Chapter 1

Logistic Regression Tree Analysis

Ordinary logistic regression (OLR) models the probability of a binary outcome. A logistic regression tree (LRT) is a machine learning method that partitions the data and fits an OLR model in each partition. This chapter motivates LRT by highlighting the challenges of OLR with respect to model selection, interpretation, and visualization on a completely observed data set. Being nonparametric, a LRT model typically has higher prediction accuracy than OLR for large data sets. Further, by sharing model complexity between the tree structure and the OLR node models, the latter can be made simple for easier interpretation and visualization.

OLR is more challenging if there are missing values in the predictor variables, because imputation must be carried out first. The second part of the chapter reviews the GUIDE method of constructing LRT models. A strength of GUIDE is its ability to deal with large numbers of variables and without the need to impute missing values. This is demonstrated on a vehicle crash test dataset for which imputation is difficult due to missing values and other problems.

`Key words and phrases:` classification and regression trees; imputation; logistic regression; machine learning; mising data; visualization

1.1 Introduction

Ordinary logistic regression (OLR) is a technique for modeling the probability of a binary outcome in terms of one or more predictor variables. Consider, for example, a data set on tree damage during a severe thunderstorm over 477,000 acres of the Boundary Waters Canoe Area Wilderness in northeastern Minnesota in July 4, 1999 (R package `alr4` [1]). Observations from 3666 trees were collected, including for each tree, whether it was blown down ($Y = 1$) or not ($Y = 0$), its trunk diameter D in centimeters, its species S, and the local

intensity L of the storm, as measured by the fraction of damaged trees in its vicinity.

Let $p = P(Y = 1)$ denote the probability that a tree is blown down. OLR approximates the logit function $\text{logit}(p) = \log(p/(1-p))$ as a function of the predictor variables linear in any unknown parameters. A *simple linear* OLR model has the form $\text{logit}(p) = \log(p/(1-p)) = \beta_0 + \beta_1 X$, where X is the only predictor variable. Solving for p yields the p-function

$$p = \frac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)} = \frac{1}{1 + \exp(-\beta_0 - \beta_1 X)}.$$

In general, if there are k predictor variables, $X_1, \ldots, X_k$, a *multiple linear* OLR model has the form $\text{logit}(p) = \beta_0 + \sum_{j=1}^{k} \beta_j X_j$. The parameters $\beta_0, \beta_1, \ldots, \beta_k$ are typically estimated by maximizing the likelihood function. Let n denote the sample size and let $(x_{i1}, \ldots, x_{ik}, y_i)$ denote the values of $(X_1, \ldots, X_k, Y)$ for the ith observation $(i = 1, \ldots, n)$. Treating each y_i as the outcome of an independent Bernoulli random variable with success probability p_i, the likelihood function is

$$\prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i} = \frac{\exp\{\sum_i y_i(\beta_0 + \sum_j \beta_j x_{ij})\}}{\prod_i \{1 + \exp(\beta_0 + \sum_j \beta_j x_{ij})\}}.$$

The maximum likelihood estimates are the values of $(\beta_0, \beta_1, \ldots, \beta_k)$ that maximize this function.

1.2 Fitting OLR models

Fitting a simple linear OLR model to the tree damage data using L yields

$$\text{logit}(p) = -1.999 + 4.407L \tag{1.1}$$

with estimated p-function shown in Fig. 1.1. The equation implies that the stronger the local storm intensity, the higher the chance that a tree is blown down. The boxplots in Figure 1.2 show that the distributions of D are skewed. To reduce the skewness, Cook and Weisberg [2] transformed D to $\log(D)$, and obtained the model

$$\text{logit}(p) = -4.792 + 1.749\log(D) \tag{1.2}$$

which suggests that larger trees are less likely to survive the thunderstorm than narrower ones. If both $\log(D)$ and L are used, the model becomes

$$\text{logit}(p) = -6.677 + 1.763\log(D) + 4.42L. \tag{1.3}$$

The relative stability of the coefficients of L and $\log(D)$ in equations (1.1)–(1.3) is due to the weak correlation of 0.168 between the two variables. If the interaction $L\log(D)$ is included, the model changes to

$$\text{logit}(p) = -4.341 + 0.891\log(D) - 1.482L + 2.235L\log(D) \tag{1.4}$$

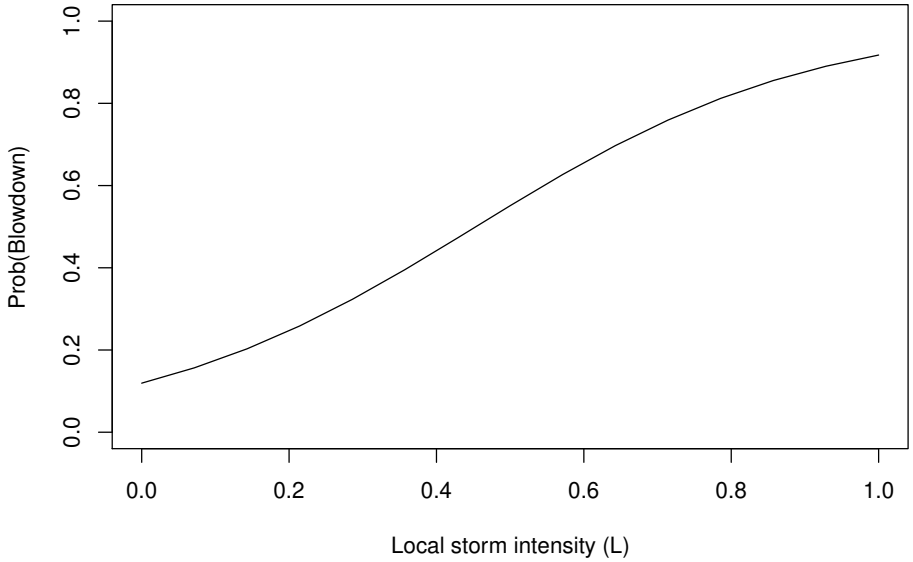

Figure 1.1: Estimated probability of blowdown computed from a simple linear logistic regression model using L as predictor

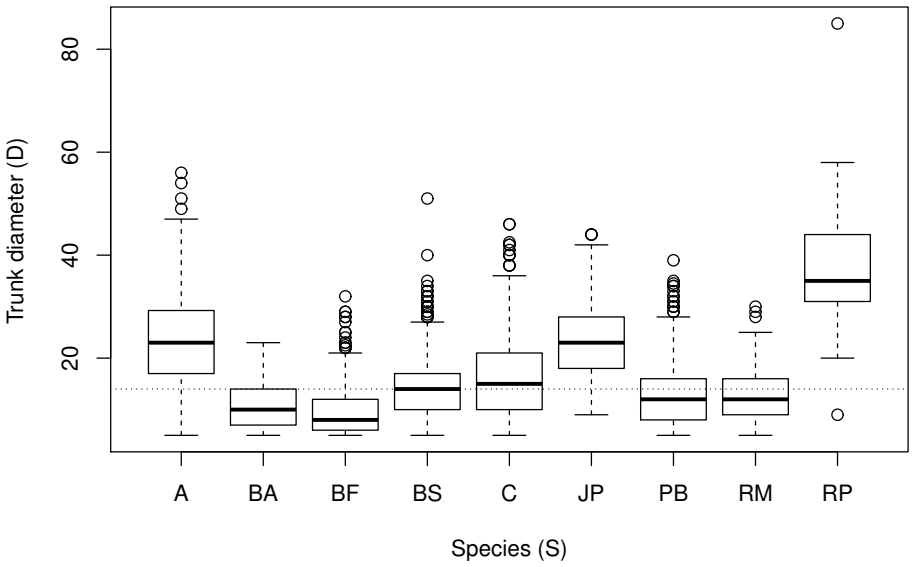

Figure 1.2: Boxplots of trunk diameter D. The dotted line marks the median value of D.

