Regression by Parts: Fitting Visually Interpretable Models with GUIDE

Wei-Yin Loh

Department of Statistics, University of Wisconsin–Madison loh@stat.wisc.edu

(Published in *Handbook of Data Visualization*, pp. 447–468, Springer, 2008)

Summary. A regression model is best interpreted visually. Because we are limited to 2D displays, one way that we can fit a non-trivial model involving several predictor variables and still visually display it, is to partition the data and fit a simple model to each partition. We show how this can be achieved with a recursive partitioning algorithm called GUIDE. Further, we use examples to demonstrate how GUIDE can (i) explain ambiguities from multiple linear regression, (ii) reveal the effect of a categorical variable hidden from a sliced inverse regression model, (iii) identify outliers in data from a large and complex but poorly designed experiment, and (iv) fit an interpretable Poisson regression model to data containing categorical predictor variables.

1 Introduction

Regression modeling often requires many subjective decisions, such as choice of transformation for each variable and the type and number of terms to include in the model. The transformations may be as simple as powers and cross-products or as sophisticated as indicator functions and splines. Sometimes, the transformations are chosen to satisfy certain subjective criteria such as approximate normality of the marginal distributions of the predictor variables. Further, model building is almost always an iterative process, with the fit of the model evaluated each time terms are added or deleted.

In statistical applications, a regression model is generally considered acceptable if it satisfies two criteria. The first is that the distribution of the residuals agrees with that specified by the model. In the case of least-squares regression, this usually means normality and variance homogeneity of the residuals. The whole subject of regression diagnostics is concerned with this problem [3]. This criterion can be hard to achieve, however, in complex datasets without the fitted model becoming unwieldy. The second criterion, which is preferred

almost exclusively in the machine learning literature, is that the model has low mean prediction squared error or, more generally, deviance.

If model selection is completely software-based, the prediction deviance of an algorithm can be estimated by V-fold cross-validation as follows:

1. Randomly divide the dataset into V roughly equal parts.
2. Leaving out one part in turn, apply the algorithm to the observations in the remaining $V - 1$ parts to obtain a model.
3. Estimate the mean prediction deviance of each model by applying the left-out data to it.
4. Average the V estimates to get a cross-validation estimate for the model constructed from all the data.

The value of V may be as small as 2 for very large datasets and as large as the sample size for small datasets. But cross-validation is impractical if the model is selected not by a computer algorithm but by a person making subjective decisions at each stage. In this case, penalty-based methods such as AIC [1] are often employed. These methods select the model that minimizes a sum of the residual deviance plus a penalty term times a measure of model complexity. Although the rationale makes sense, there is no, and probably never will be, consensus on the right value of the penalty term for all datasets.

A separate, but no less important, problem is how to build a regression model that can be interpreted correctly and unambiguously. In practice, the majority of consumers of regression models often are more interested in model interpretation than in optimal prediction accuracy. They want to know which predictor variables affect the response and how they do it. Sometimes, they also want a rank ordering of the predictors according to the strength of their effects, although this question is meaningless without a more precise formulation. Nonetheless, it is a sad fact that the models produced by most regression techniques, including the most basic ones, are often difficult or impossible to interpret. Besides, even when a model is mathematically interpretable, the conclusions can be far from unambiguous.

In the rest of this article, we use four examples to highlight some common difficulties: (i) effects of collinearity on modeling Boston housing price (Sect. 2), (ii) inclusion of a categorical predictor variable in modeling New Zealand horse mussels (Sect. 4), (iii) outlier detection amid widespread confounding in U.S. automobile crash tests (Sect. 5), and (iv) Poisson regression modeling of Swedish car insurance rates (Sect. 6). We propose a divide-and-conquer strategy to solve these problems. It is based on partitioning the dataset into naturally interpretable subsets such that a relatively simple and visualizable regression model can be fitted to each subset. A critical requirement is that the partitions be free of selection bias. Otherwise, inferences drawn from the partitions may be incorrect. Another requirement is that the solution be capable of determining the number and type of partitions by itself. In Sect. 3 we present an implementation derived from the GUIDE regression

tree algorithm [12]. At the time of this writing, GUIDE is the only algorithm that has the above properties as well as other desirable features.

2 Boston housing data—effects of collinearity

The well-known Boston housing dataset was collected by Harrison and Rubinfeld [10] to study the effect of air pollution on real estate price in the greater Boston area in the 1970s. Belsley, Kuh, and Welsch [3] drew attention to the data when they used it to illustrate regression diagnostic techniques. The data consist of 506 observations on 16 variables, with each observation pertaining to one census tract. Table 1 gives the names and definitions of the variables. We use the version of the data that incorporates the minor corrections found by Gilley and Pace [8].

Table 1. Variables in Boston housing data

Variable	Definition	Variable	Definition
ID	census tract number	TOWN	township (92 values)
MEDV	median value in $1000	AGE	% built before 1940
CRIM	per capita crime rate	DIS	distance to employ. centers
ZN	% zoned for lots > 25,000 sq.ft.	RAD	accessibility to highways
INDUS	% nonretail business	TAX	property tax rate/$10K
CHAS	1 on Charles River, 0 else	PT	pupil/teacher ratio
NOX	nitrogen oxide conc. (p.p.10^9)	B	(% black - 63)^2/10
RM	average number of rooms	LSTAT	% lower-status population

Harrison and Rubinfeld [10] fitted the linear model

$$\begin{aligned}\log(\texttt{MEDV}) = {} & \beta_0 + \beta_1 \texttt{CRIM} + \beta_2 \texttt{ZN} + \beta_3 \texttt{INDUS} + \beta_4 \texttt{CHAS} + \beta_5 \texttt{NOX}^2 + \beta_6 \texttt{RM}^2 \\ & + \beta_7 \texttt{AGE} + \beta_8 \log(\texttt{DIS}) + \beta_9 \log(\texttt{RAD}) + \beta_{10} \texttt{TAX} + \beta_{11} \texttt{PT} + \beta_{12} \texttt{B} \\ & + \beta_{13} \log(\texttt{STAT})\end{aligned}$$

whose least squares estimates, t-statistics, and marginal correlation between each regressor and log(MEDV) are given in Table 2. Note the liberal use of the square and log transformations. Although many of the signs of the coefficient estimates are reasonable and expected, those of log(DIS) and log(RAD) are somewhat surprising, because their signs contradict those of their respective marginal correlations with the response variable. For example, the regression coefficient of log(DIS) is negative but the plot in Figure 1 shows a positive slope.

To resolve the contradiction, recall that the regression coefficient of log(DIS) quantifies the linear effect of the variable after the linear effects of the other variables are accounted for. On the other hand, the correlation of log(DIS)

Table 2. Least squares estimates of the coefficients and t-statistics for the regression model for log(MEDV). The marginal correlation between the response variable and each predictor is denoted by ρ.

Regressor	β	t	ρ	Regressor	β	t	ρ
Constant	4.6	30.0		AGE	7.1E-5	0.1	-0.5
CRIM	-1.2E-2	-9.6	-0.5	log(DIS)	-2.0E-1	-6.0	0.4
ZN	9.2E-5	0.2	0.4	log(RAD)	9.0E-2	4.7	-0.4
INDUS	1.8E-4	0.1	-0.5	TAX	-4.2E-4	-3.5	-0.6
CHAS	9.2E-2	2.8	0.2	PT	-3.0E-2	-6.0	-0.5
NOX^2	-6.4E-1	-5.7	-0.5	B	3.6E-4	3.6	0.4
RM^2	6.3E-3	4.8	0.6	log(LSTAT)	-3.7E-1	-15.2	-0.8

