29. Logistic Regression Tree Analysis

This chapter describes a tree-structured extension and generalization of the logistic regression method for fitting models to a binary-valued response variable. The technique overcomes a significant disadvantage of logistic regression viz. the interpretability of the model in the face of multi-collinearity and Simpson's paradox. Section 29.1 summarizes the statistical theory underlying the logistic regression model and the estimation of its parameters. Section 29.2 reviews two standard approaches to model selection for logistic regression, namely, model deviance relative to its degrees of freedom and the Akaike information criterion (AIC) criterion. A dataset on tree damage during a severe thunderstorm is used to compare the approaches and to highlight their weaknesses. A recently published partial one-dimensional model that addresses some of the weaknesses is also reviewed.

Section 29.3 introduces the idea of a logistic regression tree model. The latter consists of a binary tree in which a simple linear logistic regression (i.e., a linear logistic regression using a single predictor variable) is fitted to each leaf node. A split at an intermediate node is characterized by a subset of values taken by a (possibly different) predictor variable. The objective is to partition the dataset into rectangular pieces according to the values of the predictor variables such that a simple linear logistic regression model adequately fits the data in each piece. Because the tree structure and the piecewise models can be presented graphically, the whole model can be easily understood. This is illustrated with the thunderstorm dataset using the LOTUS algorithm.

Section 29.4 describes the basic elements of the LOTUS algorithm, which is based on recursive partitioning and cost-complexity pruning. A key feature of the algorithm is a correction for bias in variable selection at the splits of the tree. Without bias correction, the splits can yield incorrect inferences. Section 29.5 shows an application of LOTUS to a dataset on automobile crash tests involving dummies. This dataset is challenging because of its large size, its mix of ordered and unordered variables, and its large number of missing values. It also provides a demonstration of Simpson's paradox. The chapter concludes with some remarks in Sect. 29.5.

Logistic regression is a technique for modeling the probability of an event in terms of suitable explanatory or predictor variables. For example, [29.1] use it to model the probability that a tree in a forest is blown down during an unusually severe thunderstorm that occurred on July 4, 1999, and caused great damage over 477 000 acres of the Boundary Waters Canoe Area Wilderness in northeastern Minnesota. Data from a sample of 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. The dataset may be obtained from www.stat.umn.edu/~sandy/pod.

Let $p = \Pr(Y = 1)$ denote the probability that a tree is blown down. In *linear logistic regression*, we model $\log[p/(1-p)]$ as a function of the predictor variables, with the requirement that it be linear in any unknown parameters. The function $\log[p/(1-p)]$ is also often written as $\text{logit}(p)$. If we use a single predictor such

as L, we have the *simple linear* logistic regression model

$$\text{logit}(p) = \log[p/(1-p)] = \beta_0 + \beta_1 L \qquad (29.1)$$

which can be re-expressed in terms of p as $p = \exp(\beta_0 + \beta_1 L)/[1 + \exp(\beta_0 + \beta_1 L)]$.

In general, given k predictor variables $X_1, \ldots, X_k$, a linear logistic regression model in these variables is $\text{logit}(p) = \beta_0 + \sum_{j=1}^{k} \beta_j X_j$. The parameters $\beta_0, \beta_1, \ldots, \beta_k$ are typically estimated using maximum likelihood theory. 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, we have the likelihood function

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

The *maximum likelihood estimates* (MLEs) $(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k)$ are the values of $(\beta_0, \beta_1, \ldots, \beta_k)$ that maximize this function. Newton–Raphson iteration is usually needed to compute the MLEs.

29.1 Approaches to Model Fitting

The result of fitting model (29.1) is

$$\text{logit}(p) = -1.999 + 4.407L. \qquad (29.2)$$

Figure 29.1 shows a plot of the estimated p function. Clearly, the stronger the local storm intensity, the higher the chance for a tree to be blown down.