Table 1.1: Indicator variable coding for species variable S

Species	U_1	U_2	U_3	U_4	U_5	U_6	U_7	U_8
A (aspen)	0	0	0	0	0	0	0	0
BA (black ash)	1	0	0	0	0	0	0	0
BF (balsam fir)	0	1	0	0	0	0	0	0
BS (black spruce)	0	0	1	0	0	0	0	0
C (cedar)	0	0	0	1	0	0	0	0
JP (jack pine)	0	0	0	0	1	0	0	0
PB (paper birch)	0	0	0	0	0	1	0	0
RM (red maple)	0	0	0	0	0	0	1	0
RP (red pine)	0	0	0	0	0	0	0	1

and the coefficients of $\log(D)$ and L are changed more dramatically.

So far, species S has been excluded from the models. As in linear regression, a categorical variable having m distinct values may be represented by $(m-1)$ indicator variables, $U_1, \ldots, U_{m-1}$, each taking value 0 or 1. The variables for species are shown in Table 1.1, which uses the "set-to-zero constraint" that sets all the indicator variables to 0 for the first species (aspen). A model that assumes the same slope coefficients for all species but that gives each a different intercept term is

$$\begin{aligned}\text{logit}(p) &= -5.997 + 1.581\log(D) + 4.629L \\ &\quad - 2.243U_1 + 0.0002U_2 + 0.167U_3 - 2.077U_4 \\ &\quad + 1.040U_5 - 1.724U_6 - 1.796U_7 - 0.003U_8. \end{aligned} \tag{1.5}$$

How well do models (1.1)–(1.5) fit the data? One popular way to assess fit is by means of significance tests based on the residual deviance and its degrees of freedom (df)—see, e.g., [3, p. 96] for the definitions. The residual deviance is analogous to the residual sum of squares in linear regression. For model (1.5), the residual deviance is 3259 with 3655 df. We can evaluate the fit of this model by comparing its residual deviance against that of a larger one, such as the 27-parameter model

$$\begin{aligned}\text{logit}(p) &= \beta_0 + \beta_1\log(D) + \beta_2 L + \sum_{j=1}^{8}\gamma_j U_j \\ &\quad + \sum_{j=1}^{8}\beta_{1j}U_j\log(D) + \sum_{j=1}^{8}\beta_{2j}U_j L \end{aligned} \tag{1.6}$$

that allows the coefficients of $\log(D)$ and L to vary with species. It has a residual deviance of 3163 with 3639 df. If model (1.5) fits the data well, the difference between its residual deviance and that of model (1.6) is approximately distributed as a chi-squared random variable with df equal to the difference in df of the two models. The difference in deviance is $3259 - 3163 = 96$, which is improbably large for a chi-squared random variable with $3655 - 3639 = 16$ df.

Rejection of model (1.5) does not necessarily imply that model (1.6) is satisfactory. To find out, it may be compared with a larger one, such as the 28-parameter model

$$\begin{aligned}\text{logit}(p) &= \beta_0 + \beta_1 \log(D) + \beta_2 L + \beta_3 L \log(D) + \sum_{j=1}^{8} \gamma_j U_j \\ &\quad + \sum_{j=1}^{8} \beta_{1j} U_j \log(D) + \sum_{j=1}^{8} \beta_{2j} U_j L \end{aligned} \tag{1.7}$$

that includes an interaction between L and $\log(D)$. This has a residual deviance of 3121 with 3638 df. Therefore model (1.6) is rejected because its residual deviance differs from that of (1.7) by 42 but their dfs differ only by 1. With this procedure, each of models (1.1) through (1.6) is rejected when compared against the next larger model in the sequence.

Another way to select a model employs a function such as AIC, which is residual deviance plus two times the number of estimated parameters. AIC tries to balance deviance against model complexity (see, e.g., [4, p. 234]), but it tends to over-fit the data. That is, AIC often chooses a large model. In this data set, if we apply AIC to the set of all models up to third-order, it chooses the largest, namely, the three-factor interaction model

$$\begin{aligned}\text{logit}(p) &= \beta_0 + \beta_1 \log(D) + \beta_2 L + \sum_{j=1}^{8} \gamma_j U_j \\ &\quad + \beta_3 L \log(D) + \sum_{j=1}^{8} \beta_{1j} U_j \log(D) \\ &\quad + \sum_{j=1}^{8} \beta_{2j} U_j L + \sum_{j=1}^{8} \delta_j U_j L \log(D) \end{aligned} \tag{1.8}$$

which has 36 parameters.

Models (1.7) and (1.8) are hard to graph. Plotting the estimated p-function as in Fig. 1.1 is impossible if a model has more than one predictor variable. This problem is exacerbated by the tendency of model complexity increasing with increase in sample size and number of predictors. Interpretation of the estimated coefficients is futile then, as they often change from one model to another, due to multicollinearity among the terms. For example, the coefficient for L is 4.424, -1.482, and 4.629 in models (1.3), (1.4), and (1.5), respectively.

To deal with this problem, [2] used a "partial one-dimensional model" (POD) that employs a linear function of $\log(D)$ and L as predictor variable. They found that if the observations for balsam fir (`BF`) and black spruce (`BS`) are excluded, the model $\text{logit}(p) = \beta_0 + Z + \sum_j \gamma_j U_j$, with $Z = 0.78 \log(D) + 4.1L$, fits the remaining data quite well. Now the estimated p-function can be plotted as shown in Fig. 1.3, but the graph is not as simple to interpret as that in Fig. 1.1

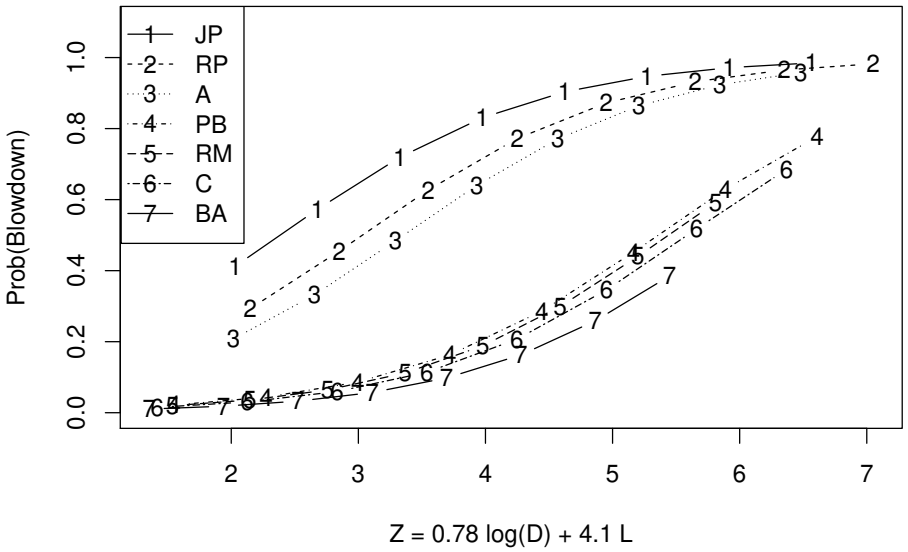

Figure 1.3: Estimated probability of blowdown for seven species, excluding balsam fir (BF) and black spruce (BS), according to model (1.9)

