Supplementary materials for this article are available online.
Please click the JCGS link at http://pubs.amstat.org.

Combining an Additive and Tree-Based Regression Model Simultaneously: STIMA

Elise DUSSELDORP, Claudio CONVERSANO, and Bart Jan VAN OS

Additive models and tree-based regression models are two main classes of statistical models used to predict the scores on a continuous response variable. It is known that additive models become very complex in the presence of higher order interaction effects, whereas some tree-based models, such as CART, have problems capturing linear main effects of continuous predictors. To overcome these drawbacks, the regression trunk model has been proposed: a multiple regression model with main effects and a parsimonious amount of higher order interaction effects. The interaction effects can be represented by a small tree: a regression trunk. This article proposes a new algorithm—Simultaneous Threshold Interaction Modeling Algorithm (STIMA)—to estimate a regression trunk model that is more general and more efficient than the initial one (RTA) and is implemented in the R-package `stima`. Results from a simulation study show that the performance of STIMA is satisfactory for sample sizes of 200 or higher. For sample sizes of 300 or higher, the 0.50 SE rule is the best pruning rule for a regression trunk in terms of power and Type I error. For sample sizes of 200, the 0.80 SE rule is recommended. Results from a comparative study of eight regression methods applied to ten benchmark datasets suggest that STIMA and GUIDE are the best performers in terms of cross-validated prediction error. STIMA appeared to be the best method for datasets containing many categorical variables. The characteristics of a regression trunk model are illustrated using the Boston house price dataset.

Supplemental materials for this article, including the R-package `stima`, are available online.

Key Words: Boston house price data; Interaction effects; Recursive partitioning; Threshold interactions.

Elise Dusseldorp is Assistant Professor, Data Theory Group, Department of Education, Faculty of Social and Behavioral Sciences, Leiden University, PO Box 9555, 2300 RB Leiden, The Netherlands and Statistician, TNO Quality of Life, PO Box 2215, 2301 CE Leiden, The Netherlands (E-mail: *elise.dusseldorp@tno.nl*). Claudio Conversano is Researcher, Department of Economics, Faculty of Economics, University of Cagliari, I-09123, Viale Frá Ignazio 17, Cagliari, Italy (E-mail: *conversa@unica.it*). Bart Jan Van Os is Assistant Professor, Data Theory Group, Department of Education, Faculty of Social and Behavioral Sciences, Leiden University, PO Box 9555, 2300 RB Leiden, The Netherlands (E-mail: *bartjan@4os.nl*).

Journal of Computational and Graphical Statistics, Volume 19, Number 3, Pages 514–530
DOI: 10.1198/jcgs.2010.06089

1. INTRODUCTION

Additive models and tree-based regression models are two main classes of statistical models used to predict the scores on a continuous response variable Y. The influence of predictor variables X_j (with $j = 1, \ldots, J$) on Y is often described by an additive model in the following way:

$$\hat{Y} = \hat{\alpha} + \sum_{j=1}^{J} \hat{f}_j(X_j) \tag{1.1}$$

in which the $f(\cdot)$'s are arbitrary univariate functions (Hastie and Tibshirani 1990); this model is also known as a main effects model. When the $f(\cdot)$'s in (1.1) are multivariate functions, we enter the field of interaction effects. In a multiple regression framework, a strict definition of interaction is applied: two predictors are said to interact in their accounting for variance in Y when "over and above any additive combination of their separate effects, they have a joint effect" (Cohen et al. 2003, p. 257). A number of factors increase the complexity of additive models with interaction effects: (a) the presence of interactions induced by categorical variables with more than two categories, (b) the presence of higher order interactions, and (c) the absence of a priori hypotheses about possible interaction effects when the data contain many predictors. In such situations, an additive regression model will soon include many parameters to be estimated and the effects will often be difficult to interpret.

Tree-based models, such as Classification and Regression Trees (CART, Breiman et al. 1984), provide a solution to these complex interactive situations. They were originally developed for assessing interaction effects; one of the earliest implementations was called Automatic Interaction Detection (AID, Morgan and Sonquist 1963). Interaction effects are represented as *threshold* interactions. Such interaction effects occur when the effect of a predictor on a response is different for observations with a score above a certain threshold value on another predictor compared to observations with a score below that threshold value. If we consider a tree of size M (i.e., M terminal nodes) and splits on predictors X_j, the corresponding model is defined as

$$\hat{Y} = \sum_{m=1}^{M} \hat{\beta}_m I((X_1, \ldots, X_J) \in R_m), \tag{1.2}$$

where $I(\cdot)$ is an indicator function with value 1 if a subject belongs to region R_m and value 0 if not. R_m is defined by the predictors used in the splits leading to that region (i.e., terminal node). Tree-based models can identify complex interactive situations more easily than additive models (Harrel Jr. 2001). However, if the data also contain main effects of some predictors, piecewise-constant trees, such as CART, have difficulty capturing these effects. CART "would take many fortuitous splits to recreate the structure, and the data analyst would be hard pressed to recognize it in the estimated tree" (Hastie, Tibshirani, and Friedman 2001, p. 274).

To overcome the above-mentioned problems concerning the identification of main and interaction effects, a number of methods have been developed to estimate models incorporating both additive and interactive influences of predictors, such as multivariate adaptive regression splines (MARS, Friedman 1991), M5 (Quinlan 1992), Treed Regression (Alexander and Grimshaw 1996), and GUIDE (Generalized Unbiased Interaction Detection and Estimation, Loh 2002).

In this article, we elaborate on a relatively new model—the regression trunk model—which integrates a multiple regression model and a regression tree (Dusseldorp and Meulman 2004). The model is especially appropriate for prediction problems with multiple predictors and a combination of linear main effects and interaction effects for the same or different sets of predictors. It is also a suitable choice for problems with no exact a priori hypotheses about the number and order of the interaction effects. In the regression trunk model, joint effects of predictors are estimated over and above their separate effects. In most situations, only a small regression tree—a so-called *regression trunk*—is needed to capture the interactive structure; hence the name of the model. While both M5 and Treed Regression estimate several linear models—one in each node of a tree—the regression trunk model is only a single linear model. The model can be defined as

$$\hat{Y} = \hat{\beta}_0 + \sum_{j=1}^{J} \hat{\beta}_j X_j + \sum_{m=1}^{M-1} \hat{\beta}_{J+m} I((X_1, \ldots, X_J) \in R_m), \tag{1.3}$$

where M denotes the total number of terminal nodes (i.e., the size of the trunk). The total number of indicator variables equals $M - 1$; the region that is not included serves as reference group. The estimated intercept for the reference region is indicated by $\hat{\beta}_0$, and the estimated intercept for region R_m by $\hat{\beta}_0 + \hat{\beta}_{J+m}$. In CART, all the subjects in a region receive the same predicted value, whereas in the regression trunk the predicted value may differ, depending on the estimated slopes $\hat{\beta}_j$ of the regression lines for the predictors. Dusseldorp and Meulman (2004) proposed a three-step algorithm to estimate (1.3), which they called the Regression Trunk Approach (RTA). We propose a new algorithm called STIMA (Simultaneous Threshold Interaction Modeling Algorithm) that simultaneously estimates a linear regression and a tree model. The simultaneous approach enables us to take into account correlations between main and interaction effects, and speeds up the cross-validation procedure, which was very time consuming in RTA. STIMA is more general than RTA in the sense that it is applicable to all types of predictors (categorical and continuous).

