Mach Learn (2009) 74: 111–131
DOI 10.1007/s10994-008-5080-8

Tree-structured model diagnostics for linear regression

Xiaogang Su · Chih-Ling Tsai · Morgan C. Wang

Received: 18 October 2005 / Revised: 25 June 2008 / Accepted: 24 July 2008 /
Published online: 27 August 2008
Springer Science+Business Media, LLC 2008

Abstract This paper studies model diagnostics for linear regression models. We propose two tree-based procedures to check the adequacy of linear functional form and the appropriateness of homoscedasticity, respectively. The proposed tree methods not only facilitate a natural assessment of the linear model, but also automatically provide clues for amending deficiencies. We explore and illustrate their uses via both Monte Carlo studies and real data examples.

Keywords AIC · CART · Heteroscedasticity · Linear models · Regression trees

1 Introduction

The linear regression model has been widely used in data analysis. Its popularity can be attributed to its simple form, sound theoretical support, fast computation in estimation, great flexibility in incorporating interactions, dummy variables, and transformations, and easy interpretation. Suppose that n independent observations $\{(y_i, \mathbf{x}_i^0),\ i = 1, \ldots, n\}$ were generated from the true model

$$y_i = f(\mathbf{x}_i^0) + \varepsilon_i.$$

The response y_i is continuous, the predictor vector $\mathbf{x}_i^0 = (x_{i1}^0, \ldots, x_{ip}^0)$ is a mixture of continuous and discrete variables, and ε_i is random error. With slight abuse of notation, we use ε_i to denote the random error across all models in the rest of paper.

Editor: David Page.

X. Su (✉) · M.C. Wang
Department of Statistics and Actuarial Science, University of Central Florida, Orlando, FL 32816, USA
e-mail: xiaosu@mail.ucf.edu

M.C. Wang
e-mail: cwang@mail.ucf.edu

C.-L. Tsai
Graduate School of Management, University of California, Davis, CA 95616, USA
e-mail: cltsai@ucdavis.edu

Suppose that the 'best' or 'near to the best' approximating linear model is

$$y_i = \mathbf{x}_i \boldsymbol{\beta} + \varepsilon_i, \tag{1}$$

where $\boldsymbol{\beta} \in \mathcal{R}^q$ is the regression parameter vector, and the row vector $\mathbf{x}_i = (x_{i1}, \ldots, x_{iq})$ contains the dummy variables, cross-product interactions, and transformations obtained from $\mathbf{x}_i^0$. Taking into account the intercept term, we assume $x_{i1} = 1$.

There are four major assumptions involved in model (1), which can be summarized as linearity, independence, normality, and homoscedasticity. Among these, the plausibility of independence can be inspected during the data collection stage and the normality assumption is usually of somewhat lesser concern for large data due to the robustness of linear regression. However, the linear specification and homoscedasticity are closely related to the bias and variance of the final estimation. The violation of either of these two assumptions may result in misleading inferences and interpretations.

In regression analysis, diagnostic plots and significance tests have become routine procedures for assessing the appropriateness of model assumptions. Diagnostic plots provide important and useful graphical evaluations of overall model plausibility and outlier identification. However, one often cannot make a decisive conclusion merely through graphical inspections. In addition, graphical plots may result in a solid black display for large data sets. Therefore, great care should be exercised when applying diagnostic plots. To complement graphical approaches, statistical significance tests are used to detect the violations of model assumptions. For example, one commonly used method is to first group the residuals and then construct chi-square type test statistics (see, e.g., Neter et al. 1996). However, the resulting conclusions are sensitive to the number of groups as well as the approach used for forming the groups. Furthermore, the testing methods usually provide little information regarding amendment of model deficiencies. Moreover, many concepts in significance testing become less meaningful in the analysis of large data sets, which make them less attractive to data miners (see Hand 1999).

To enhance diagnostic methods, we consider the tree method or recursive partitioning originated by Morgan and Sonquist (1963). By recursively bisecting the predictor space, a tree method provides a piecewise constant approximation to a targeted function. To address tree size selection as well as many practical issues in tree construction, Breiman et al. (1984) proposed the classification and regression trees (CART) algorithm. Since its inception, CART has greatly advanced the use of tree models in various fields. Their pruning idea has become the current standard approach for developing optimally-sized trees.

In this paper, we propose two tree-based procedures to assess the linearity and homoscedasticity assumptions of linear models, respectively. We start with the 'best' linear regression model (1) and then construct diagnostic trees that center around it. If the nontrivial final tree structure (i.e., containing more than one terminal node) is developed, then the adequacy of a linear model is questionable. The diagnostic tree methods not only help to check the appropriateness of linear regression, but also suggest useful clues for amending model deficiencies.

Both proposed tree methods follow the CART algorithm, which consists of three major steps: first growing a large initial tree via greedy search, then truncating it back to obtain a nested sequence of subtrees, and finally selecting the optimal tree size. The rest of this paper is organized as follows. In Sect. 2, we present a tree procedure to assess the adequacy of linearity. The basic idea is to augment linear regression with a tree structure. The proposed procedure adaptively picks up the binary split that furnishes the best augmentation to the linear model (1). It is of interest to note that Miller's (1996) employed the residual-based

CART procedure to examine model adequacy. In Monte Carlo studies, we demonstrate that our approach is often superior (or comparable) to Miller's method. In Sect. 3, we propose a tree procedure for assessing homoscedasticity. Simulated experiments are also presented to investigate its performance in detecting homoscedasticity and to make comparison with the treed variance (TV) procedure developed by (Su et al. 2006). Section 4 provides two empirical examples, the 1987 baseball salary data and the Boston housing data, to illustrate the proposed methods. Section 5 concludes the article with a brief discussion.

2 Checking adequacy of model specification

2.1 Model structure

To check whether or not the 'best' linear model, $\mathbf{x}\boldsymbol{\beta}$, provides an adequate approximation to the true regression function $f(\mathbf{x}^0)$, we consider the following hybrid model in which an augmentation tree structure T is attached to the 'best' linear model (1):

$$y_i = \mathbf{x}_i\boldsymbol{\beta} + \mathbf{z}_i^{(T)}\boldsymbol{\gamma} + \varepsilon_i, \tag{2}$$

where $\mathbf{z}_i^{(T)}$ contains dummy variables induced from a tree structure T and $\boldsymbol{\gamma}$ is the corresponding parameter vector. Given a tree structure T, let $\widetilde{T}$ and $|\widetilde{T}|$ denote the set of all terminal nodes and the size of T (i.e., the total number of terminal nodes of T), respectively. Then $\mathbf{z}_i^{(T)} = (z_{i1}, \dots, z_{i|\widetilde{T}|})$ is of dimension $1 \times |\widetilde{T}|$, where

$$z_{ik} = \begin{cases} 1 & \text{if the } i\text{-th observation falls into the } k\text{-th terminal node of } T, \\ 0 & \text{otherwise,} \end{cases}$$

for $i = 1, \dots, n$, $k = 1, \dots, |\widetilde{T}|$, and $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_{|\widetilde{T}|})'$ subject to $\sum \gamma_k = 0$ for identification. Heuristically speaking, tree T provides an augmentation to linear regression by making a piecewise constant approximation to $f(\mathbf{x}_i^0) - \mathbf{x}_i\boldsymbol{\beta}$ (see the consistency of recursive partitioning in Breiman et al. 1984).

It is worth noting that the linear regression and tree structure complement each other: the linear model captures global patterns, while the tree structure excels in detecting local properties, such as thresholds, nonlinear patterns, and complex interactions. This renders the tree structure an excellent augmentation tool for linear regression. The tree T is expected to pick up the signals overlooked by the linear model (1), and the splitting variables shown in T can be regarded as those effects that have been under-represented by the linear model. Based on the hybrid model (2), we employ the tree approach to check the adequacy of the 'best' linear model. If a nontrivial tree structure T with $|\widetilde{T}| > 1$ can be developed, then it signifies the lack-of-fit of model (1). Otherwise, the linear model provides a reasonable fit. Furthermore, the final tree structure provides useful diagnostic information for amendment.

To construct an optimal augmentation tree structure T in model (2), an intuitive approach, as considered by Miller's (1996), is to run a tree procedure such as CART (Breiman et al. 1984) directly on the residuals from fitting model (1). That is, the augmentation tree is constructed by treating the residuals $\{r_i = y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}}_0 : i = 1, \dots, n\}$ as responses and the components in $\mathbf{x}^0$ as inputs, where $\hat{\boldsymbol{\beta}}_0$ is the least squares estimate of $\boldsymbol{\beta}$ in model (1).

