross-validation in in Section 2. This is probably due to the number of extrapolation errors increasing with the test sample size.

We have also established theoretically that truncation and Winsorization do not cause any harm asymptotically, in the case of a linear model. Thus, it is safe to routinely use truncation in real applications with large training sample sizes.

The empirical results provide further evidence that no single method is best for all datasets. Even the method with the best average performance, `mmb`, is below average for one dataset. On the other hand, bagging can be expected to improve the average performance of a method. Thus `mmb` is better than `mm`, `mcb` is better than `mc`, and `rF` is better than `rpart`, on average. But bagging does not always improve a method for every dataset. For example, Figure 5 shows that `mcb` is worse than `mc` on the `Smsa` dataset, and `rF` is worse than `rpart` on the `Diamond` dataset.

A more surprising result is that the accuracy of ensemble methods is not as great as might be expected. As Figure 5 shows, the non-ensemble `mar` is only

slightly inferior on average to rF. Similarly, several GUIDE methods are better, on average, than the ensemble methods mcb and rF. To compare the benefits from ensembling with those from truncation, we show the geometric means, over the fifty-two datasets, of the reduction in MSE of the truncated versus untruncated and ensemble versus nonensemble methods in Figure 8. The ensemble method rF yields the largest average reduction (about 30 percent) over rpart. The reductions are less for mc and mm (20 and 10 percent, respectively). Overall, truncation or ensembling yields the most improvement for the least accurate methods. It remains to be seen how much ensembling can help to reduce the prediction error of the GUIDE methods.

ACKNOWLEDGMENT

We are grateful to the reviewers for their comments.

REFERENCES

AABERGE, R., COLOMBINO, U., AND STROM, S. 1999. Labor supply in Italy: An empirical analysis of joint household decisions, with taxes and quantity constraints. *J. Appl. Econom. 14*, 403–422.

AFIFI, A. AND AZEN, S. 1979. *Statistical Analysis: A Computer Oriented Approach*, 2nd ed. Academic Press, New York.

BELSLEY, D. A., KUH, E., AND WELSCH, R. E. 1980. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Wiley, New York.

BERNDT, E. R. 1991. *The Practice of Econometrics*. Addison-Wesley, New York.

BLAKE, C. AND MERZ, C. 1998. *UCI Repository of Machine Learning Databases*, http://www.ics.uci.edu/~mlearn/MLRepository.html.

BOLLINO, C. A., PERALI, F., AND ROSSI, N. 2000. Linear household technologies. *J. Appl. Econom. 15*, 253–274.

BREIMAN, L. 2001. Random forests. *Mach. Learn. 45*, 5–32.

BREIMAN, L. AND FRIEDMAN, J. 1988. Estimating optimal transformations for multiple regression and correlation. *J. Amer. Stat. Assoc. 83*, 580–597.

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A., AND STONE, C. J. 1984. *Classification and Regression Trees*. Wadsworth, Belmont, CA.

BRYANT, P. G. AND SMITH, M. A. 1996. *Practical Data Analysis: Case Studies in Business Statistics*, vol. 3. Irwin/McGraw Hill, New York.

CHATTOPADHYAY, S. 2003. Divergence in alternative Hicksian welfare measures: The case of revealed preference for public amenities. *J. Appl. Econom. 17*, 641–666.

CHU, S. 2001. Pricing the C's of diamond stones. *J. Stat. Educat. 9*. http://www.amstat.org/publications/jse.

COCHRAN, J. J. 2000. Career records for all modern position players eligible for the Major League Baseball Hall of Fame. *J. Stat. Educat. 8*. http://www.amstat.org/publications/jse.

COCHRAN, J. J. 2002. Data management, exploratory data analysis, and regression analysis with 1969–2000 Major League Baseball Attendance. *J. Stat. Educat. 10*. http://www.amstat.org/publications/jse.

COOK, D. 1998. *Regression Graphics: Ideas for Studying Regression Through Graphics*. Wiley, New York.

COOK, D. AND WEISBERG, S. 1994. *An Introduction to Regression Graphics*. Wiley, New York.

DEB, P. AND TRIVEDI, P. K. 1997. Demand for medical care by the elderly: A finite mixture approach. *J. Appl. Econom. 12*, 313–336.

DENMAN, N. AND GREGORY, D. 1998. Analysis of sugar cane yields in the Mulgrave area, for the 1997 sugar cane season. Tech. rep., MS305 Data Analysis Project, Department of Mathematics, University of Queensland, Queensland, Australia.

DELGADO, M. A. AND MORA, J. 1998. Testing non-nested semiparametric models: An application to Engel curves specification. *J. Appl. Econom. 13*, 145–162.

FERNANDEZ, C., LEY, E., AND STEEL, M. F. J. 2002. Bayesian modelling of catch in a north-west Atlantic fishery. *Appl. Stat. 51*, 257–280.

FRIEDMAN, J. 1991. Multivariate adaptive regression splines (with discussion). *Ann. Stat. 19*, 1–141.