We performed a simulation study aimed at answering the following questions: What sample size do we need to obtain sufficient power to detect threshold interaction effects? And which pruning rule should we apply to determine the correct size of the regression trunk? In other words: How do we detect interactions when we have multiple predictors available, and no a priori hypotheses about their joint effects, that is, the number, type, and order of their interaction effects? STIMA is applied to benchmark datasets and its performance is compared to reference regression methods.

2. STIMA: THE REGRESSION TRUNK ALGORITHM

STIMA works by first growing a full regression trunk, corresponding to the maximum number of splits L, and then pruning it back using V-fold cross-validation with the c standard error rule (c SE rule), where V denotes the number of subsets used in the cross-validation procedure and c is a pruning parameter (see Section 2.2 for details about the pruning procedure). The user has to provide the parameter L. The predictor for the first split of the regression trunk—for convenience denoted with X_1—can be automatically selected (default option) or can be defined by the user. The automatic selection is performed in the following way: A regression trunk with L splits is estimated using X_j as first splitting predictor ($j = 1, \ldots, J$), and the cross-validated error of the regression trunk model is computed for each size ℓ of the trunk[1] ($\ell = 1, \ldots, L$). This results in $J \times L$ regression trunks, and $J \times L$ estimated cross-validated errors. These estimates are compared, and the predictor X_j generating the minimum cross-validated error is selected as the best candidate for the first split. The user-defined selection of the first splitting predictor is motivated by a specific application of interaction research, namely treatment covariate interaction (see Dusseldorp and Meulman 2004; Dusseldorp et al. 2007), where interactions with a treatment variable indicate differential treatment effectiveness. In such a situation, the treatment variable is used for the first split.

2.1 THE TREE GROWING PROCEDURE OF STIMA

A schematic formulation of the algorithm is given in Appendix A (see online supplementary material). The algorithm starts with the estimation of a linear main effects model in the root node of the regression trunk using all predictors. Next, if the first splitting predictor X_1 is categorical, the main effect of X_1 is removed from the root node model to avoid linear dependency. If X_1 is categorical with K categories, then the regression trunk will have K nodes after the first split ($M_1 = K$). The corresponding regression model will include $K - 1$ indicator variables, as is common practice in multiple regression analysis; these indicator variables represent the main effect of X_1. The remaining splits of the regression trunk are binary. If X_1 is continuous, then all splits of the regression trunk are binary (see step 2, Algorithm 1, Appendix A, online supplementary material).

Possible splits $\ell = 1, \ldots, L$ are evaluated using the following partitioning criterion: The best split point of a predictor variable is the one that induces the highest effect size starting from the model estimated before split ℓ to the estimated model after split ℓ. The effect size is defined as $f_\ell^2 = (\rho_\ell^2 - \rho_{\ell-1}^2)/(1 - \rho_\ell^2)$ (Cohen 1988, p. 410), where the squared multiple correlation coefficient of, for example, the model after split ℓ equals $\rho_\ell^2 = \sum_i (\hat{Y}_{i\ell} - \bar{Y})^2 / \sum_i (Y_i - \bar{Y})^2$. This effect size corresponds to the relative increase in variance-accounted-for. Each new split ℓ is found by scanning through all predictors X_j, all possible split points (i.e., the observed values of X_j) at all current terminal nodes retaining the one with the largest effect size. If any predictor X_j is categorical, an extra step

[1]The size of a trunk can be expressed in terms of number of splits ℓ or number of terminal nodes M. In the case of a binary regression trunk: $M = \ell + 1$.

is performed that essentially orders the categories with respect to their average residual values, such that the subsequent split scanning is significantly reduced (see Algorithm 2, Appendix A, online supplementary material). Each subsequent split ℓ introduces one extra indicator variable to the model, after which all regression coefficients of the model are re-estimated (see step 3, Algorithm 1, Appendix A, online supplementary material). The final regression trunk model after L splits corresponds to

$$\hat{Y}_L = \hat{\beta}_{0L} + \sum_{j=1+d}^{J} \hat{\beta}_{jL} X_j + \sum_{m=1}^{M_L-1} \hat{\beta}_{J+mL} I((X_1, \ldots, X_J) \in R_{mL}), \tag{2.1}$$

where $d = 0$ if X_1 is continuous, and $d = 1$ if X_1 is categorical. Nonlinear main effects can be incorporated, for example, by adding quadratic (and cubic) terms of predictors to the input N by J data-matrix $\mathbf{X}$, or by transforming predictors a priori.

2.2 THE PRUNING PROCEDURE OF STIMA

Once the larger regression trunk has been grown, it is pruned using V-fold cross-validation (CV) in the same way as it is used in CART (Breiman et al. 1984). The user has to provide the number of subsets (V) as well as the parameter c that allows us to select the final regression trunk model using the c SE rule. The selection is made as follows: Let RE_ℓ^{cv} and SE_ℓ^{cv} denote the estimated relative cross-validated error and standard error of a regression trunk with ℓ splits:

$$\mathrm{RE}_\ell^{cv} = \sum_{v=1}^{V} \sum_{i \in S_v} \left(Y_i^{(v)} - \hat{Y}_{i\ell}^{(v)}\right)^2 \Big/ \sum_{i=1}^{N} (Y_i - \bar{Y}_i)^2,$$

$$\mathrm{SE}_\ell^{cv} = \frac{\sqrt{\sum_{i=1}^{N} [(Y_i - \hat{Y}_{i\ell}^{cv})^2 - N^{-1} \sum_{i=1}^{N} (Y_i - \hat{Y}_{i\ell}^{cv})^2]^2}}{\sum_{i=1}^{N} (Y_i - \bar{Y}_i)^2},$$

where S_v denotes the vth set playing the role of test set, and $\hat{Y}_{i\ell}^{(v)}$ denotes the predicted value of observation i in S_v. These predicted values are computed using the parameter estimates of the regression trunk of size ℓ fitted for the training set (i.e., all persons who are not in S_v). $\hat{Y}_{i\ell}^{cv}$ is a vector of length N joining the predicted values from all test sets $\hat{Y}_{i\ell}^{(v)}$.