Figure 29.2 shows boxplots of D by species. Because of the skewness of the distributions, we follow [29.1] and use $\log(D)$, the natural logarithm of D, in our analysis. With $\log(D)$ in place of L, the fitted model becomes

$$\text{logit}(p) = -4.792 + 1.749\log(D) \qquad (29.3)$$

suggesting that tall trees are less likely to survive a storm than short ones. If we use both $\log(D)$ and L, we obtain the model

$$\text{logit}(p) = -6.677 + 1.763\log(D) + 4.420L . \qquad (29.4)$$

Finally, if we include the product $L\log(D)$ to account for interactions between D and L, we obtain

$$\text{logit}(p) = -4.341 + 0.891\log(D) - 1.482L + 2.235L\log(D) . \qquad (29.5)$$

So far, we have ignored the species S of each tree in our sample. We might get a model with higher prediction accuracy if we include S. As with least-squares regression, we can include a categorical variable that takes m distinct values by first defining $m-1$ indicator variables, $U_1, \ldots, U_{m-1}$, each taking the value 0 or 1. The definitions of the indicator variables corresponding

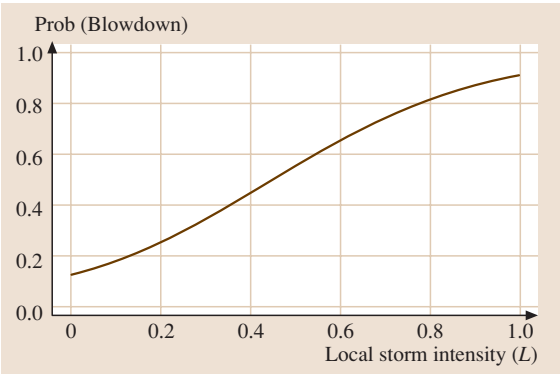

Fig. 29.1 Estimated probability of blowdown computed from a simple linear logistic regression model using L as predictor

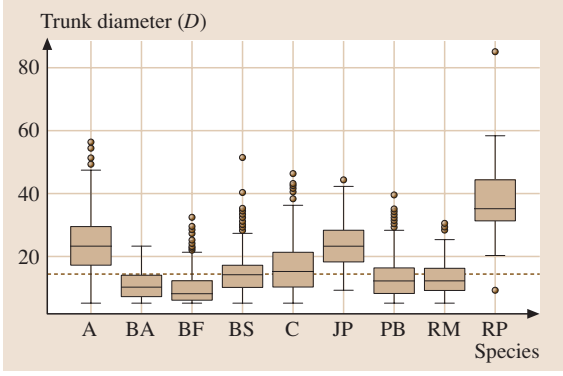

Fig. 29.2 Boxplots of trunk diameter D. The median value of 14 for D, or 2.64 for $\log(D)$, is marked with a *dotted line*

to our nine-species variable S are shown in Table 29.1. Note that we use the *set-to-zero constraint*, which sets all the indicator variables to 0 for the first category (aspen). A model that assumes the same slope coefficients for all species but that gives each a different intercept term is

$$\begin{aligned}\text{logit}(p) = &-5.997 + 1.581\log(D) + 4.629L \\ &- 2.243U_1 + 0.0002U_2 + 0.167U_3 \\ &- 2.077U_4 + 1.040U_5 - 1.724U_6 \\ &- 1.796U_7 - 0.003U_8\,. \end{aligned} \tag{29.6}$$