because Z is a linear combination of two variables. To include species BF and BS, [2] settled on the larger model

$$\begin{aligned} \text{logit}(p) &= \beta_0 + Z + \sum_{j=1}^{9} \gamma_j U_j + (\theta_1 I_{\texttt{BF}} + \theta_2 I_{\texttt{BS}}) \log(D) \\ &\quad + (\phi_1 I_{\texttt{BF}} + \phi_2 I_{\texttt{BS}}) L \end{aligned} \tag{1.9}$$

which contains separate coefficients (θ_j, ϕ_j) for BF and BS. Here $I_{(\cdot)}$ denotes the indicator function, i.e., $I_A = 1$ if species is A, and $I_A = 0$ otherwise. The model cannot be displayed graphically for species BF and BS because it is a function of three predictor variables.

1.3 Logistic regression trees

A logistic regression tree (LRT) model is a machine learning solution that simultaneously retains the graphical advantage of simple models and the prediction accuracy of more complex ones. It recursively partitions the data set and fits a simple or multiple linear OLR model in each partition. As a result, the partitions can be displayed as a decision tree [5] such as Figure 1.4, which shows a *simple linear* LRT model fitted to the tree damage data by the GUIDE algorithm [6, 7]. A terminal node represents a partition and an OLR model with a single linear predictor is fitted in each one. Beside each intermediate node is a

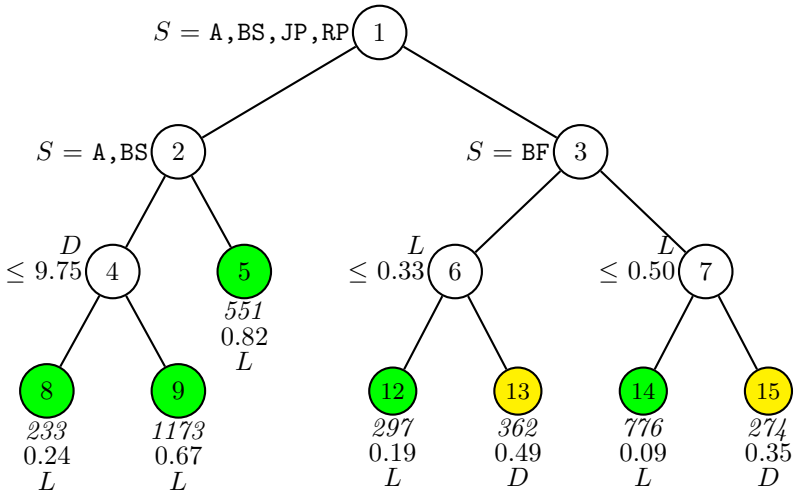

Figure 1.4: GUIDE simple linear LRT model for P(blowdown). At each split, an observation goes to the left branch if and only if the condition is satisfied. Sample size (in italics), proportion of blowdowns, and name of regressor variable are printed beneath each terminal node. Green and yellow terminal nodes have L and D, respectively, as best linear predictor.

condition stating that an observation goes to the left subnode if and only if the condition is satisfied. Below each terminal node are the sample size (in italics), the proportion of blown down trees, and the name of the best linear predictor variable. The split at the root node (labeled "1") sends observations to node 2 if and only if S is `A`, `BS`, `JP`, or `RP`. (Node labels employ the convention that a node with label k has left and right child nodes labeled $2k$ and $2k+1$, respectively.) Node 5, consisting of the `JP` and `RP` species, has the highest proportion of blown down trees at 0.82. Node 9, which consists of species `A` and `BS` trees with diameters greater than 9.75 cm, has the second highest proportion of 0.67. Variable L is the best linear predictor in all terminal nodes except nodes 13 and 15, where D is the best linear predictor. The main advantage in using one linear predictor in each node is that the fitted p-functions can be displayed graphically, as shown in Figure 1.5. It is not necessary to transform D to $\log(D)$ in the LRT.

The LRT model in Fig. 1.4 may be considered a different kind of POD model from that proposed in [2]. Whereas the word "partial" in POD refers to model (1.9) being one dimensional if restricted to certain parts of the data (species in this example), it refers to partitions of the predictor space in a LRT. In addition, whereas "one-dimensional" refers to Z being a linear combination of $\log(D)$ and L in (1.9), the OLR predictor in each node of a LRT is trivially one dimensional because it is an original variable.

GUIDE is a classification and regression tree algorithm with origins in the FACT [8], SUPPORT [9], QUEST [10], CRUISE [11, 12], and LOTUS [13] methods; see [14]. All of them split a data set recursively, choosing a single X

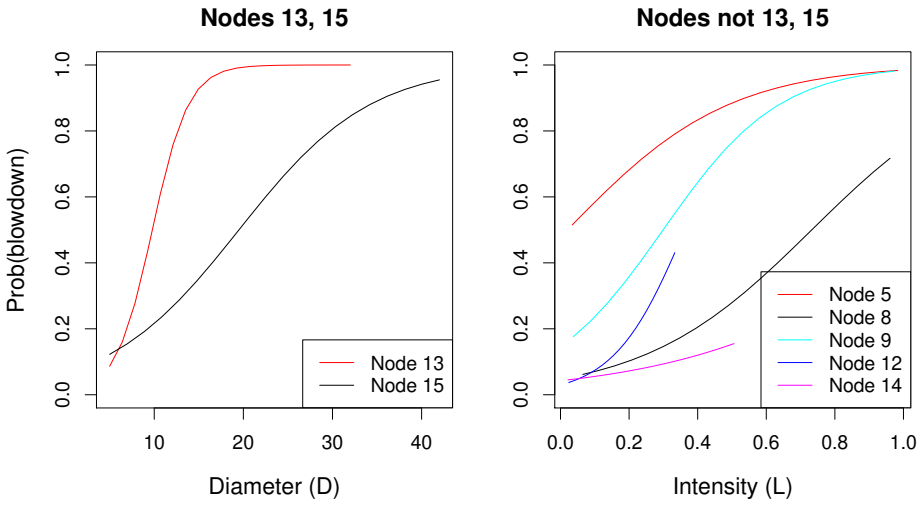

Figure 1.5: Estimated p-functions in terminal nodes of the tree in Fig. 1.4.

variable to split each node. If X is an ordinal variable, the split typically has the form $s = \{X \leq c\}$, where c is a constant. If X is a categorical variable, the split has the form $s = \{X \in \omega\}$, where ω is a subset of the values taken by X. For linear regression trees, algorithms such as AID [15], CART [16] and M5 [17] choose s to minimize the total sum of squared residuals of the regression models fitted to the two data subsets formed by s. Though seemingly innocuous, this approach is flawed as it is biased toward choosing X variables that allow more splits. To see this, suppose that X is an ordinal variable having m distinct values. Then there are $(m-1)$ ways to split the data along the X axis, with each split $s = \{X \leq c\}$ being such that c is the midpoint between two consecutively ordered distinct values of X. This creates a *selection bias* toward X variables with large values of m. In the current example, variable L has 709 unique values but D has only 87. Hence L has eight times as many opportunities as D to split the data. The bias is worse if there are high-level categorical variables, because a categorical variable having m categorical values permits $(2^{m-1} - 1)$ splits of the form $s = \{X \in \omega\}$. For example, variable S permits $(2^{9-1} - 1) = 255$ splits, which is almost three times as many splits as D allows. The earliest warning on the potential for the bias to produce misleading conclusions seems to be [18].