We use ℓ^* to denote the size of the regression trunk with the lowest RE_ℓ^{cv}. Then, the size of the pruned regression trunk (ℓ^{**}) corresponds to the minimum value of ℓ such that $\mathrm{RE}_{\ell^{**}}^{cv} \leq \mathrm{RE}_{\ell^*}^{cv} + c\mathrm{SE}_{\ell^*}^{cv}$. This pruning rule, which will be called the c SE rule, takes into account that the prediction error is estimated with error.

For small sample sizes, the estimates of RE^{cv} might vary considerably from one CV procedure to another ($0.01 \leq |\mathrm{RE}_1^{cv} - \mathrm{RE}_2^{cv}| \leq 0.10$). A more stable estimate of RE^{cv} can be achieved by repeating the CV procedure, and afterward averaging the results. Once the regression trunk model is pruned to its right size, it is possible to apply a backward selection procedure. In this procedure both main effects and interaction effects can be removed from the model. We implemented the same backward selection procedure as the one implemented in multiple regression in R. All computations in this article are based on an

implementation of STIMA within the R system for statistical computing (R Development Core Team 2009).

3. THE CHOICE OF THE PRUNING PARAMETER *c*: A SIMULATION STUDY

We compared the performance of seven pruning rules, all derived from the general c SE rule. Seven different values of c were chosen, ranging from 0 to 1. Results from a pilot study suggested that the pruning rule with $c = 0$ was too liberal: The regression trunk with the minimum RE^{cv} appeared to be too large in most cases, implying too many interaction terms in the model. The pruning rule with $c = 1$ was too conservative: The size of the selected regression trunk was too small, implying too few interaction terms. Performance of a pruning rule was indicated by its ability to detect the true size of the regression trunk (detection rate). Furthermore, by creating true models that include only linear main effects of predictors (i.e., no threshold interactions), we computed the probability that the pruning rule would select a regression trunk model that includes one or more threshold interaction terms. This probability was regarded as the Type I error.

3.1 GAUGE

The population data were generated using two different regression models: one with a threshold interaction term between two continuous variables and one with a threshold interaction term between a categorical and a continuous variable. The models are introduced in (3.1) and (3.2) respectively:

$$Y_{true} = \sum_{i=1}^{3} w_i X_i + w_4 I(X_1 > 0.00) I(X_2 > 0.50) + 0.71e, \tag{3.1}$$

$$Y_{true} = \sum_{i=1}^{3} w_i X_i + w_4 X_4 I(X_1 > 0.53) + 0.71e. \tag{3.2}$$

All predictors were continuous, standard normally distributed variables, except for predictor X_4, which consisted of two categories, coded with values 1 and 0. The number of subjects in each category of X_4 was $0.50N$. Subjects were randomly assigned to the two categories. The resulting correlation between X_4 and any other predictor was approximately zero.

The population threshold values in (3.1) and (3.2) were fixed in such a way that 15% of the subjects had a value 1 on the interaction variables. The population weights (w_1 to w_4) are set in accordance with the magnitude of the effect size of the interaction term, a design factor explained hereafter. The variable e denotes the error, randomly generated using $N(0, 1)$. An error weight of 0.71 was used, implying that the variance-accounted-for by e is about 0.50. An examination of application studies in behavioral sciences with treatment interaction effects revealed that most squared multiple correlation coefficients did not exceed a value of 0.50 (see Dusseldorp and Meulman 2004).

3.1.1 Design Factors

We systematically varied three independent variables in a full factorial design:

- The value of the *effect size* (f^2) of the threshold interaction term, having three levels: no effect, a medium effect, and a large effect. These effect sizes were realized by varying the values of weight w_4 in (3.1) and in (3.2). The other weights were adapted in such a way that the variance of Y_{true} was approximately 1. See Appendix B of the online supplementary material for the exact values of the weights. The "no effect" situation corresponds to $f^2 = 0$, and it is the situation where the true model includes only linear main effects of predictors. A medium and a large effect correspond to $f^2 = 0.10$ and $f^2 = 0.33$ (Cohen 1988, pp. 413–414).
- The sample size (N), having four levels: 100, 200, 300, and 500.
- The population correlations between the continuous predictors X_1, X_2, and X_3. These were all fixed at either 0.00 or 0.30.

For each combination of effect size × sample size × population correlation, 1000 random samples were drawn. For each random sample, STIMA was applied, using 10-fold cross-validation, a maximum of five splits ($L = 5$), and automatic selection of the first splitting variable. All variables were used in the main effects model, and were splitting candidates for the regression trunk. For each regression trunk, we applied seven pruning rules. The Type I error of a pruning rule was estimated in the situation where $f^2 = 0$. It was expressed as the proportion of the number of times the rule did not select the linear main effects model. For the continuous by continuous interaction situation in (3.1) this implied that a regression trunk with more than one terminal node was selected ($\ell^{**} > 0$). For the categorical by continuous interaction situation in (3.2) this implied that a regression trunk with more than two terminal nodes was selected ($\ell^{**} > 1$).

The power to detect the correct size of the regression trunk was estimated in the situations with a medium and a large f^2. It was expressed as the proportion of the number of times the pruning rule detected the true threshold interaction effect, that is, the true size ($\ell^{**} = 2$) and form of the regression trunk, implying the right product of indicator variables. The order of the splits was allowed to vary. Also, the split points were allowed to vary, due to random sampling fluctuations.

In a pilot study, we tested the number of times the 10-fold cross-validation procedure should be repeated for the different sample sizes to result in a stable estimate for RE^{cv} (up to two decimals). With $N = 100$, 10 times 10-fold cross-validation was satisfactory. With $N = 200$ and 300, 5 times was satisfactory; with $N = 500$, once was satisfactory.

3.2 RESULTS

Depending on which pruning rule is used, the Type I error varies considerably (see Table 1). Because the correlated predictors condition results in Type I error values similar to those of the independent predictors condition, only the latter will be displayed. The categorical by continuous situation in (3.2) results in somewhat lower Type I error values than the continuous by continuous situation in (3.1). Summarizing the results for both

Table 1. Type I error of each pruning rule. The true models are linear main effects models derived from the models given in (3.1) and (3.2). The proportion of the number of times that a linear main effects model was not selected is displayed in each cell, for independent predictors and correlated predictors. The pruning rules are variations of the c SE rule, differing in their value for c. The column labels denote the different sample sizes. Type I errors ≥ 0.05 are in boldface.