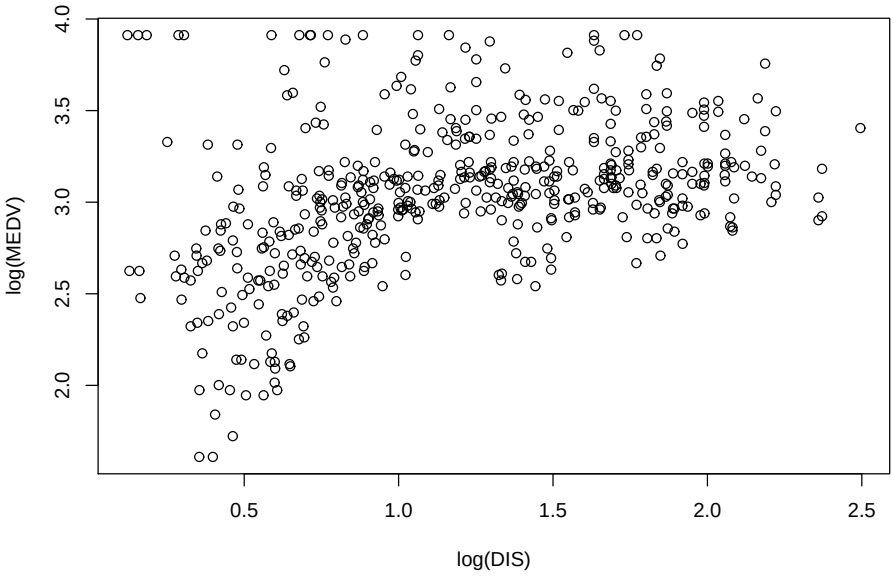

Fig. 1. Plot of log(MEDV) versus log(DIS) for Boston data

with the response variable ignores the effects of the other variables. Since it is important to take the other variables into consideration, the regression coefficient may be a better measure of the effect of log(DIS). But this conclusion requires the linear model assumption to be correct. Nonetheless, it is hard to explain the negative linear effect of log(DIS) when we are faced with Figure 1.

The problem of contradictory signs vanishes when there is only one regressor variable. Although it can occur with two regressor variables, the difficulty is diminished because the fitted model can be visualized through a contour plot. For datasets that contain more than two predictor variables, we propose a divide-and-conquer strategy. Just as a prospective buyer inspects a house

one room at a time, we propose to partition the dataset into pieces such that a visualizable model involving one or two predictors suffices for each piece. One difficulty is that, unlike a house, there are no predefined "rooms" or "walls" in a dataset. Arbitrarily partitioning a dataset makes as much sense as arbitrarily slicing a house into several pieces. We need a method that gives interpretable partitions of the dataset. Further, the number and kind of partitions should be dictated by the complexity of the dataset as well as the type of models to be fitted. For example, if a dataset is adequately described by a non-constant simple linear regression involving one predictor variable and we fit a piecewise linear model to it, then no partitioning is necessary. On the other hand, if we fit a piecewise constant model to the same dataset, the number of partitions should increase with the sample size.

The GUIDE regression tree algorithm [12] provides a ready solution to these problems. GUIDE can recursively partition a dataset and fit a constant, best polynomial, or multiple linear model to the observations in each partition. Like the earlier CART algorithm [4], which fits piecewise constant models only, GUIDE first constructs a nested sequence of tree-structured models and then uses cross-validation to select the smallest one whose estimated mean prediction deviance lies within a short range of the minimum estimate. But unlike CART, GUIDE employs lack-of-fit tests of the residuals to choose a variable to partition at each stage. As a result, it does not have the selection bias of CART and other algorithms that rely solely on greedy optimization.

To demonstrate a novel application of GUIDE, we use it to study the linear effect of log(`DIS`) after controlling for the effects of the other variables, *without* making the linear model assumption. We do this by constructing a GUIDE regression tree in which log(`DIS`) is the sole linear predictor in each partition or node of the tree. The effects of the other predictor variables, which need not be transformed, can be observed through the splits at the intermediate nodes. Figure 2 shows the tree, which splits the data into twelve nodes. The regression coefficients are between -0.2 and 0.2 in all but four leaf nodes. These nodes are colored red (for slope less than -0.2) and blue (for slope greater than 0.2). We choose the cut-off values of ± 0.2 because the coefficient of log(`DIS`) in Table 2 is 0.2. The tree shows that the linear effect of log(`DIS`) is neither always positive nor always negative—it depends on the values of the other variables. This explains the contradiction between the sign of the multiple linear regression coefficient of log(`DIS`) and that of its marginal correlation. Clearly, a multiple linear regression coefficient is, at best, an average of several conditional simple linear regression coefficients.

Figure 3 explains the situation graphically by showing the data and the twelve regression lines and their associated data points, using blue triangles and red circles for observations associated with slopes greater than 0.2 and less than -0.2, respectively, and green crosses for the others. The plot shows that, after we allow for the effects of the other variables, log(`DIS`) generally has little effect on median house price, except in four groups of census tracts (triangles and circles) that are located relatively close to employment centers

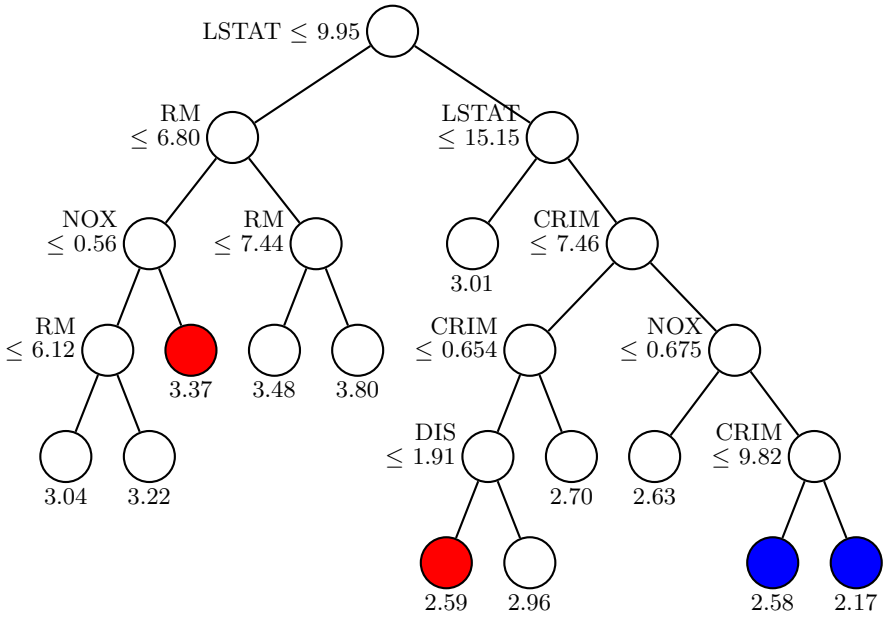

Fig. 2. GUIDE model for log(`MEDV`), using log(`DIS`) as linear predictor in each node. At each branch, a case goes to the left child node if and only if the given condition is satisfied. The sample mean of log(`MEDV`) is printed beneath each leaf node. A blue colored leaf node indicates a slope coefficient greater than 0.2. Correspondingly, a red colored leaf node is associated with a slope coefficient less than -0.2.

(log(`DIS`) < 1). According to Figure 2, the groups denoted by blue triangles are quite similar. They contain a large majority of the lower-priced tracts and have high values of `LSTAT` and `CRIM`. The two groups composed of red circles, on the other hand, are quite different from each other. One group contains tracts in Beacon Hill and Back Bay, two high-priced towns near Boston. The other group contains tracts with `DIS` lying within a narrow range and with mostly below-average `MEDV` values. Clearly, the regression coefficient of log(`DIS`) in Table 2 cannot possibly reveal such details. Unfortunately, this problem is by no means rare. Friedman and Wall [7], for example, found a similar problem that involves different variables in a subset of these data.