In contrast to the above residual-based approach, we propose a tree method that attains iterative adjustments to model (1). To this end, we select the best split by minimizing the sum of squared errors (SSE) associated with a threshold model. Then, we adopt the pruning idea of CART (Breiman et al. 1984) to determine the best tree size. We next present the detailed procedure.

2.2 Tree procedures

2.2.1 *Growing a large tree*

To split node h, we fit the following threshold model using data in node h:

$$y_i = \mathbf{x}_i\boldsymbol{\beta} + \gamma \cdot I\left\{x_{ij} \leq c\right\} + \varepsilon_i, \tag{3}$$

where the indicator function $I\{x_{ij} \leq c\}$ corresponds to a binary split, say, s, of the data according to a continuous predictor X_j. Here, X_j is used as a generic notation for the j-th predictor, where $j = 1, \ldots, q$. If X_j is discrete with values in $C = \{c_1, \ldots, c_d\}$, then the form of $I\{x_{ij} \in A\}$ is considered for any subset $A \subset C$.

Given a split s, we compute the sum of squared errors (SSE) for model (3). The best split $s^\star$, among all permissible splits, is the one associated with the smallest SSE. The data are then partitioned into two child nodes according to the best split $s^\star$. Subsequently, the same procedure is applied to partition both child nodes. Recursively doing so yields a large initial tree T_0.

Because the splitting procedure requires evaluating model (3) for every allowable split at each node, the Algorithm 1 based on QR decomposition of the design matrix $\mathbf{X}$ facilitates fast computation of the SSE of model (3) with minor updating from the results of model (1). As can be seen, the QR-decomposition, which is the most time-consuming part in the least squares fitting, is performed only once when searching over all permissible splits in node h. More details of this algorithm are given in Appendix A.

Algorithm 1 Pseudo-code for finding the best split in node h.

```
Let $\mathbf{X}$ denote the matrix with elements $x_{ij}$, for $i$ in node $h$.
Perform QR-decomposition on $\mathbf{X}$. Obtain $\mathbf{u} = Q\mathbf{y}$.
Iterate over all permissible splits $s$.
  – Define $\mathbf{x}_s = \mathbf{1}_{\{x_j \leq c\}}$ and compute $Q\mathbf{x}_s$
  – Obtain the Householder matrix $H$ (see Appendix A for details).
  – Compute $\mathbf{u}^{(s)} = H\mathbf{u}$ and $\mathrm{SSE}^{(s)}$.
Obtain $s^\star$ that minimizes $\mathrm{SSE}^{(s)}$.
```

2.2.2 *Pruning*

The pruning idea in CART (Breiman et al. 1984) is to narrow down the choices of subtrees from which the best-sized tree will be selected. Su et al. (2004) revisited CART within the maximum likelihood framework. They modified some major steps in tree construction for sound statistical justification and easy generalization. In particular, they proposed the AIC (Akaike 1973) pruning algorithm, which employs the widely used Akaike information criterion to assess tree performance. We adopt their procedure to prune augmentation trees.

For model (2) with a given augmentation tree T, the corresponding AIC is given by, up to a constant,

$$\mathrm{AIC}^{(T)} \propto n \cdot \log\left\{\mathrm{SSE}^{(T)}\right\} + 2(q + |\widetilde{T}|),$$

where $\mathrm{SSE}^{(T)} = \sum\{y_i - (\mathbf{x}_i\hat{\boldsymbol{\beta}} + \mathbf{z}_i^{(T)}\hat{\boldsymbol{\gamma}})\}^2$ is the sum of squared errors and $(q + |\widetilde{T}|)$ is the total number of parameters associated with model (2). The smaller $\mathrm{AIC}^{(T)}$, the more favorable tree T.

The pruning algorithm starts with the large initial tree T_0. For any link (i.e., internal node) $h \in T_0$, let T_h denote the branch stemming from h. To measure the quality of h, we consider its complementary subtree $T_0 - T_h$, i.e., the subtree after pruning the branch T_h. Let $\mathrm{AIC}^{(T_0 - T_h)}$ denote the AIC measure from fitting model (2) with $T_0 - T_h$ being the augmentation tree. Then the *weakest link*, $h^\star$, is the internal node that corresponds to the smallest $\mathrm{AIC}^{(T_0 - T_h)}$, i.e.,

$$\mathrm{AIC}^{(T_0 - T_{h^\star})} = \min_{h \in T_0 - \widetilde{T}_0} \mathrm{AIC}^{(T_0 - T_h)}.$$

Here, the link $h^\star$ is the weakest on the ground that its complementary subtree $T_0 - T_{h^\star}$ has the best AIC performance. Now we truncate the weakest link $h^\star$ to obtain the subtree $T_1 = T_0 - T_{h^\star}$. Subsequently, the same pruning procedure is applied to prune T_1. Repeating the above process arrives at a decreasing sequence of subtrees $T_M \prec \cdots \prec T_1 \prec T_0$, where T_M is the null tree with root node only and the notation $\prec$ means "is a subtree of".

2.2.3 Tree size selection

In this stage, a best-sized augmentation tree will be selected from the nested subtree sequence. It is convenient to use the same AIC measure to aid in tree selection. However, due to the adaptive nature of recursive partitioning, a validation method is required to obtain an honest estimate of $\mathrm{SSE}^{(T_m)}$. Since tree-based methods are not very often recommended for small samples, we assume that the available sample size is large enough so that the following test sample method can be used to validate $\mathrm{SSE}^{(T_m)}$.

To determine the tree size, we divide the whole data $\mathcal{L}$ randomly into two groups, the learning (or training) sample $\mathcal{L}_1$, and the test (or validation) sample $\mathcal{L}_2$, with the ratio of the sample sizes n_1/n_2 being approximately 2 : 1. A large tree is then grown and pruned using the learning sample $\mathcal{L}_1$ such that a decreasing sequence of subtrees $\{T_m : 0 \le m \le M\}$ is available. The test sample then is sent down to each subtree. The best-sized augmentation tree $T^\star$ is the subtree that provides the minimum AIC. That is,

$$\begin{aligned}\mathrm{AIC}^{(T^\star)} &= \min_{\{T_m :\, 0 \le m \le M\}} \mathrm{AIC}^{(T_m)} \\ &\propto \min_{\{T_m :\, 0 \le m \le M\}} n \cdot \log \mathrm{SSE}^{(T_m)} + 2\left(q + |\widetilde{T}_m|\right).\end{aligned}$$

Note that $\mathrm{SSE}^{(T_m)}$ given above is computed using data in the test sample $\mathcal{L}_2$,

$$\mathrm{SSE}^{(T_m)} = \sum_{h \in \widetilde{T}_m} \sum_{\{i : (y_i, \mathbf{x}_i) \in \mathcal{L}_2 \cap h\}} (y_i - \hat{y}_i)^2,$$

where $\hat{y}_i = \mathbf{x}_i \hat{\boldsymbol{\beta}} + \hat{\gamma}_h$ is the predicted value of y_i, and $(\hat{\boldsymbol{\beta}}, \hat{\gamma}_h)$ are the least squares estimates of $(\boldsymbol{\beta}, \gamma_h)$ computed from the learning sample $\mathcal{L}_1$. To extend the above tree size selection to small or moderate samples, one can resort to cross-validation or bootstrap resampling techniques to validate $\mathrm{SSE}^{(T_m)}$.

In addition to AIC, one may use Schwarz's (1978) Bayesian information criterion (BIC) to select the tree size. As demonstrated in Su et al. (2004), AIC (or BIC) is a good alternative to the *ad hoc* 1-SE method in CART (Breiman et al. 1984), which is aimed at correcting the overfitting problem often seen with the 0-SE method. In cases with a large sample size and strong signal, BIC tends to perform better than AIC.

2.3 Simulation studies

In this section, we employ Monte Carlo studies to compare the performance of the augmentation tree (AT) and the Miller's (1996) residual-based tree (RT) methods. For the sake of conciseness, we state two common simulation settings in advance, which are also employed in Sect. 3.3. First, all the inputs or predictors are independent and identically generated from a discrete uniform distribution over values $\{1/50, \dots, 50/50\}$. Second, the number of realizations is 500.

2.3.1 Comparing splitting statistics

We first investigate the performance of the splitting statistics employed in these two tree methods. Note that AT uses the smallest SSE associated with model (3), while RT chooses the best split that maximizes the reduction in the deviance of residuals due to a split.