	Model (3.1)				Model (3.2)			
c	100	200	300	500	100	200	300	500
0.00	**0.832**	**0.811**	**0.802**	**0.791**	**0.300**	**0.352**	**0.352**	**0.462**
0.20	**0.502**	**0.395**	**0.322**	**0.301**	**0.180**	**0.178**	**0.142**	**0.190**
0.40	**0.266**	**0.155**	**0.079**	**0.068**	**0.102**	**0.069**	0.047	0.043
0.50	**0.195**	**0.097**	0.037	0.030	**0.069**	**0.050**	0.025	0.015
0.60	**0.137**	**0.064**	0.021	0.009	0.049	0.026	0.015	0.008
0.80	**0.064**	0.024	0.004	0.002	0.020	0.013	0.003	0.001
1.00	0.030	0.008	0.001	0.000	0.008	0.005	0.001	0.000

situations: For a sample size of 300 or larger, the pruning rules with a value of $c \geq 0.50$ result in acceptable Type I errors (i.e., <0.05). For a smaller sample size a higher value of c is needed to achieve acceptable Type I errors (for $N = 200$: $c \geq 0.80$; for $N = 100$: $c = 1.00$).

Table 2 displays the estimated detection rates for the independent predictors condition. The rates for the correlated predictors condition are on average somewhat lower (0.08 for the continuous by continuous situation, and 0.01 for the categorical by continuous situation). If we regard a minimum detection rate of 0.80 as being satisfactory, then the

Table 2. Detection rate per pruning rule of true *medium-sized* and *large-sized* interaction effects ($f^2 = 0.10$ and $f^2 = 0.33$). The results are displayed for a continuous by continuous interaction, model (3.1), and for a categorical by continuous interaction, model (3.2). The pruning rules are variations of the c SE rule, differing in their value of c. The column numbers denote the sample sizes.

	Model (3.1)				Model (3.2)			
c	100	200	300	500	100	200	300	500
				Medium-sized effect				
0.00	0.225	0.531	0.690	0.714	0.303	0.608	0.750	0.739
0.20	0.190	0.515	0.713	0.852	0.285	0.636	0.812	0.863
0.40	0.154	0.430	0.669	0.911	0.244	0.591	0.800	0.924
0.50	0.135	0.382	0.607	0.892	0.230	0.541	0.775	0.918
0.60	0.116	0.334	0.545	0.855	0.207	0.474	0.689	0.894
0.80	0.080	0.232	0.417	0.730	0.151	0.349	0.509	0.786
1.00	0.044	0.142	0.274	0.566	0.107	0.246	0.383	0.637
				Large-sized effect				
0.00	0.682	0.771	0.812	0.763	0.758	0.813	0.843	0.772
0.20	0.713	0.872	0.892	0.900	0.809	0.898	0.914	0.904
0.40	0.705	0.931	0.963	0.982	0.816	0.950	0.961	0.974
0.50	0.690	0.939	0.975	0.988	0.807	0.960	0.970	0.986
0.60	0.686	0.947	0.980	0.994	0.802	0.961	0.974	0.991
0.80	0.638	0.943	0.985	0.995	0.770	0.959	0.978	0.992
1.00	0.562	0.919	0.984	0.996	0.709	0.937	0.979	0.992

required sample size is larger than 300 for a medium-sized interaction effect, and larger than or equal to 200 for a large-sized effect. In general, the pruning rules with a value of $0.20 \leq c \leq 0.60$ result in a good detection rate, provided that the sample is of the required size. For sample sizes of 100, a pruning rule with $c = 0.60$ only results in an acceptable Type I error and detection rate in the categorical by continuous situation. When we combine these results, we may conclude that in general the performance of STIMA is satisfactory for sample sizes of 200 or larger. For sample sizes of 200, a pruning rule with $c = 0.80$ can be recommended, to guarantee that the Type I error remains below 0.05. For sample sizes of 300 and larger a pruning rule with $c = 0.50$ is the best performer in terms of power and Type I error.

4. EMPIRICAL EVIDENCE

The effectiveness of STIMA was tested on 10 datasets, and its performance was compared with other well-known alternative approaches. Table 3 shows the main features of the benchmark datasets. A more detailed description can be found in Appendix C (see online supplementary material).

For each dataset, we started from the *Linear main effects Model* (*LM*), also including categorical predictors which were coded as dummy variables. Next, all interactions induced by cross-products of each dummy variable with all continuous predictors were added to the model. A stepwise variable selection was performed on the estimated LM to obtain a *stepwise linear model with interaction terms* (*Step LM*). The presence of nonlinear structures in the data was investigated with *Regression Trees* (*RPART*): The pruned tree was selected using 10-fold cross-validation with the 1 SE rule. In addition, we considered the *stepwise Generalized Additive Model* (*Step GAM*) by searching for nonlinear specifications of scatterplot smoothers (smoothing spline and lowess) associated with each predictor. A second stepwise variable selection was performed by considering all interactions induced by cross-products of predictors selected in Step GAM in order to estimate a

Table 3. The main features of the benchmark datasets.

Dataset	Source	#cases	Response	#cont.	#cat.
Abalone	UCI	4177	Rings	7	1
Baseball	Statlib	263	log(salary)	16	6
Boston	Statlib	506	Median values of houses	14	1
Cars origin	UCI	205	Price	16	9
Employee	SPSS	473	Current salary	5	3
Eurostoxx	Bloomberg	397	1-yr return	17	1
EUR/USD	Bloomberg	255	EUR/USD exchange rate	12	0
Fev	JSE	654	Forced expiratory volume	2	2
Juice	Textbook	1070	Price juice CH	4	2
Home prices	Statlib-DASL	117	Price	3	3

NOTE: #cont. and #cat. are the number of continuous and categorical predictors in each dataset; #cases is the number of observations for each dataset; UCI = UCI Machine Learning Repository (Asuncion and Newman 2007); Statlib general website: *http://lib.stat.cmu.edu/datasets*; Statlib-DASL website: *http://lib.stat.cmu.edu/DASL/*; JSE = *Journal of Statistics Education*, website: *http://www.amstat.org/publications/jse*. The Juice dataset can be found in Foster, Stine, and Waterman (1998).