3 Extension to GUIDE

The basic GUIDE procedure for fitting piecewise constant and piecewise multiple linear models is described in [12]. We present here an extension to fit piecewise simple linear models. The same ideas apply to Poisson regression and to piecewise linear two-predictor models, where the two predictors are

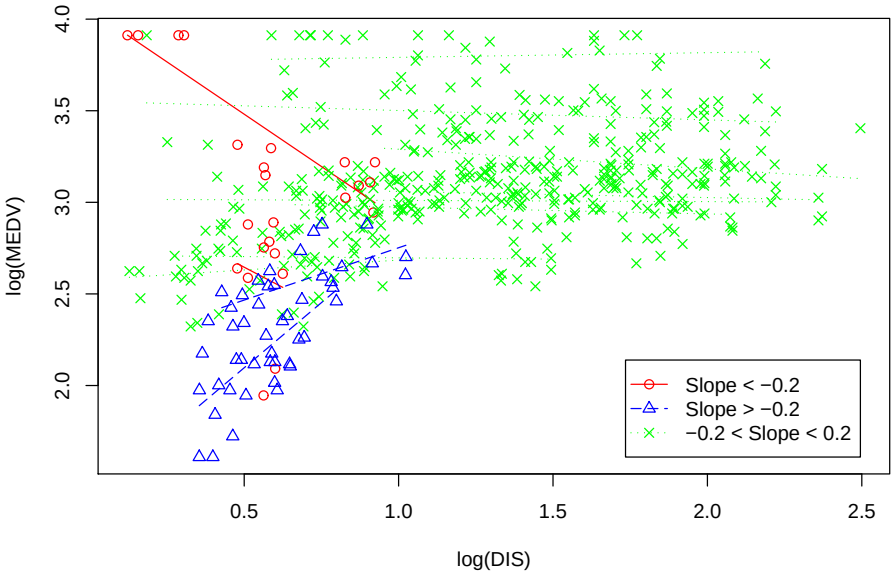

Fig. 3. Data points and regression lines in the twelve leaf nodes of the Boston data tree. The blue and red colors correspond to those in Figure 2.

chosen at each node via stepwise regression, subject to the standard F-to-enter and F-to-remove threshold values of 4.0 [13]. Our extension comprises four algorithms, starting with Algorithm 1.

Algorithm 1 (Tree construction) These steps are applied recursively to each node of the tree, starting with the root node that holds the whole dataset.

1. Let t denote the current node. Fit a simple linear regression to each predictor variable in the data in t. Choose the predictor yielding the smallest residual mean squared error and record its model R^2.
2. Stop if $R^2 > 0.99$ or if the number of observations is less than $2n_0$, where n_0 is a small user-specified constant. Otherwise, go to the next step.
3. For each observation associated with a positive residual, define the class variable $Z = 1$; else define $Z = 0$.
4. Use Algorithm 2 to find a variable X' to split t into left and right subnodes t_L and t_R.
   a) If X' is ordered, search for a split of the form $X' \le x$. For every x such that t_L and t_R contain at least n_0 observations each, find S, the smallest total sum of squared residuals obtainable by fitting a simple linear model to the data in t_L and t_R separately. Select the smallest value of x that minimizes S.
   b) If X' is categorical, search for a split of the form $X' \in C$, where C is a subset of the values taken by X'. For every C such that t_L and t_R

have at least n_0 observations each, calculate the sample variances of Z in t_L and t_R. Select the set C for which the weighted sum of the variances is minimum, with weights proportional to sample sizes in t_L and t_R.

5. Apply step 1 to t_L and t_R separately.

Algorithm 2 (Split variable selection)

1. Use Algorithms 3 and 4 to find the smallest curvature and interaction p-values $p^{(c)}$ and $p^{(i)}$ and their associated variables $X^{(c)}$ and $\{X_1^{(i)}, X_2^{(i)}\}$.
2. If $p^{(c)} \leq p^{(i)}$, define $X' = X^{(c)}$ to be the variable to split t.
3. Otherwise, if $p^{(c)} > p^{(i)}$, then:
   a) If either $X_1^{(i)}$ or $X_2^{(i)}$ is categorical, define $X' = X_1^{(i)}$ if it has the smaller curvature p-value; otherwise, define $X' = X_2^{(i)}$.
   b) Otherwise, if $X_1^{(i)}$ and $X_2^{(i)}$ are both ordered variables, search over all splits of t along $X_1^{(i)}$. For each split into subnodes t_L and t_R, fit a simple linear model on $X_1^{(i)}$ to the data in t_L and t_R separately and record the total sum of squared residuals. Let S_1 denote the smallest total sum of squared residuals over all possible splits of t on $X_1^{(i)}$. Repeat the process with $X_2^{(i)}$ and obtain the corresponding smallest total sum of squared residuals S_2. If $S_1 \leq S_2$, define $X' = X_1^{(i)}$; otherwise, define $X' = X_2^{(i)}$.

Algorithm 3 (Curvature tests)

1. For each predictor variable X:
   a) Construct a $2 \times m$ cross-classification table. The rows of the table are formed by the values of Z. If X is a categorical variable, its values define the columns, i.e., m is the number of distinct values of X. If X is quantitative, its values are grouped into four intervals at the sample quartiles and the groups constitute the columns, i.e., $m = 4$.
   b) Compute the significance probability of the chi-squared test of association between the rows and columns of the table.
2. Let $p^{(c)}$ denote the smallest significance probability and let $X^{(c)}$ denote the associated X variable.

Algorithm 4 (Interaction tests)

1. For each pair of variables X_i and X_j, carry out the following interaction test:
   a) If X_i and X_j are both ordered variables, divide the (X_i, X_j)-space into four quadrants by splitting the range of each variable into two halves at the sample median; construct a 2×4 contingency table using the Z values as rows and the quadrants as columns. After dropping any columns with zero column totals, compute the chi-squared statistic and its p-value.

   b) If X_i and X_j are both categorical variables, use their value-pairs to divide the sample space. For example, if X_i and X_j take c_i and c_j values, respectively, the chi-squared statistic and p-value are computed from a table with two rows and number of columns equal to $c_i c_j$ less the number of columns with zero totals.
   c) If X_i is ordered and X_j is categorical, divide the X_i-space into two at the sample median and the X_j-space into as many sets as the number of categories in its range—if X_j has c categories, this splits the (X_i, X_j)-space into $2c$ subsets. Construct a $2\times 2c$ contingency table with the signs of the residuals as rows and the $2c$ subsets as columns. Compute the chi-squared statistic and its p-value, after dropping any columns with zero totals.
2. Let $p^{(i)}$ denote the smallest p-value and let $X_1^{(i)}$ and $X_2^{(i)}$ denote the pair of variables associated with $p^{(i)}$.

After Algorithm 1 terminates, we prune the tree with the method described in [4, Sec. 8.5] using V-fold cross-validation. Let E_0 be the smallest cross-validation estimate of prediction mean squared error and let α be a positive number. We select the smallest subtree whose cross-validation estimate of mean square error is within α times the standard error of E_0. To prevent large prediction errors caused by extrapolation, we also truncate all predicted values so that they lie within the range of the data values in their respective nodes. The examples here employ the default values of $V = 10$ and $\alpha = 0.5$; we call this the *half-SE rule*.

Our split selection approach is different from that of CART, which constructs piecewise constant models only and which searches for the best variable to split and the best split point simultaneously at each node. This requires the evaluation of all possible splits on every predictor variable. Thus, if there are K ordered predictor variables each taking M distinct values at a node, $K(M-1)$ splits have to be evaluated. To extend the CART approach to piecewise linear regression, two linear models must be fitted for each candidate split. This means that $2K(M-1)$ regression models must be computed before a split is found. The corresponding number of regression models for K categorical predictors each having M distinct values is $2K(2^{M-1}-1)$. GUIDE, in contrast, only fits regression models to variables associated with the most significant curvature or interaction test. Thus the computational savings can be substantial. More important than computation, however, is that CART's variable selection is inherently biased toward choosing variables that permit more splits. For example, if two ordered variables are both independent of the response variable, the one with more unique values has a higher chance of being selected by CART. GUIDE does not have such bias because it uses p-values for variable selection.

4 Mussels—categorical predictors and SIR

In this section, we use GUIDE to re-analyze a dataset, previously studied by Cook [6], to show that GUIDE can deal with categorical predictor variables as naturally and easily as continuous variables. The data are from the Division of Water Science, DSIR, New Zealand [5]. They contain measurements on two hundred and one horse mussels taken from five Sites in the Marlborough Sounds, New Zealand, in December 1984. Besides Site, each mussel's Length, Width, Depth (all in mm), Gender (male, female, or indeterminate), Viscera mass, Muscle mass, and Shell mass (all in gm) were recorded, as well as the type of Peacrab (five categories) found living in its shell.

Cook [6, p. 214] used Muscle as the response variable and Length, Depth, and Shell as predictors to illustrate his approach to graphical regression. [Note: Cook used the symbols L, W, and S to denote Length, Depth and Shell, respectively.] With the aid of sliced inverse regression [11] and power transformations, he finds that the mean of Muscle can be modeled by the one-dimensional subspace defined by the variable

$$\texttt{SIR1} = 0.001\,\texttt{Length} + 0.073\,\texttt{Depth}^{0.36} + 0.997\,\texttt{Shell}^{0.11}. \tag{1}$$

Figure 4 shows the banana-shaped plot of Muscle versus SIR1.