GUIDE avoids the bias by using a two-step approach to split selection. First, it uses significance tests to select the X variable. Then it searches for c or ω for X. For linear regression trees, this is achieved by fitting a linear model to the data in the node and using a contingency table chi-squared test of the association between grouped values of each predictor variable and the signs of the residuals. If X is ordinal, the groups are intervals between certain order statistics. If X is categorical, the groups are the categorical levels. Then the X variable having the smallest chi-squared p-value is selected. Repeating this procedure recursively produces a large binary tree that is pruned to minimize a

cross-validation estimate of prediction mean squared error [16].

Let $\hat{p}(x)$ denote the estimated value of $p(x) = P(Y = 1 \mid X = x)$. The preceding split variable selection method needs modification for logistic regression, because the residual $y - \hat{p}(x)$ is positive if $y = 1$ and negative if $y = 0$, irrespective of the value of $\hat{p}(x)$. Consequently, the residual signs provide no information on the adequacy of $\hat{p}(x)$. A first attempt at a solution was proposed in [19], where the residuals $y - \hat{p}(x)$ are replaced with "pseudo-residuals" $\bar{p}(x) - \hat{p}(x)$, with $\bar{p}(x)$ being a weighted average of the y values in a neighborhood of x. Its weaknesses are sensitivity to choice of weights and neighborhoods, and difficulty in specifying the neighborhoods if the dimension of the predictor space is large or if there are missing values. LOTUS uses a trend-adjusted chi-squared test [20, 21] that effectively replaces $\bar{p}(x)$ with a linear estimate.

For logistic regression, GUIDE uses the average from an ensemble of least-squares GUIDE regression trees (called a "GUIDE forest") to form the pseudo-residuals for variable selection. The main steps are as follows.

1. Fit a least-squares GUIDE forest [22] to the data to obtain a preliminary estimate $\tilde{p}(x)$ of $p(x)$ for each observed x. (Random forest [23] cannot substitute for GUIDE forest if the data contain missing values.)

2. Beginning with the root node, carry out the following steps on the data in each node, stopping only if the number of observations is below a prespecified threshold or if all the values of the predictor variables or the Y values are constant.

   (a) For each X variable to be used in fitting an OLR model in the node, *temporarily* impute its missing values with its node sample mean.

   (b) Fit a simple or multiple linear OLR model to the imputed data in the node. If a simple linear OLR model is desired, fit one to each linear predictor variable in turn and choose the one with smallest residual deviance. Let $\hat{p}(x)$ denote the estimated value of $p(x)$ from the fitted model.

   (c) Revert the imputed values in step (2a) to their original missing state.

   (d) For each ordinal X variable, let $q_1 \le q_2 \le q_3$ denote its sample quartiles at the node and define the categorical variable $V = \sum_{j=1}^{3} I(X > q_j)$. If X is a categorical variable, define $V = X$. Add an extra "missing" category to V if X has missing values.

   (e) Form a contingency table for each X variable using the signs of $\tilde{p}(x) - \hat{p}(x)$ as rows and the values of V as columns. Find the chi-squared statistic χ^2_ν for the test of independence between rows and columns.

   (f) Let $G_\nu(x)$ denote the distribution function of a chi-squared variable with ν df and let $\epsilon = 2 \times 10^{-6}$. Convert each χ^2_ν to its equivalent one-df χ^2_1 value as follows.

      i. If $\epsilon < G_\nu(\chi^2_\nu) < 1 - \epsilon$, define $\chi^2_1 = G_1^{-1}(G_\nu(\chi^2_\nu))$.

ii. Otherwise, to avoid dealing with very small or large p-values, use the following dual application of the Wilson-Hilferty approximation [24]. Define

$$W_1 = \left\{\sqrt{2\chi^2_\nu} - \sqrt{2\nu - 1} + 1\right\}^2 /2$$

$$W_2 = \max\left(0, \left[\frac{7}{9} + \sqrt{\nu}\left\{\left(\frac{\chi^2_\nu}{\nu}\right)^{1/3} - 1 + \frac{2}{9\nu}\right\}\right]^3\right).$$

Approximate the one-df chi-squared value with

$$\chi^2_1 = \begin{cases} W_2 & \text{if } \chi^2_\nu < \nu + 10\sqrt{2\nu} \\ (W_1 + W_2)/2 & \text{if } \chi^2_\nu \geq \nu + 10\sqrt{2\nu} \text{ and } W_2 < \chi^2_\nu \\ W_1 & \text{otherwise.} \end{cases}$$

An earlier one-step approximation is used in [7]. Tables 1.2 and 1.3 show the contingency tables and corresponding chi-squared statistics for `Species`, `Intensity` and `Diameter` at the root node of the tree in Figure 1.4.

(g) Let X^* be the variable with the largest value of χ^2_1 and let `NA` denote the missing value code.

i. If X^* is ordinal, let s be a split of the form $\{X^* = \texttt{NA}\}$, $\{X^* \leq c\} \cup \{X^* = \texttt{NA}\}$, or $\{X^* \leq c\} \cap \{X^* \neq \texttt{NA}\}$.

ii. If X^* is categorical, let s be a split of the form $\{X^* \in \omega\}$, where ω is a proper subset of the values (including `NA`) of X^*.

(h) For each split s, apply steps (2a) and (2b) to the data in the left and right subnodes induced by s and let $d_L(s)$ and $d_R(s)$ be their respective residual deviances.

(i) Select the split s that minimizes $d_L(s) + d_R(s)$.

3. After splitting stops, prune the tree with the CART cost-complexity method [16] to obtain a nested sequence of subtrees.