GAM with interactions (*Step GAM_ int*). As alternative approaches, *Multivariate Adaptive Regression Spline* (*MARS*) and *Generalized, Unbiased Interaction Detection and Estimation* (*GUIDE*) were considered. The estimated MARS was pruned by backward selection. A stepwise least squares estimation with an exhaustive search split point selection method was performed for GUIDE, where the tree was pruned using the naive SE estimate and performing 10-fold cross-validation with the 1 SE rule. We used the `guide` software provided by Loh (2002) to estimate the model, whereas the R software was used to estimate all the other above-mentioned models (packages `stats`, `rpart` (Therneau and Atkinson 1997), `gam` (Hastie 2009), and `mda` (Leisch, Hornik, and Ripley 2009) were used for LM, RPART, GAM, and MARS, respectively).

The performance of these models was compared with STIMA, which was estimated by using the R-package `stima`[2]: The pruned regression trunk was selected by using 10-fold cross-validation with the 0.50 SE rule for datasets with $N \geq 300$ or with the 0.80 SE rule for small datasets and by selecting the final model through backward variable selection. The same random division in ten subsets was used to estimate the 10-fold cross-validated error for all the compared methods.

4.1 Main Results

The different approaches are compared in terms of (a) the prediction error of the model, indicated by the 10-fold cross-validated error (RE^{cv}); (b) the stability, expressed by the standard error of the cross-validated error (SE^{cv}); and (c) the model parsimony, taking into account the number of parameters included in each model. We consider a model with fewer parameters to be more parsimonious. With regard to the tree-based methods, each extra split in a tree induces an extra parameter in the model, so that the number of parameters for RPART and GUIDE corresponds to the number of terminal nodes of the tree minus one. For STIMA, the total number of parameters corresponds to the number of regression coefficients in the selected regression trunk model. The main results of this comparative analysis are summarized in Figure 1. It is worth mentioning that: (a) The prediction accuracy of STIMA is higher than that of linear models in all cases (see panel (a)) and STIMA generally results in a more parsimonious model. Its cross-validated goodness of fit particularly improves when LM and Step LM require a large number of parameters, and the models are probably overfitting the data. This is the case for the Baseball, Boston, Eurostoxx, and Juice datasets. (b) In the comparison of STIMA with Step GAM and Step GAM_int (panel (b)), STIMA is the best performer in all but two cases: Abalone and Eurostoxx. For these datasets, Step GAM is able to find important nonlinear main effects that improve the fit of the final model compared to STIMA. In addition, STIMA appears to be more sensitive to a large outlier in the Eurostoxx data than GAM. (c) STIMA always performs better than RPART (panel (c)). The default pruning rule of RPART appears to be rather restrictive, leading to small, highly inaccurate trees. When we increase the size of the tree by setting the complexity parameter to 0.001, the goodness of fit improves a bit, but the model

[2]This R-package is available online as supplementary material.

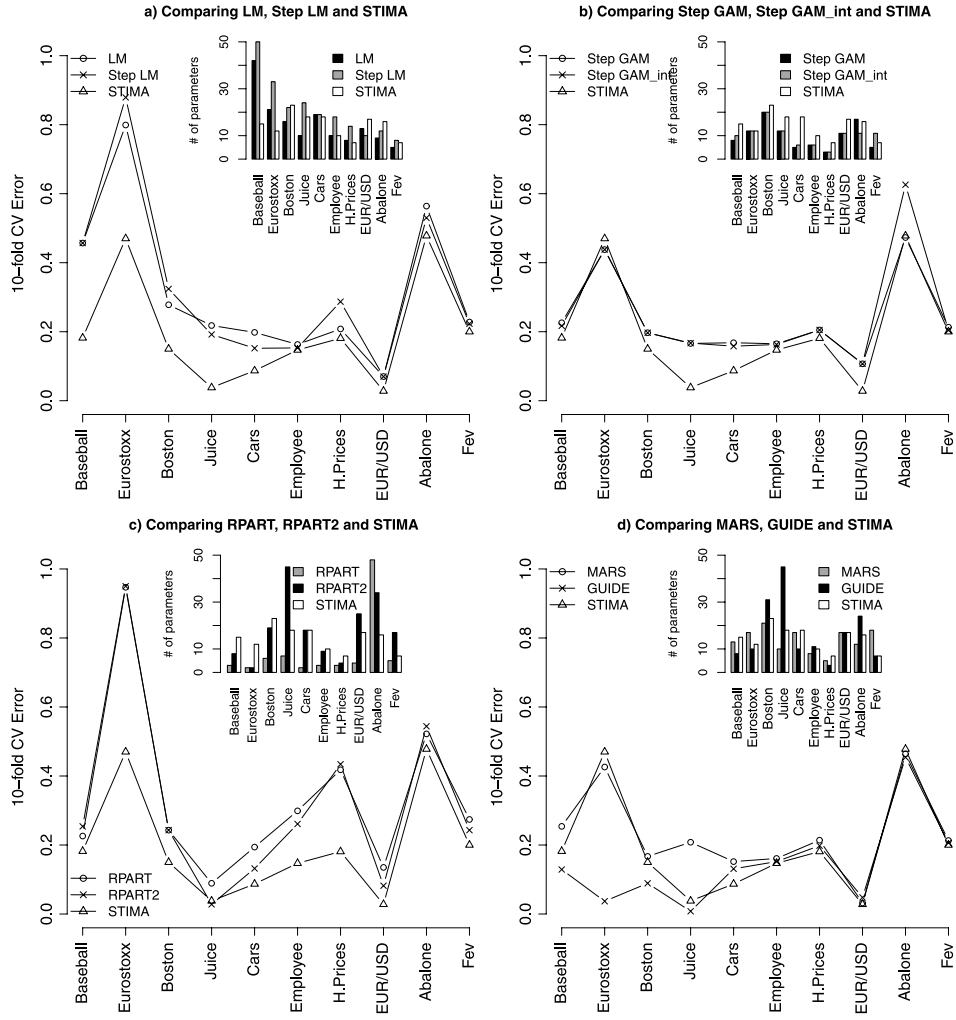

Figure 1. The performance of the compared models: for each panel, a graph of the 10-fold cross-validated error estimated by the different methods is displayed for each dataset. Each line represents a method. The barplot shown at the top of each graph provides information about the number of parameters required by each method for each dataset. In panel (c), RPART refers to the regression tree obtained by using the default options of the `rpart` package, whereas RPART2 refers to the same model obtained by decreasing the complexity parameter to 0.001.

generally overfits the data (see results of RPART2 in Figure 1). STIMA appears to provide a good balance between accuracy and parsimony. For the Abalone data, the two approaches present more or less the same accuracy. (d) The comparison of STIMA with MARS and GUIDE (panel (d)) shows that the latter performs remarkably well in the Eurostoxx example, whereas for the other datasets STIMA and GUIDE clearly outperform MARS most of the time. STIMA appears to be more sensitive to the outlier in the Eurostoxx data than GUIDE. Contrary to STIMA, GUIDE requires many additional parameters to increase its accuracy for some datasets (Boston, Juice, and Abalone).