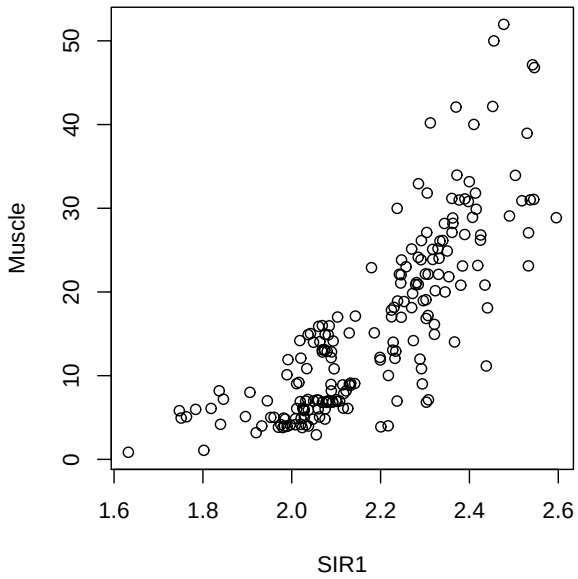

Fig. 4. Plot of Muscle versus SIR1 (slightly jittered to reduce over-plotting)

The variable Site is not used in formula (1) because, unlike GUIDE, sliced inverse regression does not easily handle categorical predictor variables. Fig-

ure 5 shows the result of fitting a GUIDE piecewise best simple linear model to the data. The tree splits first on Site. If Site is neither 2 nor 3, the tree splits further on Depth. The best simple linear predictor is Shell at two of the leaf nodes and Width at the third. Figure 6 shows the data and the fitted lines in the leaf nodes of the tree. The plots look quite linear.

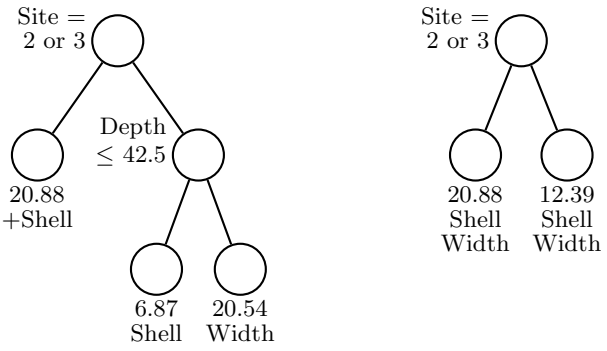

Fig. 5. Piecewise best simple linear (left) and best two-variable linear (right) least-squares GUIDE models for mussels data. At each intermediate node, a case goes to the left child node if and only if the condition is satisfied. Beneath each leaf node are the sample mean of Muscle and the selected linear predictors.

On the right side of Figure 5 is the piecewise best two-variable GUIDE model. It splits the data into two pieces, using the same top-level split as the piecewise best simple linear model. Shell and Width are selected as the best pair of linear predictors in both leaf nodes. Figure 7 shows shaded contour plots of the fitted functions and data points. Clearly, the mussels from Sites 2 and 3 tend to have greater Muscle mass than those from Sites 1, 4, and 5.

Since Site is an important predictor in the GUIDE models, we redraw the SIR plot using different symbols to indicate Site information in panel (a) of Figure 8. The banana-shaped plot is seen to be an artifact caused by combining the Sites; the data points within each Site are quite linear. Panel (b) again employs different symbols to indicate leaf node membership according to the piecewise best simple linear model in Figure 6. We see that node membership divides the data into three clusters, with the first cluster belonging to Sites 2 and 3, and the second and third clusters to Sites 1, 4, and 5, depending on whether or not Depth ≤ 42.5. The first cluster (indicated by circles) clearly exhibits the most pronounced curvature. This suggests that the nonlinear relationship between Muscle and SIR1 is mainly due to the observations from Sites 2 and 3. On the other hand, we saw in Figure 6(a) that at these two Sites, Muscle varies roughly linearly with Shell. Thus it is likely that the curvature in Figure 4 is at least partly due to the power transformation of Shell in the definition of SIR1 in formula (1).

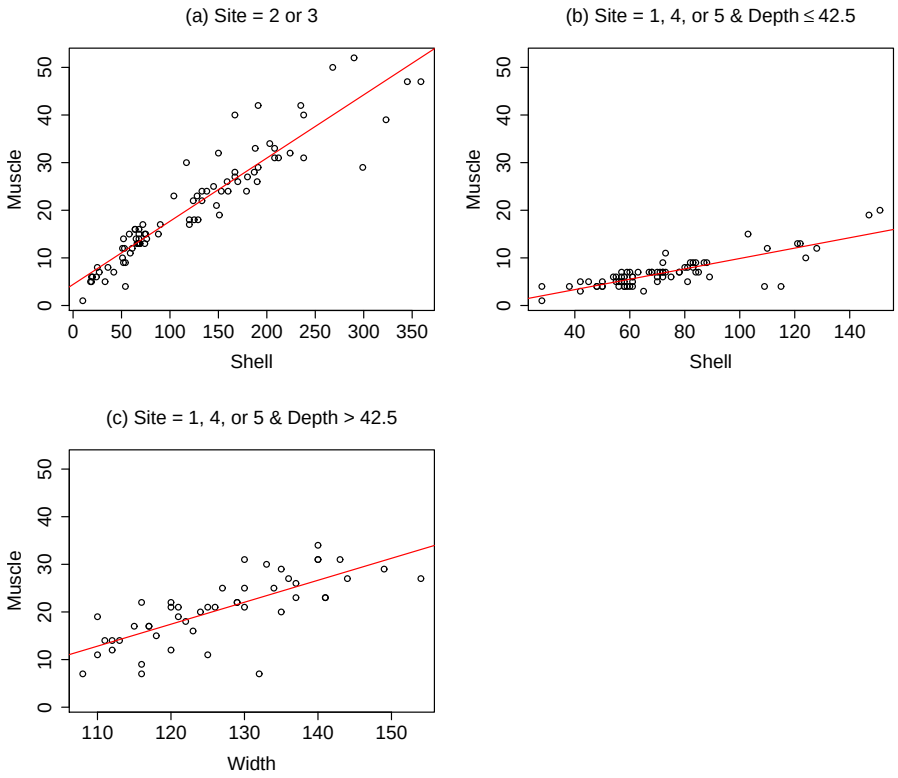

Fig. 6. Data and fitted lines for the piecewise best simple linear GUIDE model on the left side of Figure 5

5 Crash tests—outlier detection under confounding

The data in this example are obtained from 1,789 vehicle crash tests performed by the National Highway Transportation Safety Administration (NHTSA) between 1972 and 2004 (`http://www-nrd.nhtsa.dot.gov`). The response variable is the square root of the head injury criterion (`hic`) measured on a crash dummy. Values of $\sqrt{\texttt{hic}}$ range from 0 to 100, with 30 being the approximate level beyond which a person is expected to suffer severe head injury. Twenty-five predictor variables, defined in Table 3, provide information on the vehicles, dummies, and crash tests. Angular variables are measured clockwise, with -90, 0, and 90 degrees corresponding to the driver's left, front, and right sides, respectively. About one-quarter of the vehicle models are tested more than once, with the most often tested being the 1982 Chevy Citation, which was tested fifteen times.