4. Use the CART cross-validation method to estimate the prediction deviance of each subtree.

5. Select the smallest subtree whose estimated prediction deviance is within a half standard error of the minimum.

Figure 1.6 shows the LOTUS tree for the current data. MOB [25] is another algorithm that can construct a LRT, but for simple linear LRT models, it requires the linear predictor to be pre-specified and to be the same in all terminal nodes. Figure 1.7 shows the MOB tree with L as the common linear predictor. Figure 1.8 compares the values of $\hat{p}(x)$ from a GUIDE forest of 500 trees, model (1.9) and the simple linear GUIDE, LOTUS and MOB LRT models. Although there are clear differences in the values of $\hat{p}(x)$ between GUIDE,

Table 1.2: Chi-squared test for `Species` with Wilson-Hilferty χ^2_1 value

	A	BA	BF	BS	C	JP	PB	RM	RP
$\tilde{p} > \hat{p}$	413	0	239	673	0	501	0	2	47
$\tilde{p} \leq \hat{p}$	23	75	420	297	355	1	497	121	2
$\chi^2_8 = 2125$, $\chi^2_1 = 1942$									

Table 1.3: Chi-squared tests for `Intensity` and `Diameter` with quartile intervals Q_1, Q_2, Q_3, Q_4 and Wilson-Hilferty χ^2_1 values

	Intensity				Diameter			
	Q_1	Q_2	Q_3	Q_4	Q_1	Q_2	Q_3	Q_4
$\tilde{p} > \hat{p}$	327	418	527	603	14	543	637	681
$\tilde{p} \leq \hat{p}$	595	493	390	313	933	424	281	153
	$\chi^2_3 = 195$, $\chi^2_1 = 171$				$\chi^2_3 = 1378$, $\chi^2_1 = 1314$			

LOTUS and MOB, they seem to compare similarly against (1.9) and GUIDE forest. Figure 1.9 shows the corresponding results where LOTUS fits the multiple linear LRT model $\text{logit}(p) = \beta_0 + \beta_1 D + \beta_2 L$, and GUIDE and MOB fit $\text{logit}(p) = \beta_0 + \beta_1 D + \beta_2 L + \sum_{j=1}^{8} \gamma_j U_j$ in each terminal node. (LOTUS does not convert categorical variables to indicator variables to serve as regressors.) The correlations among the $\hat{p}(x)$ values are much higher.

1.4 Missing values and cyclic variables

The U.S. National Highway Traffic Safety Administration has been evaluating vehicle safety by performing crash tests with dummy occupants since 1972 (`ftp://www.nhtsa.dot.gov/ges`). We use data from 3310 crash tests where the test dummy is in the driver's seat to show how GUIDE deals with missing values and cyclic variables. Each test gives the severity of head injury (`HIC`) sustained by the dummy and the values of about 100 variables describing the vehicle, test environment and the test dummy. The response variable is $Y = 1$ if `HIC` $>$ 1000 (threshold for severe head injury), and $Y = 0$ otherwise. About half of the predictor variables are ordinal, six are cyclic, and the rest are categorical.

Three features in the data make model building particularly challenging. The first is missing data. Missing values in categorical variables are not problematic, as they can be assigned a "missing" category. Missing values in other variables, however, need to be imputed before application of OLR. This can be extraordinarily difficult if there are many missing values and the missingness patterns are complex [22, 26]. All ordinal and cyclic variables here have missing values. Table 1.4 gives the names and numbers of missing values of some of them (see [27] for the others). For example, `IMPANG`, the angle between the axis of a vehicle and the axis of another vehicle or barrier, is undefined for a rollover crash test, where there is no barrier and only one vehicle is involved. In such cases, the value of `IMPANG` is recorded as missing and imputing it with a

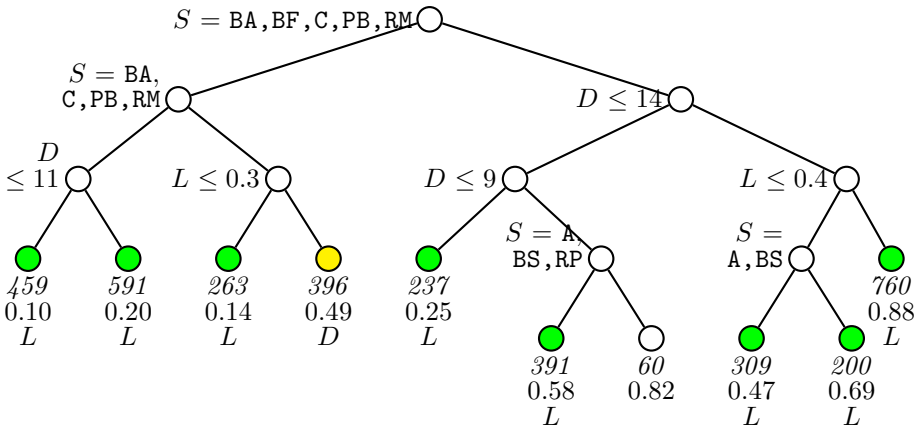

Figure 1.6: LOTUS simple linear LRT model for P(blowdown). At each split, an observation goes to the left branch if and only if the condition is satisfied. Sample size (in italics), proportion of blowdowns, and name of regressor variable (if any) are printed below nodes. Green and yellow terminal nodes have L and D, respectively, as best linear predictor.

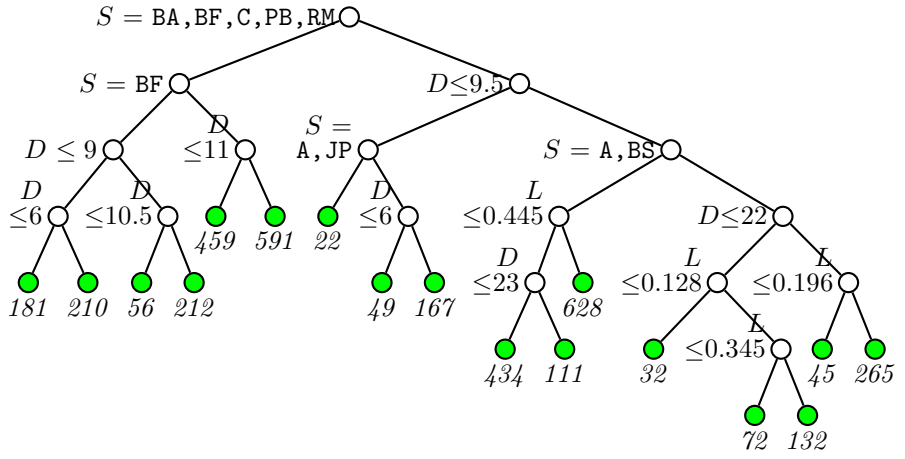

Figure 1.7: MOB simple linear LRT model with L pre-specified as the common linear predictor in all nodes. At each split, an observation goes to the left branch if and only if the condition is satisfied. Sample sizes (in italics) are printed below nodes.