The data were generated from the following two models:

$$\begin{aligned} &\text{Model A} \quad y_i = 2 + 2 \cdot x_{i1} + I\{x_{i2} \le 0.5\} + \varepsilon_i \quad \text{and} \\ &\text{Model B} \quad y_i = 2 + 2 \cdot x_{i1} + I\{x_{i1} \le 0.5\} + \varepsilon_i, \end{aligned}$$

where ε_i are i.i.d. $N(0, 1)$ random variables. Each data set has two covariates X_1 and X_2. Model A differs from Model B in that the linear regression part confounds with the threshold effect through the same covariate X_1 in Model B.

We consider two sample sizes, $n = 50$ and $n = 500$. For each generated data set, the two splitting methods were used to identify the best cutoff point. For the sake of comparison, both methods started with the 'best' model $y = \beta_0 + \beta_1 \cdot x_1 + \varepsilon$. Figure 1 depicts the relative frequencies of selected cut-points. The bar at -0.5 corresponds to the case in which the splitting variable is incorrectly chosen. For example, if X_1 is chosen to split the data in Model A, then it results in a spurious split. Figure 1 shows that both methods provide nearly identical results in the non-confounded case (Model A). However, in the confounded case (Model B), AT provides considerably more accurate splitting variable and cutoff point selections than those of RT. It is not surprising that both methods perform better with larger samples.

2.3.2 Detecting tree structure

Next, we study the effectiveness of the two tree procedures, AT and RT, in identifying the tree structure overlooked by linear regression. The following five models were considered.

$$\begin{aligned} &\text{Model } A' \quad y_i = 2 + 2 \cdot x_{i1} + 2 \cdot x_{i2} + \varepsilon_i, \\ &\text{Model } B' \quad y_i = 2 + 2 \cdot x_{i1} + 2 \cdot x_{i2} + 3 \cdot I\{x_{i1} \le 0.5 \cap x_{i2} \le 0.5\} + \varepsilon_i, \\ &\text{Model } C' \quad y_i = 2 + 2 \cdot x_{i1} + 2 \cdot x_{i2} + 3 \cdot I\{x_{i3} \le 0.5 \cap x_{i4} \le 0.5\} + \varepsilon_i, \\ &\text{Model } D' \quad y_i = 2 + 2 \cdot x_{i1} + 2 \cdot x_{i2} + \sin(3\pi x_{i1}) + \sin(3\pi x_{i2}) + \varepsilon_i, \\ &\text{Model } E' \quad y_i = 2 + 2 \cdot x_{i1} + 2 \cdot x_{i2} + \sin(3\pi x_{i3}) + \sin(3\pi x_{i4}) + \varepsilon_i, \end{aligned}$$

where $\varepsilon_i \sim N(0, 1)$ and $i = 1, \dots, n$. Each model involves four covariates, $X_1, \dots, X_4$. However, not all of them are associated with the response. Model A' is a plain linear

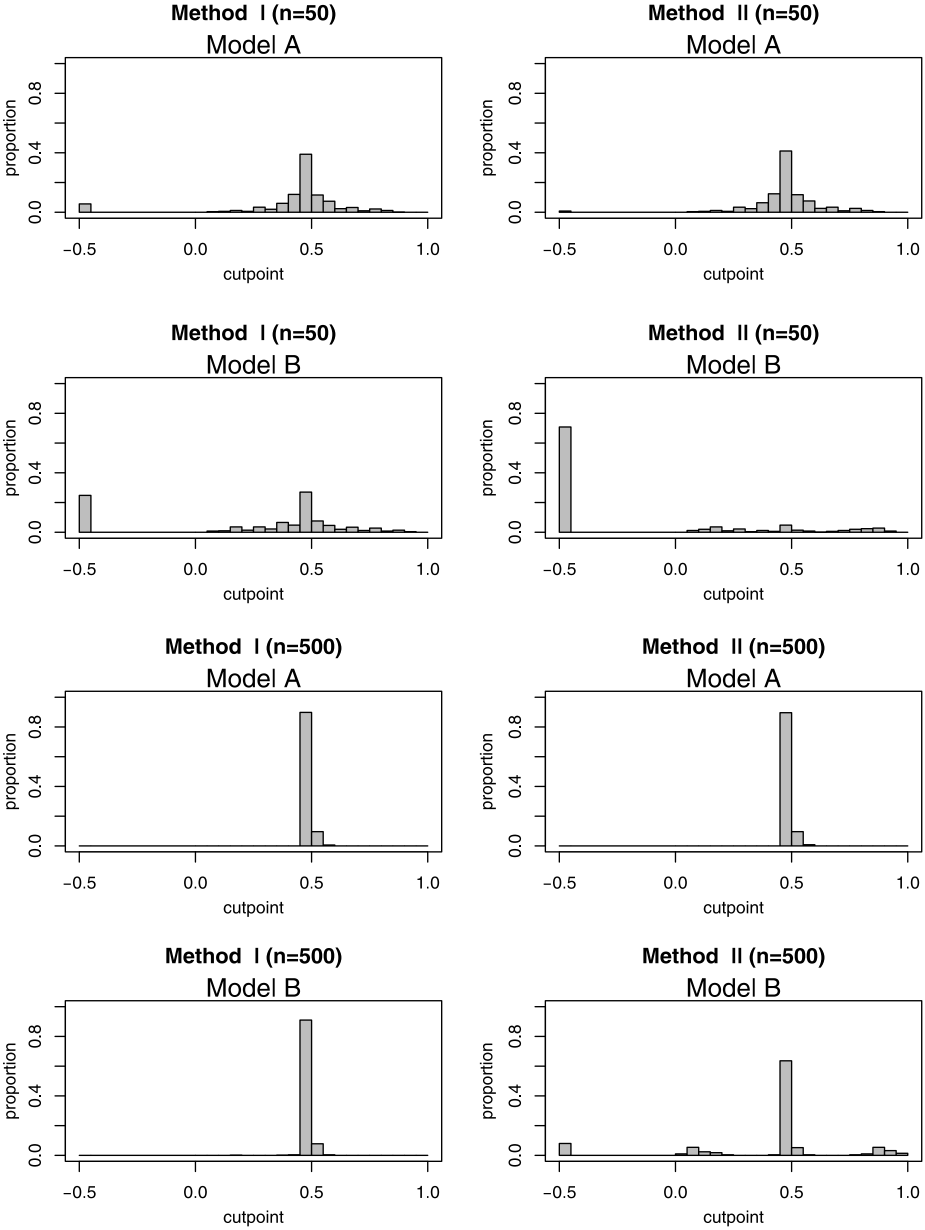


Fig. 1 Comparison of two different splitting methods: Method I, the minimum SSE obtained from fitting model (3); Method II, greedy search using least squares (LS) method based on the residuals of model (1)

model. Models B' and C' involve two interacting thresholds and fall into the general form of model (2). Models D' and E' include two additive nonlinear terms, which are different from the model form in (3). Moreover, Models B' & D' differ from Models C' & E' in that the tree structure of Models B' & D' involves the same variables (X_1 and X_2) that also appear in the linear regression model.

Table 1 Relative frequencies (as a percentage) of the final tree sizes are selected by augmentation tree (AT) and residual-based tree (RT), respectively. The last column is the percentages of the correctly splitting variable selections in the final tree structures

Model	Sample size	Tree method	Number of terminal nodes 1	2	3	4	5	≥ 6	Variable selection
A'	300	AT	**95.4**	3.6	1.0	0.0	0.0	0.0	95.4
		Miller	**98.2**	1.8	0.0	0.0	0.0	0.0	98.2
	1500	AT	**94.6**	5.4	0.0	0.0	0.0	0.0	94.6
		Miller	**98.0**	2.0	0.0	0.0	0.0	0.0	98.0
B'	300	AT	0.0	0.2	**85.4**	12.2	2.2	0.0	94.6
		Miller	15.2	0.6	**4.8**	17.6	23.4	38.4	65.2
	1500	AT	0.0	0.0	**97.0**	2.4	0.6	0.0	98.2
		Miller	0.0	0.0	**0.0**	1.0	0.8	98.2	71.4
C'	300	AT	0.0	0.0	**87.2**	11.6	0.8	0.4	100.0
		Miller	0.0	0.0	**90.8**	7.2	1.8	0.2	85.6
	1500	AT	0.0	0.0	**96.2**	3.8	0.0	0.0	97.4
		Miller	0.0	0.0	**94.8**	4.2	1.0	0.0	95.2
D'	300	AT	1.8	8.0	8.2	31.6	11.6	38.8	72.6
		Miller	6.8	0.2	34.8	2.0	8.0	48.2	53.0
	1500	AT	0.0	0.0	0.0	0.0	1.6	98.4	97.0
		Miller	0.0	0.0	0.0	0.0	0.0	100.0	82.2
E'	300	AT	4.0	6.8	19.8	16.4	20.2	32.8	58.8
		Miller	5.2	0.0	25.0	6.4	19.0	44.4	63.2
	1500	AT	0.0	0.0	0.0	0.0	0.0	100.0	100.0
		Miller	0.0	0.0	0.0	0.0	0.0	100.0	93.8