Our goal is to identify the vehicle models for which the `hic` values are unusually high, after allowing for the effects of the predictor variables. Since

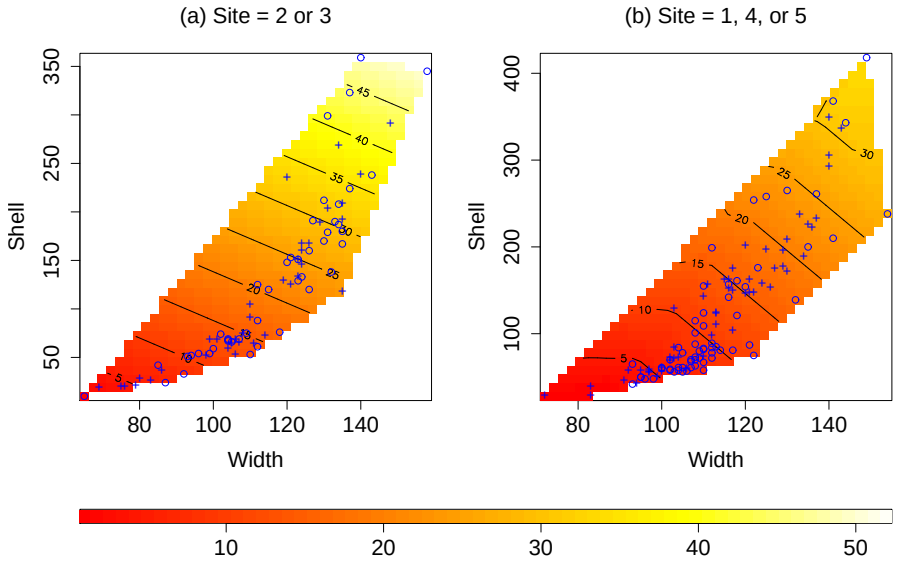

Fig. 7. Shaded contour plots of fitted functions and data points for the piecewise best two-variable linear model on the right side of Figure 5

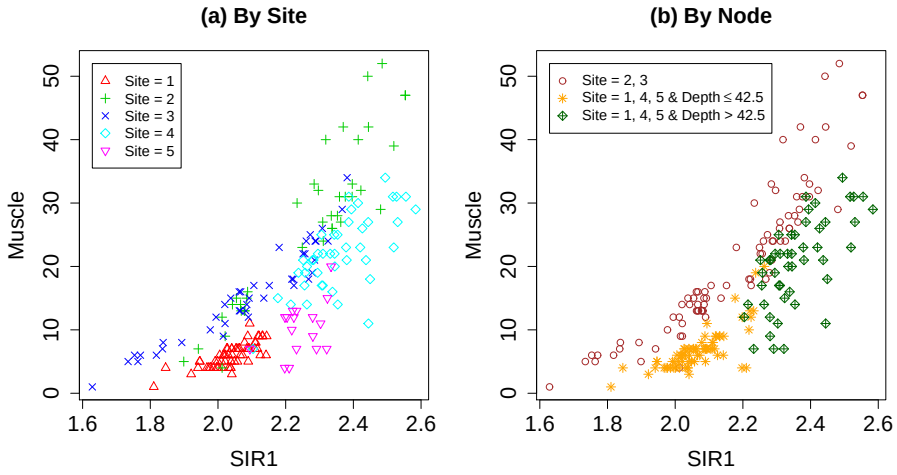

Fig. 8. Muscle versus SIR1 by Site and by the nodes of the tree in Figure 5(a)

Table 3. Variables for NHTSA data

Name	Description	Name	Description
`hic`	Head injury criterion	`make`	Car manufacturer (62 values)
`year`	Car model year	`mkmodel`	Car model (464 values)
`body`	Car body type (18 values)	`transm`	Transmission type (7 values)
`engine`	Engine type (15 values)	`engdsp`	Engine displacement (liters)
`vehtwt`	Vehicle total weight (kg)	`colmec`	Collapse mechanism (11 values)
`vehwid`	Vehicle width (mm)	`modind`	Car modification indicator (5 values)
`vehspd`	Vehicle speed (km/h)	`crbang`	Crabbed angle (degrees)
`tksurf`	Track surface (5 values)	`pdof`	Principal direction of force (degrees)
`tkcond`	Track condition (6 values)	`impang`	Impact angle (degrees)
`occtyp`	Occupant type (10 values)	`dumsiz`	Dummy size (6 values)
`seposn`	Seat position (5 values)	`barrig`	Barrier rigidity (rigid/deformable)
`barshp`	Barrier shape (14 values)	`belts`	Seat belt type (none/2pt/3pt)
`airbag`	Airbag present (yes/no)	`knee`	Knee restraint present (yes/no)

almost all the tests involve two or more crash dummies, we will give two separate analyses, one for the driver and another for the front passenger dummies. After removing tests with incomplete values, we obtain 1,633 and 1,468 complete tests for driver and front passenger, respectively. The tests for driver dummies involve 1,136 different vehicle models. Figure 9 shows a histogram of the $\sqrt{\texttt{hic}}$ values for the driver data (the histogram for front passenger is similar). There are twenty-two vehicle models with $\sqrt{\texttt{hic}}$ values greater than 50. They are listed in Table 4, arranged by model year, with the total number of times tested and (within parentheses) the $\sqrt{\texttt{hic}}$ values that exceed 50. For example, the 2000 Nissan Maxima was tested eight times, of which five gave $\sqrt{\texttt{hic}}$ values greater than 50.

To identify the outliers after removing the effects of the predictor variables, we need to regress the response values on the predictors. The regression model must be sufficiently flexible to accommodate the large number and mix of predictor variables and to allow for nonlinearity and interactions among them. It must also be suitable for graphical display, as the outliers will be visually identified. These requirements are well-satisfied by a piecewise simple linear GUIDE model, which is shown in Figure 10. The tree has three leaf nodes, partitioning the data according to `vehspd`. Beneath each leaf node is printed the sample mean response for the node and the selected signed linear predictor. We see that model year is the most important linear predictor in two of the three leaf nodes, and `impang` in the third. In the latter (Node 4), injury tends to be more severe if the impact occurs on the driver side ($\texttt{impang} = -90$).

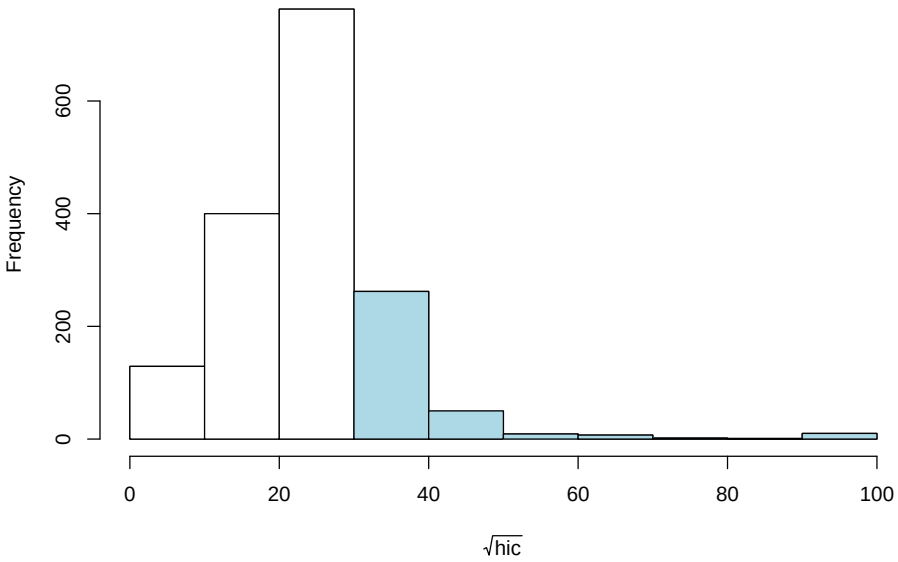

Fig. 9. Histogram of $\sqrt{\texttt{hic}}$ for driver dummy data. Shaded areas correspond to $\sqrt{\texttt{hic}} > 30$.

Table 4. Vehicles with $\sqrt{\texttt{hic}}$ (in parentheses) greater than 50 registered on driver dummies. The column labeled # gives the total number of each model tested. For example, four out of eight 2000 Nissan Maxima's tested had $\sqrt{\texttt{hic}} > 50$.

# Model	# Model
1 1979 Dodge Colt (96)	2 1983 Renault Fuego (57)
12 1979 Honda Civic (53)	1 1984 Ford Tempo (54)
1 1979 Mazda B2000 Pickup (55)	1 1988 Chevy Sportvan (61)
1 1979 Peugeot 504 (68)	1 1990 Ford Clubwagon MPV (51)
2 1979 Volkswagen Rabbit (62)	4 1995 Honda Accord (94)
3 1980 Chevy Citation (65)	5 2000 Nissan Altima (100)
1 1980 Honda Civic (52)	8 2000 Nissan Maxima (69, 72, 100, 100, 100)
1 1980 Honda Prelude (55)	4 2000 Saab 38235 (72)
2 1981 Mazda GLC (51)	4 2000 Subaru Legacy (100, 100)
2 1982 Chrysler Lebaron (51)	11 2001 Saturn L200 (68, 87, 100)
2 1982 Renault Fuego (61)	9 2002 Ford Explorer (100)

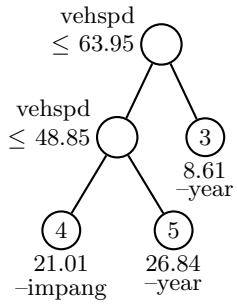

Fig. 10. Piecewise-simple linear GUIDE model for driver data. At each intermediate node, a case goes to the left child node if and only if the condition is satisfied. Beneath each leaf node are the sample mean of $\sqrt{\texttt{hic}}$ and the selected signed linear predictor.