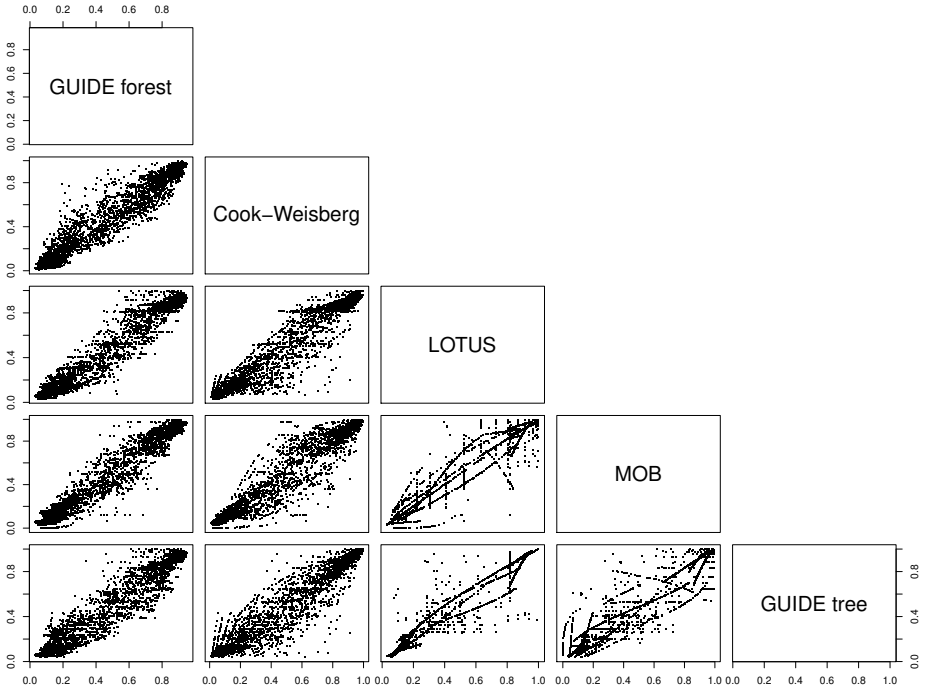

Figure 1.8: Comparison of fitted values $\hat{p}$ of Cook-Weisberg model (1.9) and GUIDE forest versus simple linear LRT models

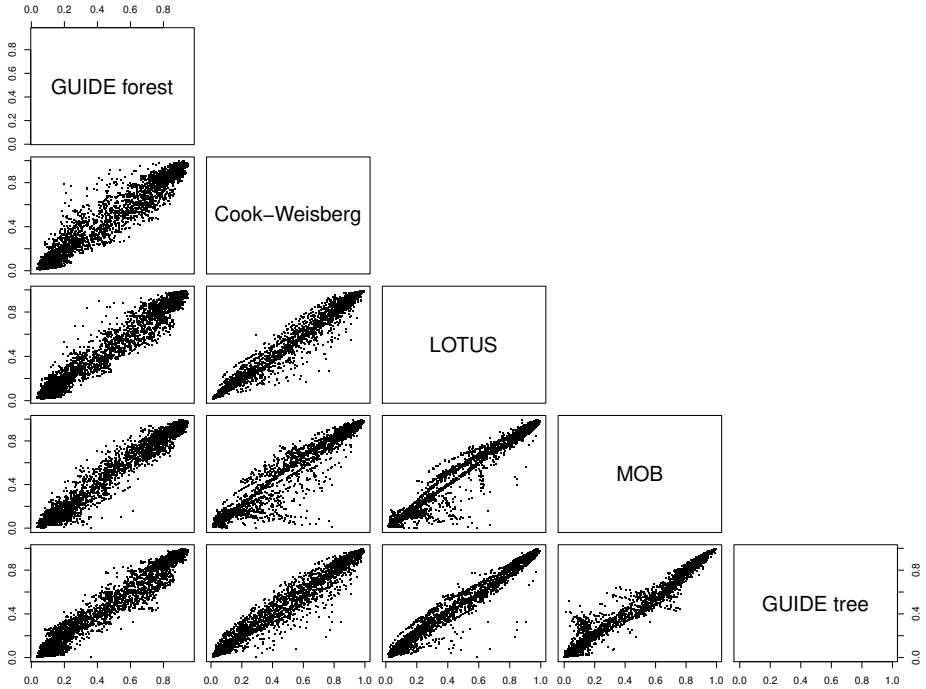

Figure 1.9: Comparison of fitted values $\hat{p}$ of Cook-Weisberg model (1.9) and GUIDE forest versus multiple linear LRT models

Table 1.4: Definitions and numbers of missing values of some predictor variables in the crash-test data

Variable	Description	Missing
BARRIG	rigid or deformable barrier	1
BARSHP	barrier shape (21 types)	0
BX2	rear surface of vehicle to front of engine	288
BX5	rear surface of vehicle to upper leading edge of left door	288
CARANG	angle between surface of rollover test cart and ground	991
COLMEC	steering column collapse mechanism (9 types)	248
ENGDSP	engine displacement	24
IMPANG	angle between axis of vehicle 2 and axis of vehicle 1 or barrier (0 degrees is perpendicular to barrier)	4
CLSSPD	closing speed: relative velocity of approach of two centers of gravity before contact	2
VEHSPD	resultant speed of vehicle before impact	1
VEHTWT	vehicle test weight	4
VEHWID	vehicle width	90
WHLBAS	vehicle or impactor's wheelbase	30
YEAR	vehicle model year	4

number is inappropriate. The situation is worse for variable CARANG, which has 991 missing values. Given that the crash tests are carefully monitored and have been performed for years, it is unlikely for so many observations to be missing by chance.

For split selection, GUIDE sends all missing values in the selected ordinal or cyclic variable either to the left or to the right subnode, depending on which split gives a smaller sum of residual deviances in the two subnodes. Hence no imputation is carried out in this step. To fit an OLR model to a node, GUIDE imputes missing values in the selected predictor variable with its node mean.

A second challenging feature is the presence of cyclic variables that are angles with periods of 360 degrees. These variables are traditionally transformed to sines and cosines but splits on one of them at a time are not as meaningful as splits on the angles themselves. The problem is more difficult if the variable has missing values. Should we impute the angles and then compute the sines and cosines of the imputed values or should we impute the sines and cosines directly? GUIDE avoids imputation entirely by restricting cyclic variables to split the nodes. If a cyclic variable is selected, the split takes the form of a sector "$X \in [\theta_1, \theta_2]$", where θ_1 and θ_2 are angles, and missing values are sent to the left or right subnode in the same fashion as non-cyclic variables.

The third challenging feature is that, apparently by design, high-speed crash tests are more often carried out on deformable barriers and low-speed tests more often on rigid barriers. This is evident from the boxplots of CLSSPD by BARRIG in Fig. 1.10, where half of the tests with deformable barriers are above closing speeds of 60 km/h, but less than one quarter of those with rigid barriers are

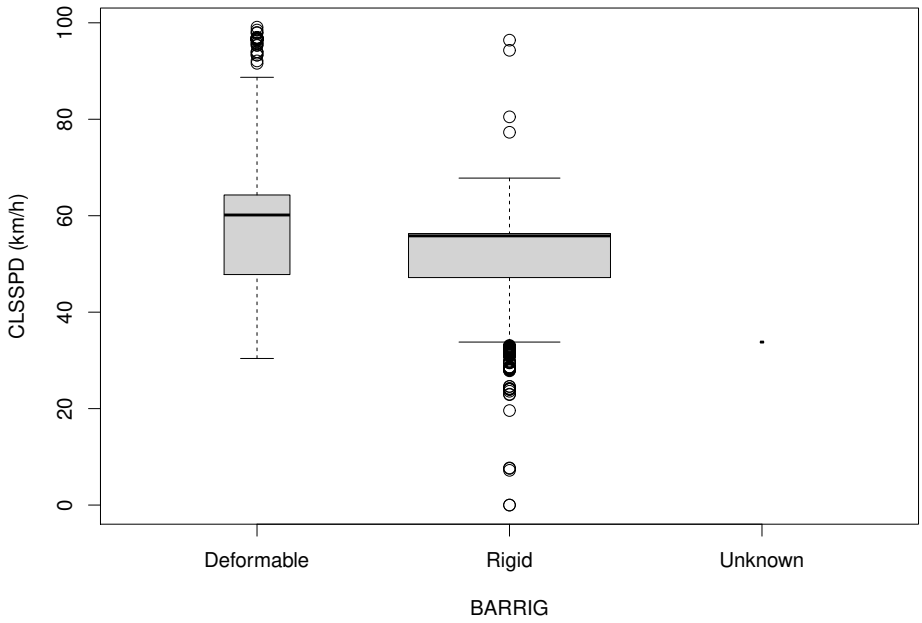

Figure 1.10: Boxplots of closing speed by barrier rigidity for the crash-test data, with box width proportional to square root of sample size

above 60 km/h. Presumably, crashes into rigid barriers are not performed at high speeds because the outcomes are predictable, but this confounds the effects of `CLSSPD` and `BARRIG` in an OLR model.