In this study, we consider two sample sizes, $n = 300$ and $n = 1,500$, and the ratio of the training sample versus the test sample is $2:1$. For each generated sample, both AT and RT start with the 'best' linear model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$. Subsequently, we record the Final tree size (i.e., the number of terminal nodes) selected by each of the two methods (for example, three terminal nodes are desired to represent the tree structures of Models B' & C'). In addition, we expect that the final tree structure extracts useful diagnostic information as well as identifies variables whose effects have been under-represented by the linear regression model. To this end, we adopt Miller's (1996) idea of recording the correctly splitting variable selections in the final tree structure. For example, both X_1 and X_2, but neither X_3 nor X_4, are expected to correctly show up in the final tree structure of Model B'.

Based on above settings, we examine AT's performance. Table 1 (rows 1 and 3) shows that if a linear model is sufficient, then it is rather unlikely for AT to come up with a nontrivial tree structure. It is noteworthy that the frequency for the plain linear Model A' roughly corresponds to the size or 'false positive' rate in hypothesis testing. As for Models $B' - E'$, AT also successfully signals the deficiencies of the linear model by producing nontrivial tree structures. Accordingly, AT performs fairly well in identifying the true model structure and capturing correct variables when the underlying effects could not be adequately captured by the linear model.

We further compare the performance of AT versus RT. Table 1 indicates that AT performs similarly to RT when the tree structure and the linear model contain different sets of variables. However, AT significantly outperforms RT when the tree structure is confounded with the linear model by the same variables. Moreover, the performance of both methods improves in the larger sample size. The above findings are consistent with the results reported in Su and Tsai (2005), who studied Cox proportional hazards models.

Remark 1 To explain the inferiority of method II, we think this is mainly because the initial 'best' linear model is an underfitted one, which leads to biased estimates of $\boldsymbol{\beta}$ (e.g., Chap. 7.9 of Rencher 2000). Such biases may become overwhelming, especially in the confounded cases such as Models B' and D' where predictors that have been left out are highly correlated with those included in the model.

Remark 2 We implement the AT method via R (http://www.r-project.org/), and employ the existing `tree` packages (e.g., `tree`, `prune.tree`, and `cv.tree`) in R (Venables and Ripley 1999) to execute RT. To achieve a better tree size determination, BIC is applied. We also explore the computational efficiency. For the sake of illustration, we consider Model E' with $n = 300$. It took 40 seconds to complete computations of RT on a 1.2 GHz personal computer with 2.00 GB of RAM, whereas it took 3 minutes and 34 seconds to finish the computations of AT. Thus, the computational cost of AT is higher than that of RT. However, we recommend AT for practical use due to its superior performance.

3 Assessing heteroscedasticity

3.1 Model structure

In addition to linearity, another important assumption in model (1) is that of equal variance or homoscedasticity, i.e.,

$$\mathrm{var}(\varepsilon_i) = \sigma_i^2 \equiv \sigma^2.$$

When non-constant variances or heteroscedasticity occurs, the statistical inferences and predictions via the ordinary least squares method are no longer reliable. Therefore, it is crucial to check homoscedasticity. Furthermore, estimating the error variance function itself is often of keen interest in many fields such as economics, finance, engineering, and biological science.

In statistical literatures, heteroscedasticity is commonly modelled as a known function of predictors (see e.g. Rutemiller and Bowers 1968 and Harvey 1976) or the expected response (see e.g., Bickel 1978 and Box 1988). In practice, this specification encounters a serious challenge from the difficulty of effectively and adequately selecting variables from a set consisting of categorical and continuous predictors or their interactions, and the expected response variable. This motivated us to consider a tree procedure to model the error variance and assess homoscedasticity. The proposed model can be formulated as

$$\sigma_i^2 = \sigma^2 \cdot \exp\left(\mathbf{w}_i^{(T_{\mathrm{V}})}\boldsymbol{\theta}\right) \tag{4}$$

for $i = 1, \ldots, n_t$, where $\mathbf{w}_i^{(T_{\mathrm{V}})} = (w_{i1}, \ldots, w_{i|\widetilde{T}_{\mathrm{V}}|})$ is the dummy vector induced by a tree structure T_{V}, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{|\widetilde{T}_{\mathrm{V}}|})'$ is the corresponding parameter vector satisfying $\sum \theta_k = 0$, the subscript "v" in notation T_{V} indicates that the tree structure is built for variance, and the exponential transformation is taken to guarantee the non-negativeness of the variance. Under the formulation of model (4), observations in the same terminal node have a common constant variance. That is, if the i-th observation falls in the terminal node t, then $\sigma_i^2 = \sigma_t^2 = \sigma^2 \cdot \exp\{\theta_t\}$.

3.2 Tree procedure

To construct tree T_{V}, we propose applying a tree procedure such as CART on the squared residuals. Specifically, obtain the residuals r_i's from the ordinary least squares fit of model (1) with equal variances. Then, construct the tree T_{V} by treating the squared residual r_i^2 and the components in $\mathbf{x}_i^0$ as responses and inputs, respectively. To incorporate the situation where the variance depends on the mean response, we suggest splitting the fitted values $\{y_i = \mathbf{x}_i \hat{\boldsymbol{\beta}}_{\mathbf{0}} : i = 1, \ldots, n\}$.

To justify the above procedure, we examine a splitting statistics, which is derived from the score test for heteroscedasticity. Specifically, for a given split s, we consider the following model

$$\sigma_i^2 = \sigma^2 \cdot \exp(\theta w_i), \tag{5}$$

where w_i is a 1-0 binary dummy variable induced by split s and θ is the corresponding parameter. To assess heteroscedasticity, we apply the score test (see Rao 1947 and Cox and Hinkley 1974) for testing $H_0 : \theta = 0$. This test is particularly appealing in the recursive partitioning setting because it is evaluated under the null model. See Appendix B for an outline of derivations under the normality assumption. To make the score test more robust against non-normal error distributions (see Simonoff and Tsai 1994), we adopt Koenker's (1981) studentized approach. After algebraic simplifications, the resulting studentized score test for the given split s (see Appendix B) has the following simple form:

$$\mathrm{ST}(s) = n \cdot \{\mathrm{corr}(\mathbf{w}, \mathbf{q})\}^2, \tag{6}$$

where $\mathbf{w} = (w_1, \ldots, w_{n_t})$ is the dummy vector associated with split s, $\mathbf{q} = (r_1^2, \ldots, r_{n_t}^2)'$, and $\mathrm{corr}(\mathbf{u}, \mathbf{q})$ is the Pearson sample correlation coefficient between $\mathbf{w}$ and $\mathbf{q}$. Among all permissible splits, the best split, $s^\star$, is the one corresponding to the maximum score test statistic, i.e.,

$$\mathrm{ST}(s^\star) = \max_s \mathrm{ST}(s),$$

or, equivalently, the maximum of $\{\mathrm{corr}(\mathbf{w}, \mathbf{q})\}^2$. Recall that, if a simple linear regression model is fit by using r_i^2 as response and w_i as predictor, the resultant coefficient of determination R^2 is equal to the squared correlation coefficient, i.e., $\mathrm{ST}(s) = n_t \cdot \mathrm{corr}^2 = n_t \cdot R^2$. Nevertheless, $R^2 = (\mathrm{SST} - \mathrm{SSE})/\mathrm{SST}$, is a monotone function of deviance or sum of squared errors (SSE) since the total sum of squares (SST) is invariant to splits. Therefore, the best split identified by the maximum score test statistic or the maximum R^2 is the same as the one identified by maximum reduction of deviance in r_i^2, which is exactly the splitting criterion employed in CART (Breiman et al. 1984). The direct implication of this insightful observation is that we can construct the tree T_{V} in model (4) for variance by running a CART type procedure through r_i^2.