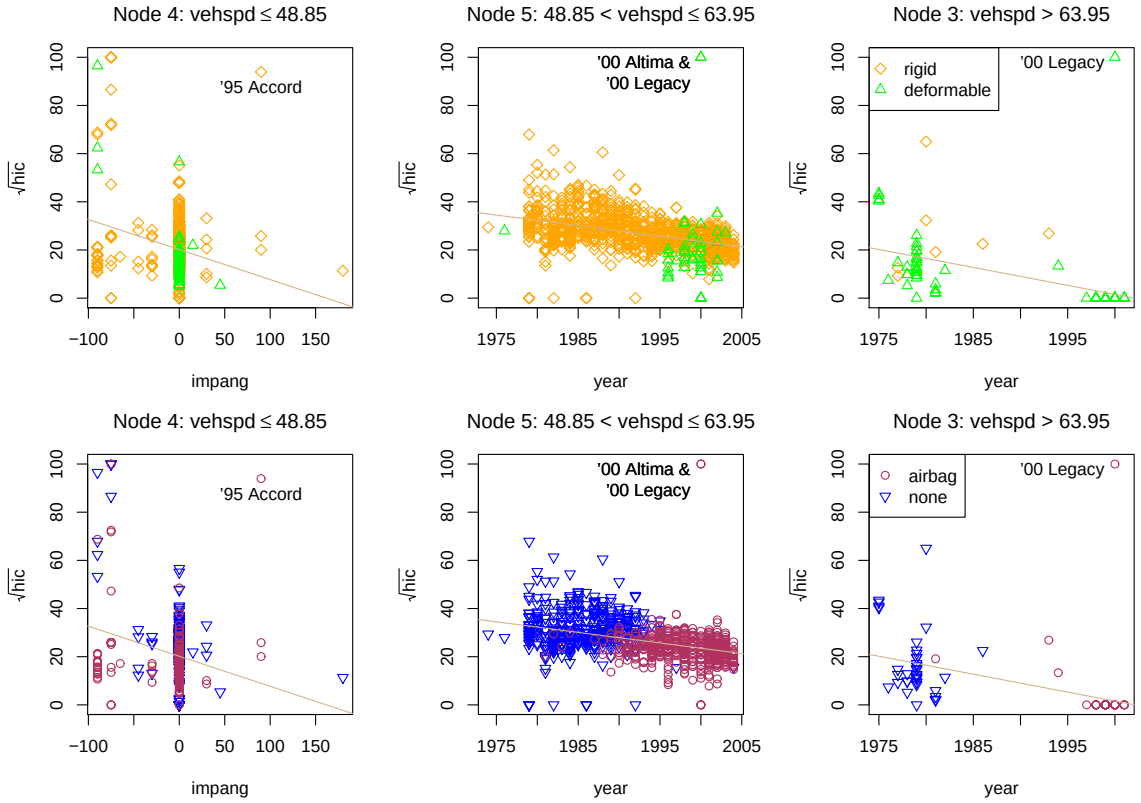

Fig. 11. Data and fitted regression functions in the leaf nodes of the tree model in Figure 10, using different symbols for `barrig` (top) and `airbag` (bottom) values

A very interesting feature of the tree is that the sample mean response is lowest in Node 3, which has the highest values of `vehspd` (> 63.95). At first glance, this does not make sense because injury severity should be positively correlated with vehicle speed. It turns out that the design of the experiment causes some variables to be confounded. This is obvious from the upper row of plots in Figure 11, which show the data and regression lines in the three leaf nodes of the tree, using different symbols to indicate whether a vehicle is crashed into a rigid or a deformable barrier. We see that the proportion of tests involving deformable barriers is much greater at high speeds (Node 3) than at low speeds. This would explain the lower injury values among the high-speed tests.

Another variable confounded with `vehspd` is `airbag`. This can be seen in the second row of plots in the same Figure, where different symbols are used to indicate whether a vehicle is equipped with an airbag or not. We see that almost all vehicles manufactured from 1990 onwards have airbags and that their presence is associated with lower `hic` values. Since there is a fair number of such vehicles in Node 3, this could also account for the low sample mean response.

Finally, a third confounding variable is evident in Figure 12, which shows barplots of the proportions of barrier shape type (`barshp`) within each leaf node of the tree. Node 3, whose bars are colored green, stands out in that barrier shapes EOB, GRL, IAT, MBR, and SGN practically never appear in the other two nodes. For some reason, the testers seem to prefer these barrier shapes for high speed crashes. Thus barrier shape is yet another possible explanation for the low mean response value in the node.

Despite these difficulties, it is clear from the plots that three vehicle models stand out as outliers: 1995 Honda Accord, 2000 Nissan Altima, and 2000 Subaru Legacy. All are foreign imports. The 2000 Subaru Legacy appears as an outlier in two separate tests, one at moderate speed and one at high speed.

Figure 13 shows the corresponding tree model for the front passenger data. Now airbag and barrier rigidity appear as split variables after the top-level split on `vehspd`. The plots of the data in the leaf nodes are presented in Figure 14. Everything seems to make sense: injury is less severe when a vehicle is equipped with airbags and when it is crashed into a deformable barrier, and also if impact occurs on the driver side (Node 6). It is interesting to note that in Node 5, where `vehspd` ≤ 48.25 and the vehicles are equipped with airbags, rigid barriers are used for the higher speeds and deformable barriers for the lower speeds. This may exaggerate the effect of `vehspd` in this node. The outliers for these data turn out to be all domestic models: 1978 Chevy Blazer, 1982 Chevy Citation, 1994 Ford Ram 150, 1998 Ford Contour, and 1999 Dodge Intrepid.

The good news from both analyses is that no obvious outliers are found among vehicles newer than the 2000 model year.

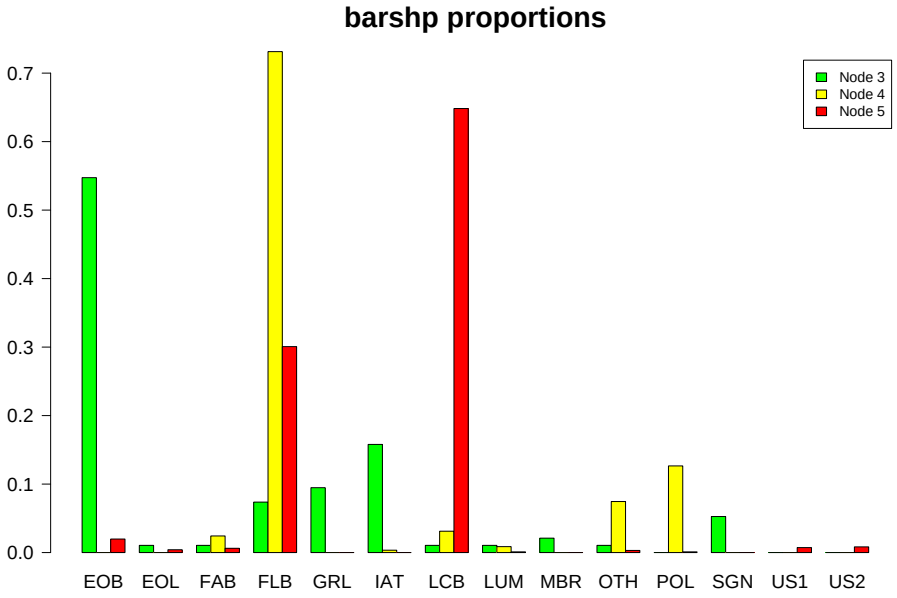

Fig. 12. Proportions of different barrier shapes within the three leaf nodes of the tree model in Figure 10. The lengths of the bars sum to one for each color.