We say that X is an "`s`" variable if it can be used to *split* the nodes and an "`f`" variable if it can be used to *fit* OLR models in the nodes. To limit the amount of imputation in this example, we restrict ordinal variables with more than 20 percent missing values to serve as `s` variables only. Cyclic and categorical variables are also restricted to splitting nodes.

Figure 1.11 shows the LRT where a simple linear OLR model is fitted in each node. The root node is split on `COLMEC`, which is steering wheel collapse mechanism. Observations with `COLMEC` equal to `BWU` (behind wheel unit), `EMB` (embedded ball), `EXA` (extruded absorber), `NON` (none), or `OTH` (other) go to node 2. Otherwise, if `COLMEC` is `CON` (convoluted tube), `CYL` (cylindrical mesh tube), `NAP` (not applicable), `UNK` (unknown), or missing, observations go to node 3. At node 2, observations go to node 4 if `BX2` $\leq$ 3496.5 or missing (the asterisk beside the inequality sign in the figure indicates that missing values go to the left node). At node 3, observations go to node 6 if `BARSHP` is `LCB` (load cell barrier), `POL` (pole), `US2`, or `US3` (different barrier types). Node 6 is split on impact angle `IMPANG`, where 0 degrees indicates impact is head-on. If an observation has `IMPANG` between 284 and 286 degrees inclusive (i.e., driver-side), it goes to node 12. The two-degree range may seem narrow, but there are 67 observations in the node, suggesting that the tests were by design. Below each

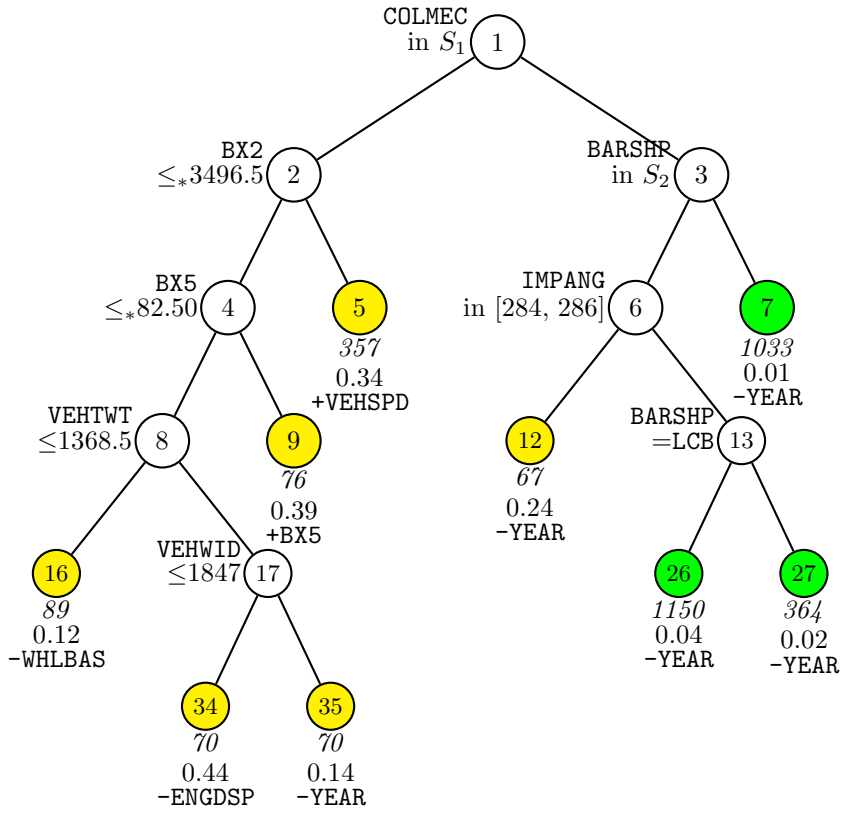

Figure 1.11: GUIDE piecewise simple linear LRT for crash-test data. At each split, an observation goes to the left branch if and only if the condition is satisfied. The symbol '$\leq_*$' stands for '$\leq$ or missing'. Set S_1 = {BWU, EMB, EXA, NON, OTH}. Set S_2 = {LCB, POL, US2, US3}. Sample size (*in italics*), proportion of cases with $Y = 1$, and sign and name of regressor variable printed below nodes. Terminal nodes with proportions of $Y = 1$ above and below value of 0.08 at root node are colored yellow and green, respectively.

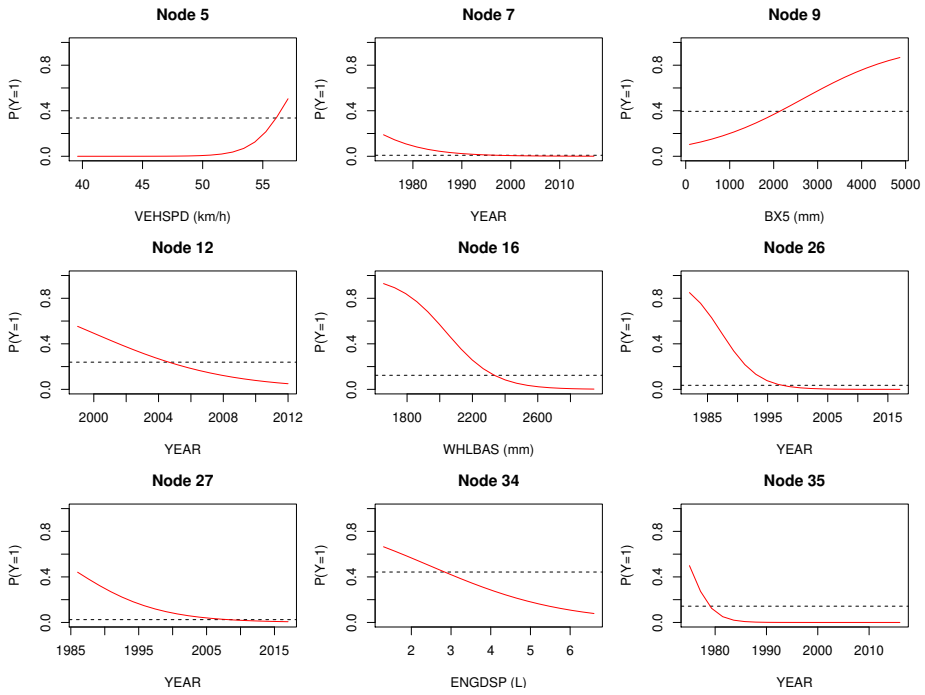

Figure 1.12: Fitted logistic regression curves in terminal nodes of Fig. 1.11; horizontal dotted lines indicate proportion of severe injury in the node

terminal node are the sample size (in *italics*), proportion of $Y = 1$, and the selected OLR predictor variable, with the sign of its estimated coefficient.