To assess the homoscedasticity assumption, we inspect whether a nontrivially final variance tree structure can be obtained or not. If a nontrivial tree structure is indeed developed, then the validity of the equal variance assumption is doubtful. Moreover, the resulting tree structure may yield useful clues on modeling heteroscedasticity. In contrast, if the final tree structure contains the root node only, then one may tentatively conclude that variance is constant.

Supposing that a nontrivial variance tree structure has been identified, we pool the learning sample with the test sample together and then fit the whole data set with model (1) together with the variance function (4), via an iteratively weighted least squares (see, e.g., Carroll and Ruppert's 1988) approach. In this tree-based modeling, the error variance is approximated by piecewise constant functions. Its main advantages include: it naturally incorporates modeling based on the mean response and addresses the variable selection issue in an automatic manner; it has the capability to handle different types of predictors in variance modeling; and the hierarchical variance tree structure carries meaningful interpretation.

To model heteroscedasticity, Su et al. (2006) proposed the treed variance (TV) procedure, which also employed the score test as the splitting statistic. However, their procedure requires an adaptive adjustment for each internal node. In addition, they employed the log-likelihood score associated with models (1) and (4) to aid in tree size determination. Accordingly, the difference between TV and our proposed residual-based tree variance (RTV) approach is analogous to the difference between AT and RT in Sect. 2. Based on Monte Carlo studies in the next subsection, TV and RTV have rather comparable performance.

3.3 Simulations

We conduct simulated experiments to evaluate the performance of RTV in identifying the true variance structure and make comparisons with TV. The data were generated from the model

$$y_i = \mu_i + \varepsilon_i = 2 + 2x_{i1} + 2x_{i2} + \varepsilon_i,$$

where ε_i's are independent and identically normal variables with mean zero and variance σ_i^2. As given in Sect. 2.3.2, we consider two sample sizes, $n = 300$ and $n = 1,500$, and the ratio of the training sample versus the test sample is $2:1$. For the sake of illustration, we present seven structures for error variance σ_i^2 listed below, where the scale parameter σ^2 is implicitly assumed to be 1

$$\begin{array}{ll}
\text{Model } A'' & \sigma_i^2 = 1, \\
\text{Model } B'' & \sigma_i^2 = \exp\{3 \cdot I(x_{i1} \le 0.5 \cap x_{i2} \le 0.5)\}, \\
\text{Model } C'' & \sigma_i^2 = \exp\{3 \cdot I(x_{i3} \le 0.5 \cap x_{i4} \le 0.5)\}, \\
\text{Model } D'' & \sigma_i^2 = \exp\{I(x_{i1} \le 0.5 \cap x_{i2} \le 0.5)\}, \\
\text{Model } E'' & \sigma_i^2 = \exp\{I(x_{i3} \le 0.5 \cap x_{i4} \le 0.5)\}, \\
\text{Model } F'' & \sigma_i^2 = \exp\{8 \cdot x_{i1}x_{i2}\}, \\
\text{Model } G'' & \sigma_i^2 = \exp\{2I(\mu_i > 2.5) + I(\mu_i > 5.0)\}.
\end{array}$$

Each of above models involves four covariates, $X_1, \ldots, X_4$, but not all of them are relevant. Model A'' is a null model with constant variance. Models B''–E'' involve interacting thresholds with different signal strength. In Models B'' & D'', the variables in the variance are the same as those in the mean regression, while in Models C'' & E'' they are different. Model F'' includes an interaction of the original covariates, and Model G'' is only a function of the regression mean. The error variance in Models B'' to E'' and G'' can be fully represented by a tree structure that has three terminal nodes. However, Model F'' needs a large tree to represent the structure. For the sake of convenience, the mean responses in the simulation studies are $E(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$.

Table 2 reports the relative frequencies of final tree sizes and correct selections of splitting variables obtained via TV and RTV, respectively. Note that the fourth column of Table 2 presents the percentage of null final trees found for each model. Results with the null model

Table 2 The relative frequencies of final tree sizes are selected by residual-based tree variance (RTV) and treed variance (TV; Su et al. 2006), respectively. The last column is the percentages of the correctly splitting variable selections in the final tree structures

Model	Sample	Tree	Final tree size						Variable
	size	method	1	2	3	4	5	≥ 6	selection
A''	300	RTV	**96.2**	2.8	0.4	0.4	0.2	0.0	96.2
		TV	**97.4**	1.6	0.6	0.2	0.2	0.0	97.4
	1500	RTV	**97.0**	1.2	1.4	0.4	0.0	0.0	97.0
		TV	**98.8**	1.2	0.0	0.0	0.0	0.0	98.8
B''	300	RTV	26.2	2.8	**45.4**	21.8	3.8	0.0	46.8
		TV	0.0	3.2	**49.2**	24.0	12.4	11.2	80.6
	1500	RTV	0.0	0.0	**91.2**	1.8	4.2	2.8	93.4
		TV	0.0	0.0	**71.4**	22.4	2.2	4.0	98.8
C''	300	RTV	14.8	0.8	**51.4**	29.0	3.8	0.2	56.0
		TV	0.0	1.2	**55.6**	27.2	11.4	4.6	87.0
	1500	RTV	0.0	0.0	**88.6**	6.4	1.2	3.8	91.0
		TV	0.0	0.0	**69.2**	23.0	5.6	2.2	99.0
D''	300	RTV	59.0	7.6	**29.2**	3.2	1.0	0.0	31.2
		TV	55.2	15.4	**23.8**	3.0	2.0	0.6	24.6
	1500	RTV	3.2	0.2	**88.0**	3.4	4.2	1.0	90.0
		TV	0.0	0.0	**92.8**	3.8	2.4	1.0	97.0
E''	300	RTV	53.4	8.8	**26.4**	10.2	0.4	0.8	31.6
		TV	58.6	12.8	**22.6**	3.0	1.4	1.6	24.6
	1500	RTV	2.0	0.4	**88.4**	4.0	4.2	1.0	90.2
		TV	0.0	1.2	**96.0**	1.8	0.8	0.2	97.0
F''	300	RTV	42.2	22.2	19.0	16.6	0.0	0.0	21.8
		TV	0.8	0.0	2.2	4.2	20.8	72.0	91.4
	1500	RTV	10.4	1.0	8.6	8.0	19.2	52.8	55.8
		TV	0.0	0.0	0.0	0.0	0.0	100.0	98.4
G''	300	RTV	17.4	52.2	**22.8**	6.6	1.0	0.0	76.2
		TV	0.0	6.0	**47.6**	24.8	16.0	5.6	94.8
	1500	RTV	0.0	34.4	**58.2**	5.0	2.0	0.4	93.2
		TV	0.0	0.0	**66.8**	20.0	10.4	2.8	97.0

A'' suggests that the 'false positive' rates for both methods are rather low. For Models B''–G'', it can be seen that both TV and RTV perform reasonably well, while TV outperforms RTV in most of cases for signaling heteroscedasticity.

In identifying the true variance structures and selecting the desired variables, Table 2 indicates that TV and RTV are comparable. It is not surprising that the performance of both methods improves in the larger sample size. Although TV often outperforms RTV, RTV yields considerably more correct tree size selections in Models B'' & C'' when the signal is strong and the sample size is large. It is worth noting that RTV is computationally faster and much easier to implement. For example, it took 47 seconds for RTV to complete 500 runs for Model B'' while TV spent 12 minutes and 27 seconds.

4 Data examples

To illustrate, we consider two well-known data sets: one is the 1987 baseball salary data and the other is the Boston housing data. Both data sets, as well as variable description and other related information, are available from StatLib (http://lib.stat.cmu.edu). For conciseness, we will describe them very briefly and refer readers to StatLib for details.

4.1 1987 baseball salary data

The 1987 baseball salary data set originally comes from the 1988 American Statistical Association (ASA) graphics poster session. After deleting observations with missing values, the remaining data set contains the salary information for 263 major league hitters. The response variable is the log transformation of salary and there are 22 input variables, which are performance measures for each hitter.

This baseball data set has been widely studied in the literature. Hoaglin and Velleman (1995) (HV), provided a nice overview on analysis results using various statistical methods. They found that the following model structure yields good model fit and leads to sensible interpretations:

$$\begin{aligned}\log(\texttt{salary}) = \beta_0 + \beta_1 \frac{\texttt{runcr}}{\texttt{yrs}} + \beta_2\sqrt{\texttt{run86}} \\ + \beta_3 \min[(\texttt{yrs} - 2)_+, 5] + \beta_4(\texttt{yrs} - 7)_+ + \varepsilon. \qquad (7)\end{aligned}$$

The segmentations on year are roughly based on a player's eligibility for arbitration and free agency, respectively. HV's model fitting (labelled as Model I) is given in Table 3. We next apply the proposed tree diagnostic methods to assess the adequacy of Model I.