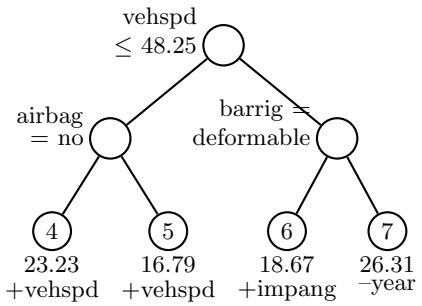

Fig. 13. Piecewise-simple linear GUIDE model for front passenger data. At each intermediate node, a case goes to the left child node if and only if the condition is satisfied. Beneath each leaf node are the sample mean of $\sqrt{\texttt{hic}}$ and the selected signed linear predictor.

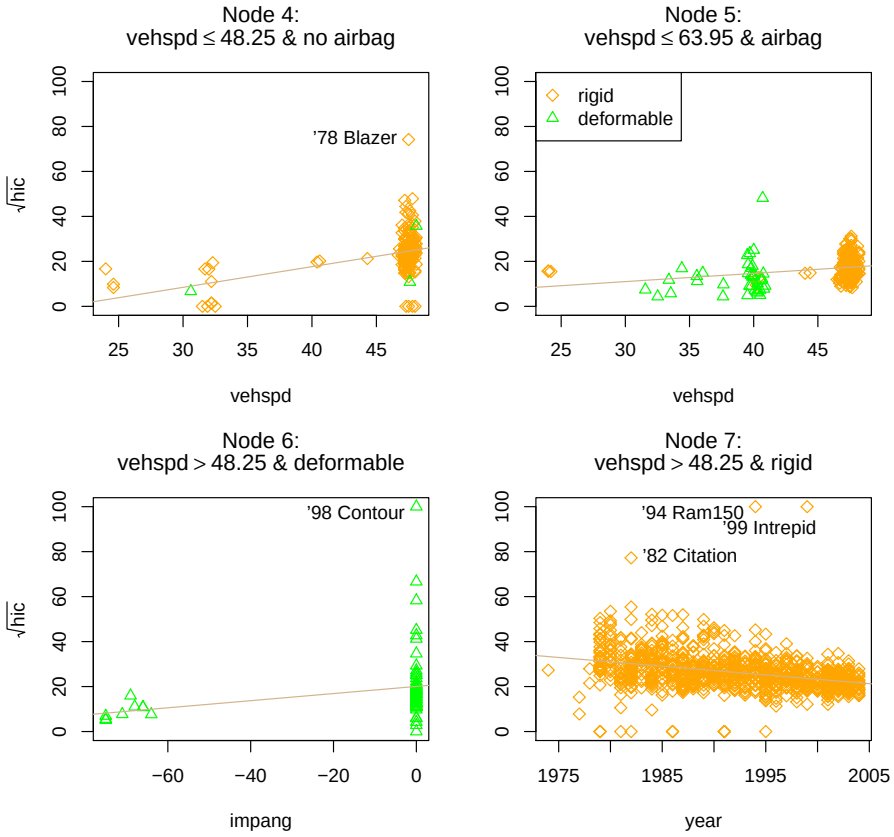

Fig. 14. Data and fitted regression functions in the leaf nodes of the tree model in Figure 13, with different symbols for `barrig` type

6 Car insurance rates—Poisson regression

The data are from Statlib. A subset of it is given in Andrews and Herzberg [2, pp. 415–421]. The original data consist of information on more than two million third-party automobile insurance policies in Sweden for the 1977 year. For each policy was recorded the annual mileage, bonus class (on a seven-point scale), geographical zone (seven categories), and make of car (nine categories). Annual mileage is discretized into five categories — (1) less than 10,000 km/yr, (2) 10,000–15,000 km/yr, (3) 15,000–20,000 km/yr, (4) 20,000–25,000 km/yr, and (5) more than 25,000 km/yr, see [9]. These four explanatory variables yield a $5 \times 7 \times 7 \times 9$ table with 2205 possible cells. For each cell, the following quantities were obtained:

1. total insured time in years,
2. total number of claims,

3. total monetary value of the claims.

Twenty-three cells are empty.

We will model claim rate here. According to [2, p. 414], a Swedish Analysis of Risk group decided that a multiplicative model (i.e., an additive Poisson loglinear model) for claim rate is fairly good, and that any better model is too complicated to administer. To challenge this conclusion, we will use GUIDE to fit a piecewise-additive Poisson loglinear model for number of claims, using the log of number of claims as offset variable. Bonus class and mileage class are treated as continuous, and zone and make as categorical variables.

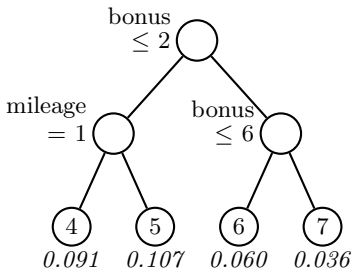

Fig. 15. GUIDE multiple linear Poisson regression tree for car insurance data. At each intermediate node, a case goes to the left child node if and only if the condition is satisfied. The number in italics beneath each leaf node is the sample claim rate.

Figure 15 shows the GUIDE tree, which has four leaf nodes and an estimated prediction deviance (based on ten-fold cross-validation) about 25% lower than that of a single additive loglinear model. From the sample average claim rates printed beneath the leaf nodes, we see that bonus classes 1 and 2 tend to yield rates two to three times as large as the other bonus classes.

The estimated regression coefficients in the leaf nodes are given in Table 5, where the coefficients of the dummy variables corresponding to the first levels of each categorical variable are set to zero. As may be expected, mileage has a positive slope coefficient in all three nodes where it is not constant. The slope for bonus is, however, negative wherever it is not constant. Thus the higher the bonus class, the lower the claim rate tends to be.

For make, the coefficient for level 4 has a larger negative value than the coefficients for the other make levels, uniformly across all the nodes. Hence this level of make is likely to reduce claim rate the most. In contrast, the coefficient for level 5 of make is positive in all nodes and is larger than the coefficients for all other levels in three nodes—it is second largest in the remaining node. This level of make is thus most likely to increase claim rate. The situation is quite similar for zone: since all its coefficients are negative except for level 1, which is set to zero, that level is most likely to increase claim rate, across all four nodes. The zone level most likely to decrease claim rate is 7, which has

Table 5. Regression estimates for GUIDE model, using set-to-zero constraints for the first levels of make and zone

	Node 4	Node 5	Node 6	Node 7
Constant	-0.8367	-1.0639	-2.3268	-3.3725
mileage	aliased	0.0427	0.1425	0.1439
bonus	-0.5202	-0.4500	-0.0992	aliased
make=2	-0.1705	-0.0356	0.0756	0.1375
make=3	-0.2845	-0.2763	-0.2038	-0.2247
make=4	-1.0964	-0.7591	-0.6555	-0.4595
make=5	0.0892	0.1685	0.1468	0.1308
make=6	-0.5971	-0.5437	-0.3274	-0.2563
make=7	-0.3330	-0.2900	-0.0405	0.0214
make=8	-0.0806	-0.0848	0.0233	-0.0584
make=9	-0.4247	-0.2097	-0.0592	0.0039
zone=2	-0.3306	-0.2735	-0.2525	-0.1837
zone=3	-0.5220	-0.3905	-0.4046	-0.3303
zone=4	-0.8298	-0.5692	-0.5986	-0.5120
zone=5	-0.4683	-0.3927	-0.3533	-0.2384
zone=6	-0.7414	-0.5437	-0.5830	-0.4273
zone=7	-0.8114	-0.8538	-0.7760	-0.6379

the largest negative coefficient in three of the nodes, and the second largest negative coefficient in the fourth node. Figure 16 presents the results more vividly by showing barplots of the coefficients for make and zone by node. The relative sizes of the coefficients are fairly consistent between nodes.

Because rate of change of log claim rate with respect to bonus and mileage class depends on the levels of make and zone, the best way to visualize the effects is to draw a contour plot of the fitted model for each combination of make and zone. This is done in Figure 17 for four level combinations, those corresponding to the best and worst levels of make and zone. We see that claim rate is highest when mileage class is 5, bonus class is 1, make is 5, and zone is 1. The lowest claim rates occur for make level 4 and zone level 7, more or less independent of mileage and bonus class.