To summarize the overall results of the comparative analysis, the methods were ranked with respect to the above-mentioned measures and the average rank was computed: STIMA and GUIDE perform better than the other methods. They present an average rank (with respect to the prediction error) of 2.0 and 1.7, respectively. Both methods estimate the best model in five out of ten cases. For all the datasets with a relatively large percentage ($\geq$50%) of categorical predictors (i.e., Cars, Employee, Fev, and Home Prices), STIMA appears to be the best performer. Detailed results are reported in Appendix D (see online supplementary material).

Finally, to investigate the effectiveness of the pruning rule arrived at in the previous section, we plotted the RE^{cv} with respect to different sizes of the regression trunk. For brevity's sake, Figure 2 only shows the results of four out of the ten datasets. The results suggest that the decrease in the RE^{cv} is less smooth in examples involving either a large number of observations (Abalone) or a large number of predictors (Eurostoxx). In such

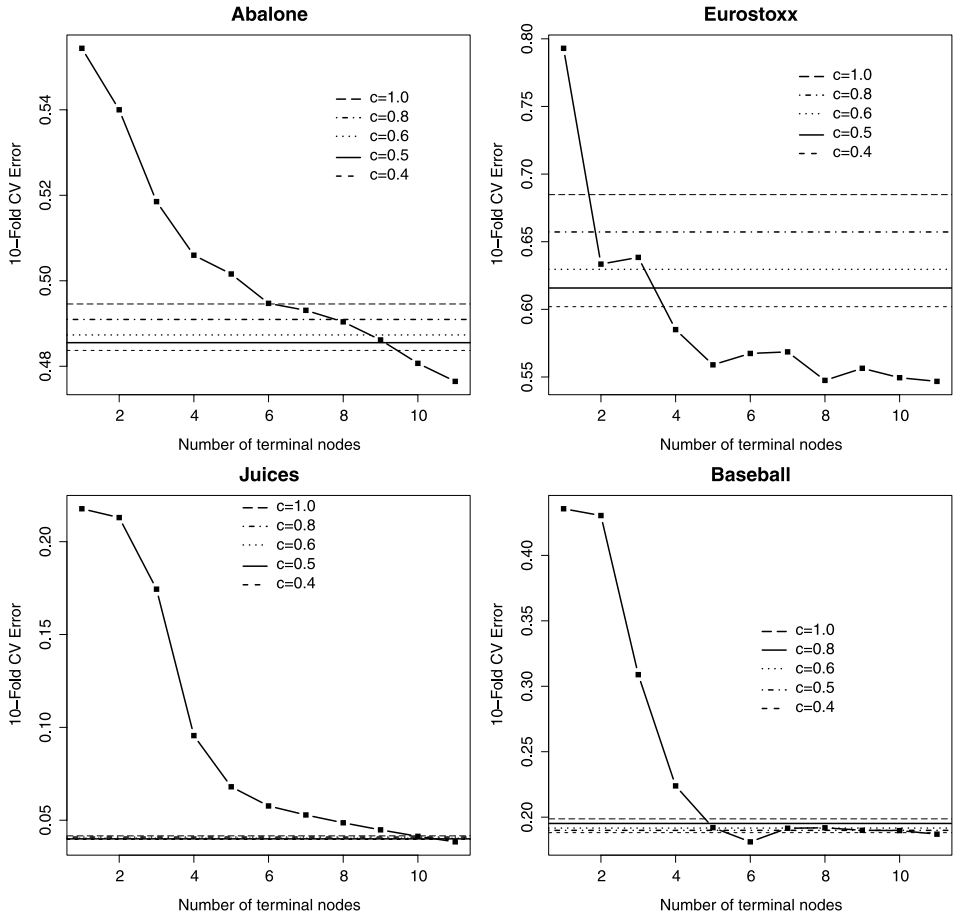

Figure 2. Investigating the effectiveness of the pruning rule: each plot shows the 10-fold cross-validated error with respect to the size of the regression trunk (expressed in terms of number of terminal nodes) for the Abalone, Eurostoxx, Juices, and Baseball datasets. Each horizontal line provides information about the final model obtained by specifying different values of c in the c SE pruning rule.

cases, an appropriate value of c must be specified to avoid overfitting of the data. In other situations (e.g., Juices and Baseball) the RE^{cv}'s obtained through the application of different pruning rules are very close to each other. In other words, the specification of c is less important for these datasets. Furthermore, the final RE^{cv} is much lower (between 0.01 and 0.20) for these datasets than for the Abalone and Eurostoxx datasets (between 0.48 and 0.69).

4.2 A CLOSER LOOK AT THE BOSTON HOUSE PRICE DATASET

We describe the results of the analysis of the Boston house price dataset in order to highlight the main features of the regression trunk model, paying particular attention to the interpretation of the interaction terms. The dataset is composed of 14 continuous predictors and 1 binary predictor observed with respect to median values of houses situated in different census tracts of Boston. The regression trunk model has been built in order to estimate the Corrected Median value of owner-occupied houses (*CMedv*) measured in 1000s of USD (for a complete overview of the variables, see the R-package `stima`, available online as supplementary material, and enter help(boston)). STIMA automatically selected *Lstat* as first splitting variable. The maximum number of splits L of the regression trunk was fixed at 10. The STIMA algorithm grew regression trunks with 2 up to 11 terminal nodes and estimated the RE^{cv} and the SE^{cv} for each model. The lowest value of RE^{cv} was reached for a regression trunk with ten splits ($RE^{cv} = 0.144$, $SE^{cv} = 0.018$). Applying the 0.50 SE pruning rule, a regression trunk with seven splits ($RE^{cv} = 0.150$, $SE^{cv} = 0.019$) was considered to be the best model. This model was the smallest model with an RE^{cv} lower than 0.153 ($0.144 + 0.50 \times 0.018$).