Figure 2a shows the final augmentation tree structure obtained by AT, which has three terminal nodes. The first split of the data is according to hitcr $\leq$ 450 (i.e., whether the total number of career hits for a player is no more than 450). For those players with hitcr $>$ 450 (i.e., observations in node 12), their salaries are further differentiated by puto86 $\leq$ 570. These two threshold effects can also be roughly visualized via the partial residual plots (see, e.g., Mansfield and Conerly 1987) in Fig. 3. Specifically, Figs. 3a and 3b depict the partial residuals associated with HV's model versus the values of hitcr and puto86, respectively. Because the threshold effect of puto86 is identified at node 12, it is natural to further examine the partial residual plot at node 12. The resulting Fig. 3c indicates that the threshold effect of puto86 becomes more prominent in comparison to Fig. 3b. In contrast to AT, we also employ Miller's (1996) RT to obtain a final tree structure, which has two terminal nodes split by {puto86 $\leq$ 462.5} (see Fig. 2b). Because Monte Carlo studies show that AT is often superior (or comparable) to RT, we mainly focus on exploring and interpreting the final tree structure found by AT.

The AT procedure suggests that the effects of hitcr and puto86 on salary have been under-represented in HV's model. It is interesting to note that hitcr is a measure of offensive performance of a player in his career, while puto86 is a measure of defensive performance of a player in the preceding year. Therefore, in addition to the average number of runs per year, the total number of runs scored in the previous season, a boost in salary after year two, and an inverse relation to time in the league after year seven, there are additional bonuses to a hitter's salary. Specifically, the players are rewarded with further monetary enhancement to their salaries if they are able to combine offensive and defensive baseball expertise. To take into account tree effects, we add two dummy variables via the

Table 3 Three fitted models for the 1987 baseball salary data: Mode I—Hoaglin and Velleman (1995); Model II—Augmentation Tree (AT); Model III—Residual-based Tree (RT)

Model	Adjusted R^2				
I	0.815		estimate	s.e.	z
		intercept	3.529	0.113	31.350
		runcr/yrs	0.016	0.002	9.597
		$\sqrt{\texttt{run86}}$	0.082	0.020	4.071
		$\min[(\texttt{yrs}-2)_+, 5]$	0.347	0.015	23.066
		$(\texttt{yrs}-7)_+$	−0.040	0.009	−4.434
II	0.848		estimate	s.e.	z
		intercept	3.704	0.108	34.378
		runcr/yrs	0.014	0.002	8.067
		$\sqrt{\texttt{run86}}$	0.075	0.018	4.093
		$\min[(\texttt{yrs}-2)_+, 5]$	0.286	0.019	14.998
		$(\texttt{yrs}-7)_+$	−0.047	0.008	−5.543
		node 121	0.320	0.081	3.954
		node 122	0.731	0.101	7.225
III	0.830		estimate	s.e.	z
		intercept	3.798	0.124	30.534
		runcr/yrs	0.016	0.002	9.894
		$\sqrt{\texttt{run86}}$	0.080	0.019	4.123
		$\min[(\texttt{yrs}-2)_+, 5]$	0.346	0.015	23.773
		$(\texttt{yrs}-7)_+$	−0.042	0.009	−4.747
		node 12	0.285	0.064	4.445

final augmentation tree into HV's model, and then fit all data together. The resulting hybrid model fit (Model II) is given in Table 3. Similarly, We obtain a hybrid Model III by augmenting HV's model with the final RT tree structure. Table 3 shows that the coefficients of Model II are all significant and Model II performs slightly better than Model III in terms of the adjusted R^2.

We next employ RT and RTV to examine the homoscedasticity of HV's model. Both result in a trivial final tree structure that includes the root node only, which supports the equal-variance assumption. To gain further confirmation, we conduct additional tree procedures 10 times with different training and test samples. This is similar to the idea of bagging (Breiman 1996), which is particularly useful for fully extracting diagnostic information. Because all 10 additional runs yield the trivial tree structure, we conclude that the constant variance assumption of HV's model is valid. One possible explanation of this finding is that the logarithm transformation on `salary` stabilizes the error variance.

4.2 Boston housing data

The Boston housing price data were collected by Harrison and Rubinfeld (1978) to study the effect of air pollution on real estate price in the greater Boston area in the 1970s. The data consist of 506 observations on 16 variables, with each observation pertaining to one census tract.

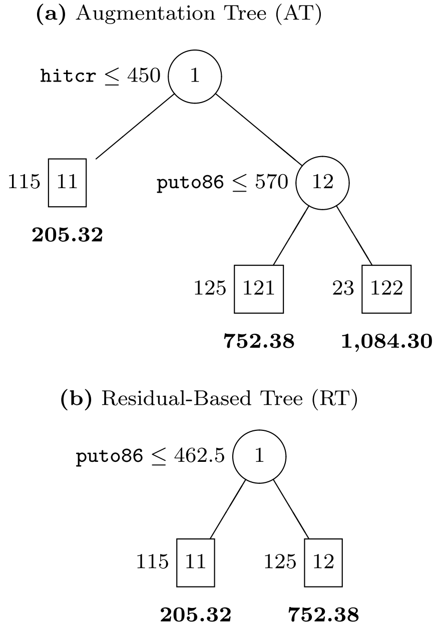


Fig. 2 The final tree structures found by AT and RT, respectively, which are augmentations to HV's (Hoaglin and Velleman 1995) model for the Baseball Salary Data. To the left of each terminal node is the number of observations in each terminal node, and underneath is the mean salary (in thousand of dollars)

To understand causality, Harrison and Rubinfeld (1978) fitted a multiple linear model that includes all predictors with some transformations. Since many of them are quite insignificant, we conduct a stepwise variable selection to obtain the following 'best' linear model:

$$\log(\mathtt{MEDV}) = \beta_0 + \beta_1\mathtt{NOX}^2 + \beta_2\mathtt{DIS} + \beta_3\mathtt{PTRATIO} + \beta_4\log(\mathtt{LSTAT}) + \varepsilon. \tag{8}$$

The ordinary least squares (OLS) estimates of the β's and their associated standard errors are given in Table 4.

We first employ AT and RT to assess the adequacy of model (8). Neither shows a non-trivial tree structure. We also conduct 10 additional runs by using different learning and test samples, and note that all result in the trivial tree structure.

We next apply RTV and TV to check heteroscedasticity. RTV results in a final tree with two terminal nodes, while TV leads to a final tree with eight terminal nodes (see Fig. 4). Hence, the equal variance assumption seems problematic for the OLS fit of model (8). We then refit model (8) by incorporating these two tree structures for error variance. Table 4 indicates that all three methods have very similar parameter estimates. This finding is not surprising, as all yield unbiased estimates. However, both TV and RTV have much smaller standard errors of parameter estimates than those of the OLS estimates. It is worth noting that TV utilizes a substantially bigger tree than that of RTV (see Fig. 4), which is often unstable (see Breiman et al. 1984). Hence, TV is less favorable in this example. To diagnose heteroscedasticity, we finally follow Carroll and Ruppert's (1988) suggestion and plot the cubic root of the squared studentized residuals versus the predicted values. Figure 5 shows that the split $\{\mathtt{CRIM} \leq 12.148\}$ can be roughly visualized.