The tree shows that nodes 5, 9, and 34 have the highest proportions of severe head injury, at 34, 39, and 44 percent, respectively. Vehicles in these nodes have certain steering wheel collapse mechanisms and they tend to be longer (`BX2` $>$ 3496.5 or `BX5` $>$ 82.5) or are heavy (`VEHTWT` $>$ 1368.5) and narrow (`VEHWID` $\leq$ 1846). Figure 1.12 shows the fitted logistic regression curves in the terminal nodes. The proportion of tests with severe head injury is indicated by a dotted line in each plot.

1.5 Conclusion

Logistic regression is a technique for estimating the probability of an event in terms of the values of one or more predictor variables. If there are missing values among the predictor variables, they need to be imputed first. Otherwise, the observations or variables containing the missing values would need to be deleted. Neither solution is attractive. In practice, finding a logistic regression model with good prediction accuracy is seldom automatic; it usually requires trial-and-error selection of variables, choice of transformations, and estimation of the

accuracy of numerous models. Even when a model with good estimated accuracy is found, interpretation of the regression coefficients is not straightforward if there are two or more predictor variables.

A logistic regression tree is a piecewise logistic regression model, with the pieces obtained by recursively partitioning the space of predictor variables. Consequently, if there is no over-fitting, it may be expected to possess higher prediction accuracy than a one-piece logistic regression model. Recursive partitioning has two advantages over a search of all partitions: it is computationally efficient and it allows the partitions to be displayed as a decision tree. At a minimum, a logistic regression tree can serve as an informal goodness of fit test of whether a one-piece logistic model is adequate for the whole sample. A non-trivial pruned tree would indicate that a one-piece logistic model has lower prediction accuracy, possibly due to unaccounted interactions or nonlinearities among the variables. Ideally, an effective tree-growing and pruning algorithm would automatically account for the overlooked effects, making it unnecessary to specify interaction and higher-order terms. It would also allow the models in the terminal nodes to be as simple as desired (such as fitting a single linear predictor in each node).

Tree pruning is very important for prediction accuracy. Many methods adopt the AIC-type approach of selecting the tree that minimizes the sum of the residual deviance and a multiple, K, of the number of terminal nodes. There being no value of K that works for all data sets [16], the advantage of this approach is mainly computational speed. Our experience indicates that it is inferior to a pruning approach that uses cross-validation to estimate prediction accuracy.

Despite a binary decision tree being intuitive to interpret, a poor split selection method can yield misleading conclusions. A common cause is selection bias. The greedy approach used by CART and many other algorithms is known to prefer variables that permit more splits of the data. Consequently, it is hard to know if a variable is chosen due to its predictive power or because it has more ways to partition the data. LOTUS and GUIDE avoid the bias by selecting variables with chi-squared tests. At the time of completion of this article, GUIDE is the only tree algorithm that can deal with cyclic variables and with two or more missing value codes [22]. The GUIDE software and manual may be obtained from `www.stat.wisc.edu/~loh/guide.html`.

1.6 Acknowledgment

Part of this work was done in the summer of 2019 during the author's visit to the National Chung Cheng and National Tsing Hua Universities, Taiwan, under the auspices of the National Center for Theoretical Sciences.

Bibliography

[1] S. Weisberg: *Applied Linear Regression*, Fourth edn. (Wiley, Hoboken NJ 2014)

[2] R. D. Cook, S. Weisberg: Partial one-dimensional regression models, American Statistician **58**, 110–116 (2004)

[3] A. Agresti: *An Introduction to Categorical Data Analysis* (Wiley, New York 1996)

[4] J. M. Chambers, T. J. Hastie: *Statistical Models in S* (Wadsworth, Pacific Grove 1992)

[5] W.-Y. Loh: Classification and regression trees, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **1**, 14–23 (2011) doi:10.1002/widm.8

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

[7] W.-Y. Loh: Improving the precision of classification trees, Annals of Applied Statistics **3**, 1710–1737 (2009)

[8] W.-Y. Loh, N. Vanichsetakul: Tree-structured classification via generalized discriminant analysis (with discussion), Journal of the American Statistical Association **83**, 715–728 (1988)

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

[10] W.-Y. Loh, Y.-S. Shih: Split selection methods for classification trees, Statistica Sinica **7**, 815–840 (1997)

[11] H. Kim, W.-Y. Loh: Classification trees with unbiased multiway splits, Journal of the American Statistical Association **96**, 589–604 (2001)

[12] H. Kim, W.-Y. Loh: Classification trees with bivariate linear discriminant node models, Journal of Computational and Graphical Statistics **12**, 512–530 (2003)

[13] K.-Y. Chan, W.-Y. Loh: LOTUS: An algorithm for building accurate and comprehensible logistic regression trees, Journal of Computational and Graphical Statistics **13**, 826–852 (2004)

[14] W.-Y. Loh: Fifty years of classification and regression trees (with discussion), International Statistical Review **34**, 329–370 (2014)

[15] J. N. Morgan, J. A. Sonquist: Problems in the analysis of survey data, and a proposal, Journal of the American Statistical Association **58**, 415–434 (1963)

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

[17] J. R. Quinlan: Learning with continuous classes. In: *Proceedings of AI'92 Australian National Conference on Artificial Intelligence* (World Scientific, Singapore 1992) pp. 343–348

[18] P. Doyle: The use of Automatic Interaction Detector and similar search procedures, Operational Research Quarterly **24**, 465–467 (1973)

[19] P. Chaudhuri, W.-D. Lo, W.-Y. Loh, C.-C. Yang: Generalized Regression Trees, Statistica Sinica **5**, 641–666 (1995)

[20] W. G. Cochran: Some methods of strengthening the common χ^2 tests, Biometrics **10**, 417–451 (1954)

[21] P. Armitage: Tests for linear trends in proportions and frequencies, Biometrics **11**, 375–386 (1955)

[22] W.-Y. Loh, J. Eltinge, M. J. Cho, Y. Li: Classification and regression trees and forests for incomplete data from sample surveys, Statistica Sinica **29**, 431–453 (2019)

[23] L. Breiman: Random Forests, Machine Learning **45**(1), 5–32 (2001)

[24] E. B. Wilson, M. M. Hilferty: The distribution of chi-square, Proceedings of the National Academy of Sciences of the United States of America **17**, 684–688 (1931)

[25] A. Zeileis, T. Hothorn, K. Hornik: Model-based recursive partitioning, Journal of Computational and Graphical Statistics **17**, 492–514 (2008)

[26] W.-Y. Loh, Q. Zhang, W. Zhang, P. Zhou: Missing data, imputation and regression trees, Statistica Sinica **30**, 1697–1722 (2020)

[27] NHTSA: Test Reference Guide Version 5, `https://one.nhtsa.gov/Research/Databases-and-Software/NHTSA-Test-Reference-Guides`, **1** (2014)