7 Conclusion

We have given four examples to illustrate the uses of GUIDE for building visualizable regression models. We contend that a model is best understood if it can be visualized. But in order to make effective use of current visualization techniques, namely scatter and contour plots, we will often need to fit models to partitions of a dataset. Otherwise, we simply cannot display a model involving more than two predictor variables in a single 2D graph. The data partitions, of course, should be chosen to build as parsimonious a model

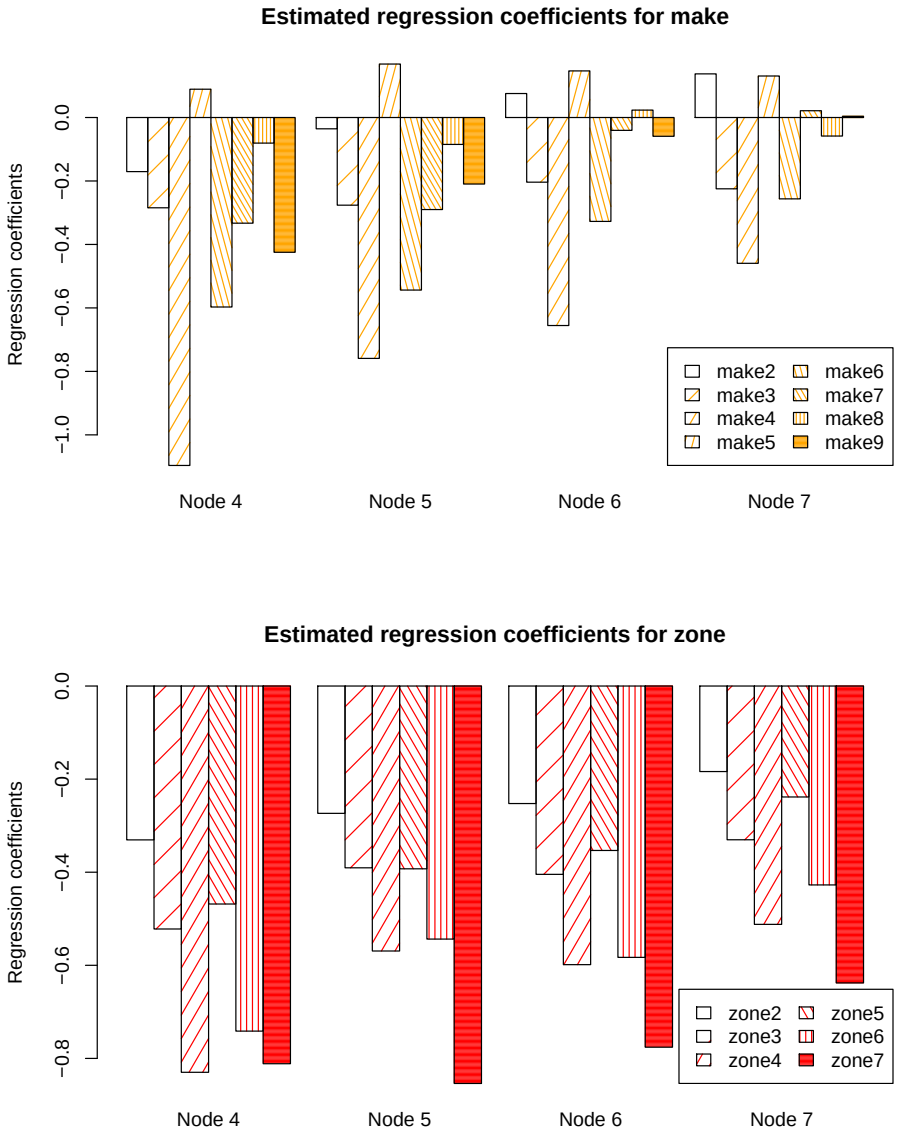

Fig. 16. Estimated regression coefficients for make (above) and zone (below)

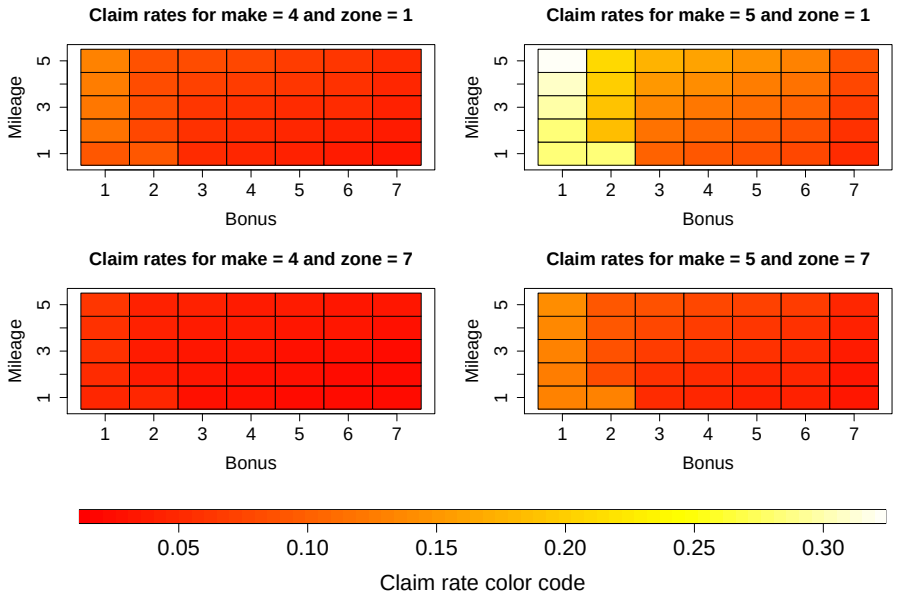

Fig. 17. Estimated claim rates for selected values of make and zone

as possible. The GUIDE algorithm does this by finding partitions that break up curvature and interaction effects. As a result, it avoids splitting a partition on a predictor variable whose effect is already linear. Model parsimony as a whole is ensured by pruning, which prevents the number of partitions from being unnecessarily large.

After pruning is finished, we can be quite confident that most of the important effects of the predictor variables are confined within the one or two selected linear predictors. Thus it is safe to plot the data and fitted function in each partition and to draw conclusions from them. As our examples showed, such plots usually can tell us much more about the data than a collection of regression coefficients. An obvious advantage of 2D plots is that they require no special training for interpretation. In particular, the goodness of fit of the model in each partition can be simply judged by eye instead of through a numerical quantity such as AIC.

The GUIDE computer program is available for Linux, Macintosh, and Windows computers from `www.stat.wisc.edu/%7Eloh/`.

Acknowledgments

The author is grateful to a referee for comments that led to improvements in the presentation. This research was partially supported by grants from the U.S. Army Research Office and the National Science Foundation.

References

1. H. Akaike. Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csàki, editors, *Second International Symposium on Information Theory*, pages 267–281, Budapest, 1973. Akademia Kiadó.
2. D. F. Andrews and A. M. Herzberg. *Data: A Collection of Problems from Many Fields for the Student and Research Worker*. Springer, New York, 1985.
3. D. A. Belsley, E. Kuh, and R. E. Welsch. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Wiley, New York, 1980.
4. L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth, Belmont, 1984.
5. M. Camden. The data bundle. New Zealand Statistical Association, Wellington, 1989.
6. D. Cook. *Regression Graphics: Ideas for Studying Regression Through Graphics*. Wiley, New York, 1998.
7. L. Friedman and M. Wall. Graphical views of suppression and multicollinearity in multiple linear regression. *American Statistician*, 59:127–136, 2005.
8. O. W. Gilley and R. Kelley Pace. On the Harrison and Rubinfeld data. *Journal of Environmental Economics and Management*, 31:403–405, 1996.
9. M. Hallin and J.-F. Ingenbleek. The Swedish automobile portfolio in 1977. A statistical study. *Scandinavian Actuarial Journal*, pages 49–64, 1983.
10. D. Harrison and D. L. Rubinfeld. Hedonic prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5:81–102, 1978.
11. K.-C. Li. Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association*, 86:316–342, 1991.
12. W.-Y. Loh. Regression trees with unbiased variable selection and interaction detection. *Statistica Sinica*, 12:361–386, 2002.
13. A. Miller. *Subset Selection in Regression*. Chapman & Hall, 2nd edition, 2002.