Fig. 3 Partial residual plot for `hitcr` (**a**) and `puto86` (**b** and **c**) in the 1987 baseball salary data. Superimposed on each plot is the cutoff point selected by the tree procedure and a smooth curve given by locally linear fit (*lowess*)

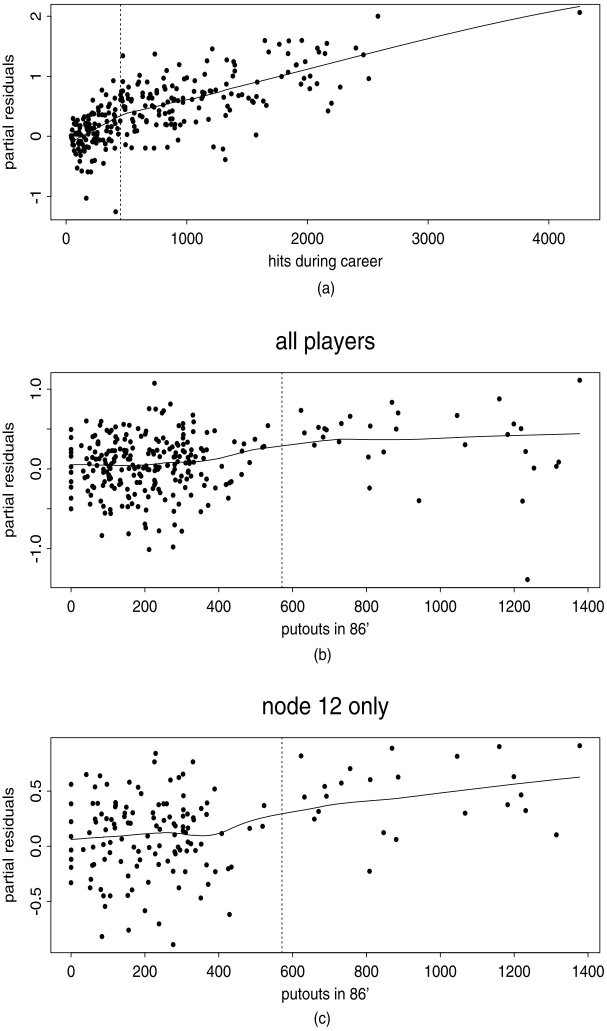


Table 4 The comparison of RTV and TV versus OLS in fitting Boston housing data

	OLS		RTV		$\frac{\text{s.e.}\hat{\beta}_{\text{OLS}}}{\text{s.e.}\hat{\beta}_{\text{RTV}}}$	TV		$\frac{\text{s.e.}\hat{\beta}_{\text{OLS}}}{\text{s.e.}\hat{\beta}_{\text{TV}}}$
	$\hat{\beta}_{\text{OLS}}$	s.e	$\hat{\beta}_{\text{RTV}}$	s.e		$\hat{\beta}_{\text{TV}}$	s.e	
`intercept`	7.807	0.345	7.823	0.227	1.520	7.158	0.121	2.859
$\texttt{NOX}^2$	−1.452	0.354	−1.487	0.234	1.512	−0.966	0.161	2.193
`DIS`	−0.053	0.022	−0.064	0.015	1.514	−0.056	0.007	3.069
`PTRATIO`	−0.070	0.016	−0.062	0.011	1.523	−0.043	0.006	2.860
log(`LSTAT`)	−0.605	0.069	−0.639	0.047	1.487	−0.544	0.024	2.949

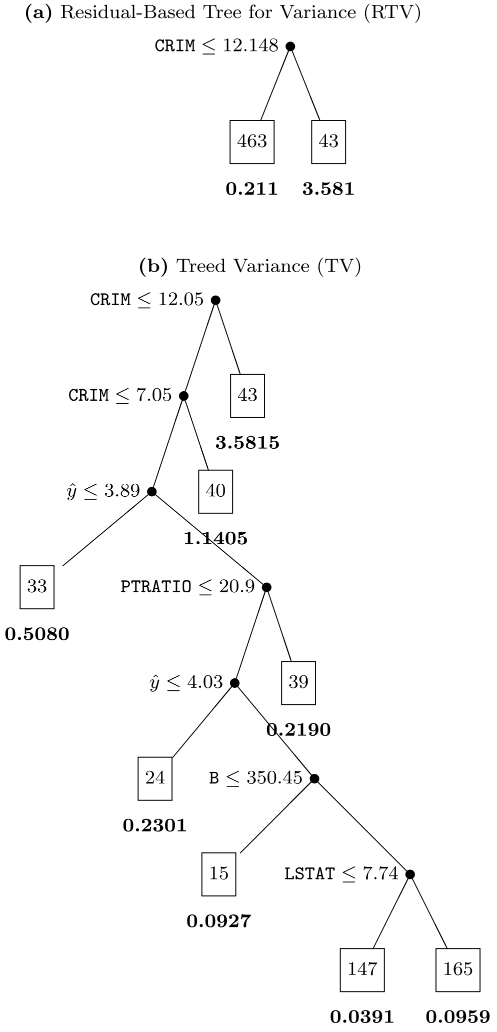


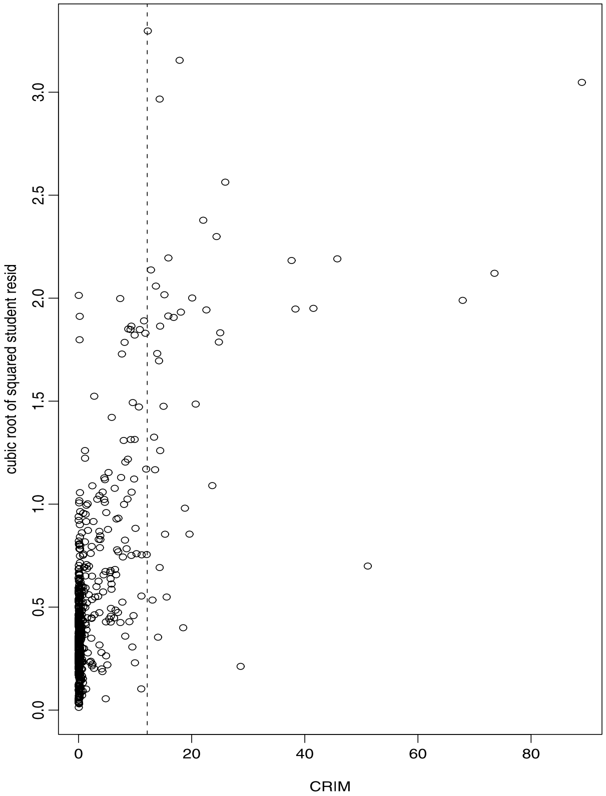
Fig. 4 The final tree structure for heteroscedasticity in model (8) with Boston housing data. Inside each terminal node t is the node size n_t, and underneath is the maximum likelihood estimate of σ_t^2, $\hat{\sigma}_t^2 = \sum_{i\in t}(y_i - \hat{y}_i)^2/n_t$

5 Discussion

In this paper, we obtain two diagnostic tree procedures for detecting the adequacy of linear regression models. One is designed to assess the linearity and the other is used to evaluate homoscedasticity. If the resulting diagnostic tree is nontrivial, then the assumption of linearity (or homoscedasticity) may not be valid. Our proposed methods not only detect possible deficiencies but also provide clues for amendments. Furthermore, these methods are particularly useful in dealing with large data sets, which often occur in data mining or machine learning.

One could extend the current work to include assessing the adequacy of logistic regression models for classification. Accordingly, the issues of lack-of-fit and over-dispersion should be considered. It is also of interest to follow an anonymous referee's suggestion to simultaneously examine the mean and variance specifications in linear regression models.

Fig. 5 The cubic root of squared studentized residuals versus `CRIM` for the Boston housing data. Superimposed on the plot is the cutoff point selected by the RTV



We believe that these efforts would further enhance the usefulness of tree-structured model diagnostics in machine learning and data analysis.

Acknowledgements We thank Professor David Page and two anonymous referees for their insightful and constructive comments, which led to a substantially improved manuscript.

Appendix A: A computationally efficient splitting method

We first rewrite model (1) as $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\mathbf{y} = (y_1, \ldots, y_n)'$ is a $n \times 1$ vector and X is an $n \times q$ matrix with i-th row $\mathbf{x}_i$. There must exist an $n \times n$ orthogonal matrix Q that triangularizes X such that

$$QX = \begin{pmatrix} R \\ 0_{(n-q)\times q} \end{pmatrix},$$

where R is $q \times q$ upper-triangular and $0_{(n-q)\times q}$ is an $(n-q) \times q$ matrix with elements 0. After applying the same orthogonal transformation to the response vector $\mathbf{y}$, we partition the resultant vector into two components

$$Q\mathbf{y} = \mathbf{u} = \begin{pmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{pmatrix},$$

where $\mathbf{u}_1$ is a $q \times 1$ vector and $\mathbf{u}_2$ is an $(n-q) \times 1$ vector. Then the sum of squared errors (SSE) for model (1) can be computed as SSE $= \| \mathbf{u}_2 \|^2$ (see, e.g., Kennedy and Gentle 1980).

Next, we consider model (3) which involves the split $\{s : x_j \le c\}$, $\mathbf{y} = X\boldsymbol{\beta} + \gamma \cdot \mathbf{1}_{\{x_j \le c\}} + \boldsymbol{\varepsilon}$. Let $X^{(s)}$ denote its design matrix, $X^{(s)} = (X|\mathbf{x}_s)$ where $\mathbf{x}_s = \mathbf{1}_{\{x_j \le c\}}$. We assume that $X^{(s)}$ is of full column rank with dimension $q+1$. To compute the SSE associated with model (3) efficiently, we utilize the available QR decomposition of X to obtain an orthogonal matrix $Q^{(s)}$ that triangularizes $X^{(s)}$. The detailed procedures are given below.