HAIR, J. F. ANDERSON, R. E., TATHAM, R. L., AND BLACK, W. C. 1998. *Multivariate Data Analysis*. Prentice Hall, Englewood Cliffs, NJ.

HALLIN, M. AND INGENBLEEK, J.-F. 1983. The Swedish automobile portfolio in 1977: A statistical study. *Scand. Actuarial J. 83*, 49–64.

HAMPEL, F. R., RONCHETTI, E. M., ROUSSEEUW, P. J., AND STAHEL, W. A. 1986. *Robust Statistics: The Approach Based on Influence Functions*. Wiley, New York.

HARRELL, JR., F. E. 2001. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. Springer-Verlag, New York.

HASTIE, T. AND TIBSHIRANI, R. 1990. *Generalized Additive Models*. CRC Press.

HORRACE, W. C. AND SCHMIDT, P. 2000. Multiple comparisons with the best, with economic applications. *J. Appl. Econom. 15*, 1–26.

KENKEL, D. S. AND TERZA, J. V. 2001. The effect of physician advice on alcohol consumption: Count regression with an endogenous treatment effect. *J. Appl. Economet. 16*, 165–184.

KIM, H., LOH, W.-Y., SHIH, Y.-S., AND CHAUDHURI, P. 2007. A visualizable and interpretable regression model with good prediction power. *IIE Transactions 39*, 565–579.

LAI, T. L., ROBBINS, H., AND WEI, C. Z. 1977. Strong consistency of least squares estimates in multiple regression. *Proc. Nat. Acad. Sci., USA 75*, 3034–3036.

LAROQUE, G. AND SALANIE, B. 2002. Labor market institutions and employment in France. *J. Appl. Econom. 17*, 25–28.

LIU, Z. AND STENGOS, T. 1999. Non-linearities in cross country growth regressions: A semiparametric approach. *J. Appl. Econom. 14*, 527–538.

LOH, W.-Y. 2002. Regression trees with unbiased variable selection and interaction detection. *Stat. Sinica 12*, 361–386.

LUTKEPOHL, H., TERASVIRTA, T., AND WOLTERS, J. 1999. Investigating stability and linearity of a German M1 money demand function. *J. Appl. Econom. 14*, 511–525.

MARTINS, M. F. O. 2001. Parametric and semiparametric estimation of sample selection models: An empirical application to the female labour force in Portugal. *J. Appl. Econom. 16*, 23–40.

NETER, J., KUTNER, M. H., NACHTSHEIM, C. J., AND WASSERMAN, W. 1996. *Applied Linear Statistical Models*, 4th ed. Irwin.

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. *J. Appl. Econom. 13*, 543–565.

ONOYAMA, K., OHSUMI, N., MITSUMOCHI, N., AND KISHIHARA, T. 1998. Data analysis of deer-train collisions in eastern Hokkaido, Japan. In *Data Science, Classification, and Related Methods*, (Tokyo, Japan) C. Hayashi, N. Ohsumi, K. Yajima, Y. Tanaka, H.-H. Bock, and Y. Baba, Eds. Springer-Verlag, New York, 746–751.

PACE, R. K. AND BARRY, R. 1997. Sparse spatial autoregressions. *Stat. Probab. Lett. 33*, 291–297.

PENROSE, K., NELSON, A., AND FISHER, A. 1985. Generalized body composition prediction equation for men using simple measurement techniques. *Med. Sci. Sports Exer. 17*, 189.

QUINLAN, J. R. 1992. Learning with continuous classes. In *Proceedings of the Australian Joint Conference on Artificial Intelligence* (Singapore), World Scientific, 343–348.

R DEVELOPMENT CORE TEAM. 2005. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, (Vienna, Austria). ISBN 3-900051-07-0.

RAWLINGS, J. O. 1988. *Applied Regression Analysis: A Research Tool*. Wadsworth & Brooks/Cole Advanced Books & Software.

SCHAFGANS, M. M. 1998. Ethnic wage differences in Malaysia: Parametric and semiparametric estimation of the Chinese-Malay wage gap. *J. Appl. Econom. 13*, 481–504.

SIMONOFF, J. 1996. *Smoothing Methods in Statistics*. Springer-Verlag, New York.

TORGO, L. 1999. *Inductive Learning of Tree-Based Regression Models*. PhD thesis, Department of Computer Science, Faculty of Sciences, University of Porto.

WANG, Y. AND WITTEN, I. 1997. Inducing model trees for continuous classes. In *Proceedings of the Poster Papers of the European Conference on Machine Learning* (Prague).

WEISS, S. AND INDURKHYA, N. 1995. Rule-based machine learning methods for functional prediction. *J. Artif. Int. Res. 3*, 383–403.

WITTEN, I. AND FRANK, E. 2005. *Data Mining: Practical Machine Learning Tools and Techniques with JAVA Implementations*, 2nd ed. Morgan Kaufmann, San Fransico, CA. http://www.cs.waikato.ac.nz/ml/weka.

Received August 2006; revised April 2007; accepted April 2007