The equation of the pruned regression trunk model with standardized regression coefficients is

$$\begin{aligned} CMedv = {} & 0.00 - 0.03Chas - 0.06Long + 0.06Lat - 0.06Crim + 0.11Zn \\ & - 0.09Indus + 0.64Nox + 0.13Rm - 0.07Age - 0.13Dis \\ & + 0.64Rad - 0.23Tax - 0.13Ptratio + 0.07B - 0.25Lstat \\ & + 0.69R_2 + 0.83R_3 + 1.13R_4 + 0.65R_5 + 0.64R_6 + 0.84R_7 + 0.30R_8, \end{aligned} \quad (4.1)$$

where R_2 through R_8 are defined as

$$\begin{aligned} R_2 &= I(Lstat \leq 4.65, Rm \leq 7.37), \\ R_3 &= I(Lstat \leq 4.65, Rm > 7.37), \\ R_4 &= I(Lstat > 4.65, Nox \leq 0.50, Rm \leq 6.92), \\ R_5 &= I(Lstat > 4.65, Nox \leq 0.50, Rm > 6.92), \\ R_6 &= I(4.65 \leq Lstat < 9.54, 0.50 \leq Nox < 0.67), \\ R_7 &= I(Lstat > 9.54, 0.50 \leq Nox < 0.67, Rad \leq 16), \\ R_8 &= I(Lstat > 9.54, 0.50 \leq Nox < 0.67, Rad > 16). \end{aligned}$$

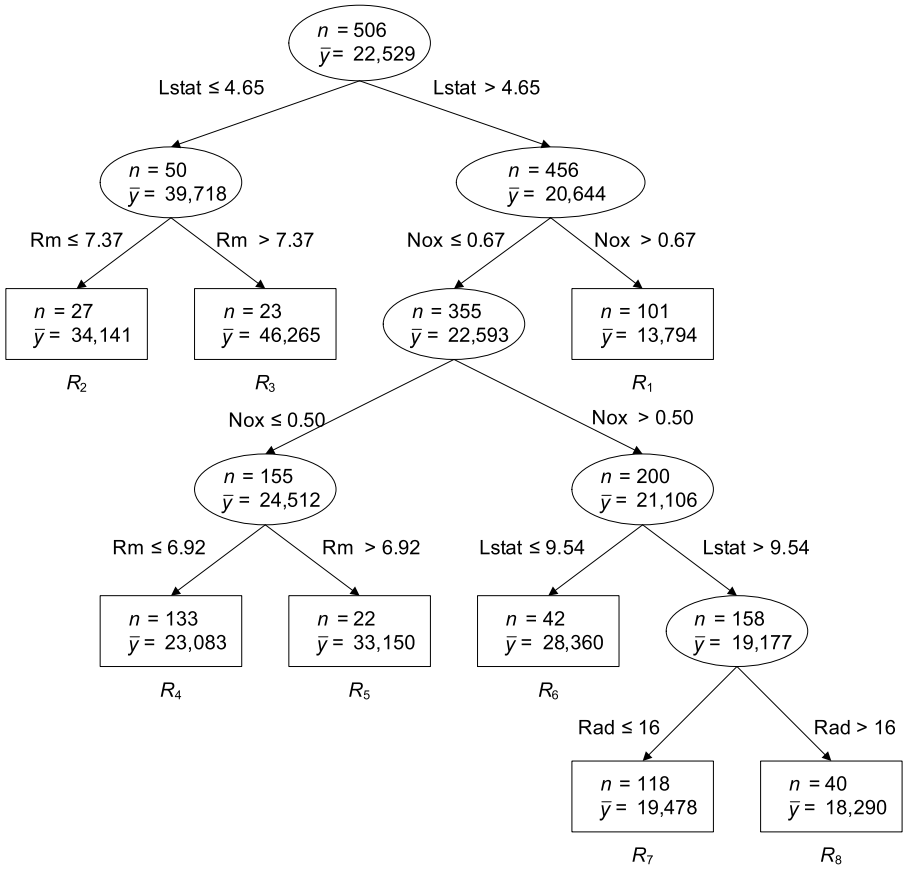

Figure 3. Best size of the regression trunk for the Boston house price data. For each node, the number of subjects is displayed and their average value on the response variable Corrected Median Value of owner-occupied houses.

The term R_1 does not appear in the model equation as it plays the role of reference category. To interpret the pruned regression trunk model, we inspected both the plot of the regression trunk (Figure 3) and the regression coefficients given in (4.1). The first splitting predictor in the regression trunk is *Lstat* and the split point is 4.65 (see Figure 3). It allows us to separate the 50 districts with a higher real estate value from the remaining 456. The group of houses with a higher value is further split according to *Rm*: houses with more than 7.37 rooms present the highest average value, whereas houses with fewer than 7.37 rooms have a medium price. The right branch of the trunk is split further with *Nox*, *Rm*, *Lstat*, and *Rad* as splitting variables.

A comparison of the standardized regression coefficients in (4.1) reveals that the main effects of air pollution (*Nox*), the accessibility to radial highways (*Rad*), and the interaction effects defined by the regression trunk (R_2 through R_7 versus R_1) are very important. Houses in the R_2 through R_7 regions have a significantly higher value than those in R_1, adjusted for the effect of the other predictors. Region R_1 consists of the census tracts with a combination of a relatively high percentage of lower class population (*Lstat* $>$ 4.65) and

a high level of air pollution ($Nox > 0.67$). Our results suggest that the combined influence of a higher percentage of lower class population and a high level of air pollution results in the lowest average house prices.

5. DISCUSSION

The present study proposes a new algorithm to fit a regression trunk model: the Simultaneous Threshold Interaction Modeling Algorithm (STIMA). In STIMA, the main effects and threshold interaction effects are optimized simultaneously. This is the main innovation with respect to the regression trunk approach of Dusseldorp and Meulman (2004), where the main effects estimation and the search for threshold interactions are two separate procedures.

Our simulation results show that the performance of STIMA is satisfactory for sample sizes of 200 or larger. For sample sizes of 200, the 0.80 SE pruning rule can be recommended. This pruning rule ensures that the chance of finding spurious interaction effects, the Type I error, is smaller than 0.05. For sample sizes of 300 and larger, the 0.50 SE pruning rule is the best performer in terms of power (i.e., detection rate is $\geq$0.80) and Type I error. For sample sizes of 100, the power of STIMA is only satisfactory for large-sized two-way categorical by continuous interactions. For comparison, the required sample size for a power of 0.80 to detect a moderate two-way cross-product interaction in multiple regression is 100 to 150 (if the reliability of each predictor is 0.88; Cohen et al. 2003).