Note that $QX^{(s)} = Q(X|\mathbf{x}_s) = \left(\begin{smallmatrix} R & \mathbf{v}_1 \\ 0_{(n-q)\times q} & \mathbf{v}_2 \end{smallmatrix}\right)$, where $\left(\begin{smallmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \end{smallmatrix}\right) = Q\mathbf{x}_s = (v_1, \ldots, v_n)'$. To accomplish the triangularization of $QX^{(s)}$, we apply Householder's (1958) one-step transformation approach. Let

$$\xi^2 = \sum_{i=q+1}^{n} v_i^2,$$

$$a = \xi^2 + |v_{q+1}| \cdot \xi, \quad \text{and}$$

$$\mathbf{b} = (0, \ldots, 0, v_{q+1} + \operatorname{sign}(v_{q+1})\xi, v_{q+2}, \ldots, v_n)'.$$

Define the Householder matrix $H = I_n - \mathbf{b}\mathbf{b}'/a$. Then we have

$$H(QX^{(s)}) = \begin{pmatrix} R & \mathbf{v}_1 \\ \mathbf{0}' & -\operatorname{sign}(v_{q+1})a \\ 0_{(n-q-1)\times q} & \mathbf{0}_{(n-q-1)\times 1} \end{pmatrix},$$

an upper-triangular matrix. Let $Q^{(s)} = HQ$. Accordingly, $Q^{(s)}$ triangularizes $X^{(s)}$ and is an orthogonal matrix. Therefore, the sum of squared errors of model (3) can be computed as

$$\text{SSE}^{(s)} = \| \mathbf{u}_2^{(s)} \|^2,$$

where $\mathbf{u}_2^{(s)}$ consists of the last $n-(q+1)$ elements of the vector $\mathbf{u}^{(s)} = Q^{(s)}\mathbf{y} = HQ\mathbf{y} = H\mathbf{u}$.

Appendix B: Derivation of the score test

Let $w_i = I\{x_{ij} \le c\}$, $n_1 = \sum_{i=1}^{n} w_i$, and $n_2 = n - n_1$. The log-likelihood for the model jointly given by (1) and (5) is

$$l(\boldsymbol{\beta}, \sigma^2, \theta) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{\theta n_1}{2} - \sum_{i=1}^{n} \frac{(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2}{2\sigma^2 \exp(\theta w_i)}.$$

After algebraic simplifications, the score function and the expected information matrix are

$$U = \left.\frac{\partial l}{\partial \theta}\right|_{(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2, \theta = 0)} = -\frac{n_1}{2} + \frac{\sum w_i}{2\hat{\sigma}^2} = \frac{\sum (w_i - \bar{w})(q_i - \bar{q})}{2\hat{\sigma}^2},$$

and

$$J = -\operatorname{E}\left\{ \frac{\partial^2 l}{\partial(\boldsymbol{\beta}, \sigma^2, \theta)\, \partial(\boldsymbol{\beta}, \sigma^2, \theta)'} \right\}\Bigg|_{(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2, \theta = 0)}$$

$$= \left[\begin{array}{cc|c} \mathbf{X}'\mathbf{X}/\hat{\sigma}^2 & \mathbf{0} & \mathbf{0} \\ \mathbf{0}' & n/(2\hat{\sigma}^4) & n_1/(2\hat{\sigma}^2) \\ \hline \mathbf{0}' & n_1/(2\hat{\sigma}^2) & n_1/2 \end{array}\right] = \left[\begin{array}{c|c} J_{11} & J_{12} \\ \hline J_{21} & J_{22} \end{array}\right],$$

respectively, where $\bar{w} = n_1/n$, $\hat{\sigma}^2 = \sum q_i/n = \bar{q}$ and $\mathbf{X} = (x_1', \ldots, x_n')'$.

Adopting the approach in Breusch and Pagan's (1979) (see also Cox and Hinkley 1974, p. 324), we obtain the score test for testing $H_0: \theta = 0$ versus $H_1: \theta \neq 0$. The resulting test statistic for a given split s is

$$\mathrm{ST}^*(s) = U^2 \cdot V = \frac{\sum (q_i - \bar{q})^2}{2\hat{\sigma}^4} \left\{ \frac{\sum (w_i - \bar{w})(q_i - \bar{q})}{\sqrt{\sum (w_i - \bar{w})^2 \sum (q_i - \bar{q})^2}} \right\}^2,$$

where $V = (J_{22} - J_{21} J_{11}^{-1} J_{21})^{-1} = \frac{2n}{n_1 n_2} = \frac{2}{\sum (w_i - \bar{w})^2}$, and $\sum w_i = \sum w_i^2 = n_1$.

A drawback of the score test $\mathrm{ST}^*(s)$ is that it depends on the assumption that ε is normally distributed. Hence, we adopt Koenker's (1981) approach to replace $2\hat{\sigma}^4$ in $\mathrm{ST}^*(s)$ by $\sum (q_i - \bar{q})^2/n$ and obtain the studentized score test given in (6).

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Czaki (Eds.), *2nd int. symp. inf. theory* (pp. 267–281). Budapest: Akad Kiado.

Bickel, P. J. (1978). Using residuals robustly i: Tests for heteroscedasticity, nonlinearity. *Annals of Statistics*, *6*, 266–291.

Box, G. (1988). Signal-to-noise ratios, performance criteria, and transformation. *Technometrics*, *29*, 1–17.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*, 123–140.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees*. Belmont, CA: Wadsworth International Group.

Breusch, T. S., & Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica*, *47*, 1287–1294.

Carroll, R. J., & Ruppert, D. (1988). *Transformation and weighting in regression*. New York, NY: Chapman and Hall.

Cox, D. R., & Hinkley, D. V. (1974). *Theoretical statistics*. New York: Chapman and Hall.

Hand, D. J. (1999). Statistics and data mining: intersecting disciplines. *ACM SIGKDD*, *1*, 16–19.

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

Harvey, A. C. (1976). Estimating regression models with multiplicative heteroscedasticity. *Econometrica*, *44*, 461–465.

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

Householder, A. S. (1958). Unitary triangularization of a nonsymmetric matrix. *Journal of the Association for Computing Machinery*, *5*, 339–342.

Kennedy, W. J., & Gentle, J. E. (1980). *Statistical computing*. New York: Marcel Dekker, Inc.

Koenker, R. (1981). A note on studentizing a test for heteroscedasticity. *Journal of Econometrics*, *17*, 107–112.

Mansfield, E. R., & Conerly, M. D. (1987). Diagnostic value of residual and partial residual plots. *American Statistician*, *41*, 107–116.

Miller, T. W. (1996). Putting the cart after the horse: tree-structured regression diagnostics. In *1996 proceedings of the statistical computing section, American statistical association* (pp. 150–155).

Morgan, J., & Sonquist, J. (1963). Problems in the analysis of survey data and a proposal. *Journal of the American Statistical Association*, *58*, 415–434.

Neter, J., Kutner, M., Wasserman, W., & Nachtsheim, C. J. (1996). *Applied linear statistical models* (4th ed.). Boston, MA: McGraw-Hill.

Rao, C. R. (1947). Large sample tests of statistical hypotheses concerning several parameters with application to problems of testing. *Proceeding of the Cambridge Philosophical Society*, *44*, 50–57.

Rencher, A. C. (2000). *Linear models in statistics*. New York: Wiley

Rutemiller, H. C., & Bowers, D. A. (1968). Estimation in a heteroscedastic regression model. *Journal of American Statistical Association*, *63*, 552–557.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.

Simonoff, J. S., & Tsai, C.-L. (1994). Improved tests for nonconstant variance in regression based on the modified profile likelihood. *Journal of the Royal Statistical Society, Series C*, *43*, 357–370.

Su, X. G., & Tsai, C.-L. (2005). Tree-augmented cox proportional hazards models. *Biostatistics*, *6*, 486–499.

Su, X. G., Wang, M., & Fan, J. (2004). Maximum likelihood regression trees. *Journal of Computational and Graphical Statistics*, *13*, 586–598.

Su, X. G., Tsai, C.-L., & Yan, X. (2006). Treed variance. *Journal of Computational and Graphical Statistics*, *15*, 356–371.

Venables, W. N., & Ripley, B. D. (1999). *Modern applied statistics with S-plus* (3rd ed.). New York: Springer.