In the true models used to generate the population data, we added a large percentage—50%—of irreducible error; that is, 50% of the variance in Y was due to other nonobserved factors. We checked whether the results for sample sizes of 100 would be considerably better for true models that included only 20% of error variance, while keeping the effect sizes equal. This was not the case. A plausible reason for this phenomenon is that the value of the error variance is taken into account in the calculation of the effect size, in the denominator. The effect size for regression is computed as the absolute increase in variance-accounted-for divided by the error variance. Therefore, the absolute increase in variance-accounted-for by the interaction terms in the population models was much lower in the 20% error situation than in the 50% error situation. As a consequence, the interaction terms were nearly as hard to detect by STIMA in the 20% error situation as in the 50% error situation. A limitation of the simulation study is that the reliability of the predictors was assumed to be 1.0 in the population. Future research should determine the influence of the amount of random error of the predictor variables on the power of the regression trunk approach.

The results of the application study are promising. They suggest that among the tested regression methods, both STIMA and GUIDE are worth considering when the researcher's interest lies in assessing interaction effects. These methods resulted in the best model (in terms of estimated prediction error) in five out of ten cases. STIMA performed slightly better than GUIDE in terms of parsimony; most solutions of STIMA used fewer parameters than GUIDE, whereas for GUIDE a more stringent pruning rule was used (the 1 SE rule) than for STIMA (the 0.50 SE rule). STIMA outperformed the other regression methods for

datasets containing interaction effects between both categorical and continuous variables. For datasets with 50% or more categorical predictor variables in particular, STIMA was the best method. The performance of STIMA was worse than some other methods (GUIDE and GAM) for datasets with outliers or nonlinear main effects. The default main effects model of STIMA is linear. Nonlinear main effects can be incorporated, but they need to be specified a priori.

There are three other limitations of STIMA that need to be acknowledged. First, STIMA cannot handle datasets where the number of predictor variables is larger than the sample size. Second, if the first splitting variable is categorical with a large number of levels, the performance of STIMA may suffer. This would be because STIMA will split the root node into as many pieces as the number of levels of a categorical variable. For example, if the first splitting variable is political orientation with as many categories as political parties in a country, the regression trunk will start by separating all parties. In such a situation, it may be better to reduce the number of parties a priori into, for example, left-wing, center, and right-wing. Third, the current implementation of STIMA is limited to modeling a subset of complete observations if there are missing values.

We compared the tested regression methods in terms of parsimony, focusing on the number of estimated parameters in the models. It should be noted that this number does not indicate the number of degrees of freedom for the tree-based regression models. For these models, the approximated number of degrees of freedom is lower than the sample size minus the number of parameters (i.e., terminal nodes), because each terminal node is the result of an optimal split search.

Results of the comparative study show that there is not a single best method for every dataset in terms of prediction error. One might argue, however, whether it is necessary to search for a single best method. In our view, it is very important in practical situations to apply several regression methods to a specific dataset and compare the results in line with the objectives and hypotheses of the prediction problem at hand.

SUPPLEMENTAL MATERIALS

STIMA: R-package `stima` containing code to perform the STIMA analyses described in the article. The package also contains some of the example datasets from the article. The file consists of a Windows binary version of the package. (stima_1.0.1.zip)

Appendices: Appendix A through Appendix D. (AppendixABCD.pdf)

ACKNOWLEDGMENTS

This study was supported by the Netherlands Organization for Scientific Research grant 451-02-058 to the first author and by the research fund awarded to the second author by University of Cagliari within the "Young Researchers Start-Up Programme 2007." The authors thank Jerome H. Friedman and the referees for their helpful and valuable suggestions, which improved the overall quality of the article.

[Received July 2006. Revised April 2010.]

REFERENCES

Alexander, W. P., and Grimshaw, S. D. (1996), "Treed Regression," *Journal of Computational and Graphical Statistics*, 5, 156–175. [516]

Asuncion, A., and Newman, D. J. (2007), "UCI Machine Learning Repository," available at *http://www.ics.uci.edu/~mlearn/MLRepository.html*. [522]

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984), *Classification and Regression Trees*, London: Chapman & Hall. [515,518]

Cohen, J. (1988), *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.), Hillsdale, NJ: Lawrence Erlbaum. [517,520]

Cohen, J., Cohen, P., West, S. G., and Aiken, L. S. (2003), *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences* (3rd ed.), Mahwah, NJ: Lawrence Erlbaum. [515,528]

Dusseldorp, E., and Meulman, J. J. (2004), "The Regression Trunk Approach to Discover Treatment Covariate Interaction," *Psychometrika*, 69, 355–374. [516,517,519,528]

Dusseldorp, E., Spinhoven, P., Bakker, A., Van Dyck, R., and Van Balkom, A. J. L. M. (2007), "Which Panic Disorder Patients Benefit From Which Treatment: Cognitive Therapy or Antidepressants?" *Psychotherapy and Psychosomatics*, 76, 154–161. [517]

Foster, D. P., Stine, R. A., and Waterman, R. P. (1998), *Business Analysis Using Regression: A Casebook*, New York: Springer. [522]

Friedman, J. H. (1991), "Multivariate Adaptive Regression Splines" (with discussion), *The Annals of Statistics*, 19, 1–141. [516]

Harrel, F. E., Jr. (2001), *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*, New York: Springer. [515]

Hastie, T. J. (2009), "GAM: Generalized Additive Models," R package version 1.0.1, available at *http://CRAN.R-project.org/package=gam*. [523]

Hastie, T. J., and Tibshirani, R. J. (1990), *Generalized Additive Models*, London: Chapman & Hall. [515]

Hastie, T. J., Tibshirani, R. J., and Friedman, J. H. (2001), *The Elements of Statistical Learning*, New York: Springer. [515]

Leisch, F., Hornik, K., and Ripley, B. D. (2009), "MDA: Mixture and Flexible Discriminant Analysis," R package version 0.4-1, available at *http://CRAN.R-project.org/package=mda*. [523]

Loh, W. Y. (2002), "Regression Trees With Unbiased Variable Selection and Interaction Detection," *Statistica Sinica*, 12, 361–386. [516,523]

Morgan, J. N., and Sonquist, J. A. (1963), "Problems in the Analysis of Survey Data, and a Proposal," *Journal of the American Statistical Association*, 58, 415–434. [515]

Quinlan, J. R. (1992), "Learning With Continuous Classes," in *AI'92: Proceedings of the 5th Australian Joint Conference on Artificial Intelligence*, eds. A. Adams and L. Sterling, Singapore: World Scientific, pp. 343–348. [516]

R Development Core Team (2009), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing. Available at *http://www.R-project.org*. [519]

Therneau, T. M., and Atkinson, E. J. (1997), "An Introduction to Recursive Partioning Using the rpart Routines," available at *http://mayoresearch.mayo.edu/mayo/research/biostat/upload/61.pdf*. [523]