How well do the models (29.2–29.6) fit the data? One popular method of assessing fit is by means of significance tests based on the model deviance and its *degrees of freedom* (DF)—see, e.g., [29.2] for the definitions. The deviance is analogous to the residual sum of squares in least-squares regression. For the model (29.6), the deviance is 3259 with 3655 DF. We can evaluate the fit of this model by comparing its deviance against that of a larger one, such as the 27-parameter model

$$\begin{aligned}\text{logit}(p) = &\beta_0 + \beta_1\log(D) + \beta_2 L + \sum_{j=1}^{8}\gamma_j U_j \\ &+ \sum_{j=1}^{8}\beta_{1j}U_j\log(D) + \sum_{j=1}^{8}\beta_{2j}U_j L \end{aligned} \tag{29.7}$$

which allows the slope coefficients for $\log(D)$ and L to vary across species. Model (29.7) has a deviance of 3163 with 3639 DF. If the model (29.6) provides a suitable fit to the data, statistical theory says that the difference in deviance should be approximately distributed as a chi-square random variable with DF equal to the difference in the DF of the two models. For our example, the difference in deviance of $3259 - 3163 = 96$ is much too large to be explained by a chi-square distribution with $3655 - 3639 = 16$ DF.

Rejection of model (29.6) does not necessarily imply, however, that the model (29.7) is satisfactory. To find out, we need to compare it with a larger model, such as the 28-parameter model

$$\begin{aligned}\text{logit}(p) = &\beta_0 + \beta_1\log(D) + \beta_2 L + \beta_3 L\log(D) \\ &+ \sum_{j=1}^{8}\gamma_j U_j + \sum_{j=1}^{8}\beta_{1j}U_j\log(D) \\ &+ \sum_{j=1}^{8}\beta_{2j}U_j L \end{aligned} \tag{29.8}$$

which includes an interaction between $\log(D)$ and L. This has a deviance of 3121 with 3638 DF. Model (29.7) is therefore rejected because its deviance differs from that of (29.8) by 42 but their DFs differ only by 1. It turns out that, using this procedure, each of models (29.2–29.7) is rejected when compared against the next larger model in the set.

A second approach chooses a model from a given set by minimizing some criterion that balances model fit with model complexity. One such is the AIC criterion, defined as the deviance plus twice the number of estimated parameters [29.3]. It is well known, however, that the AIC criterion tends to overfit the data. That is, it often chooses a large model. For example, if we apply it to the set of all models up to third order for the current data, it chooses the largest, i. e., the 36-parameter model

$$\begin{aligned}\text{logit}(p) = &\beta_0 + \beta_1\log(D) + \beta_2 L + \sum_{j=1}^{8}\gamma_j U_j \\ &+ \beta_3 L\log(D) + \sum_{j=1}^{8}\beta_{1j}U_j\log(D) \\ &+ \sum_{j=1}^{8}\beta_{2j}U_j L + \sum_{j=1}^{8}\delta_j U_j L\log(D)\,. \end{aligned} \tag{29.9}$$

Table 29.1 Indicator variable coding for the 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

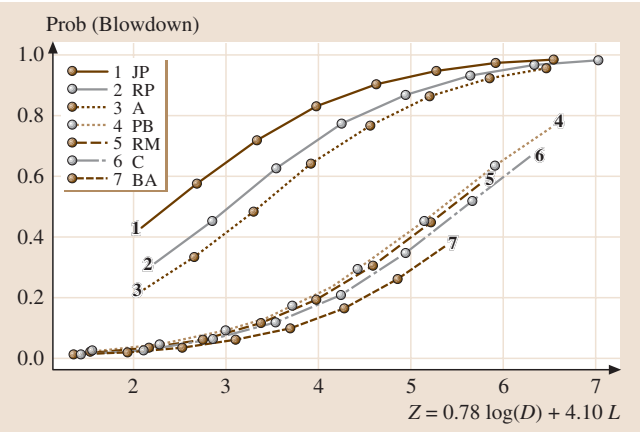

Fig. 29.3 Estimated probability of blowdown for seven speciesTable 29.1, excluding balsam fir (BF) and black spruce (BS), according to model (29.10)

Graphical interpretation of models (29.8) and (29.9) is impossible. The simple and intuitive solution of viewing the estimated p-function by a graph such as Fig. 29.1 is unavailable when a model involves more than one predictor variable. This problem is exacerbated by the fact that model complexity typically increases with increasing sample size or number of predictors. Interpretation of the estimated coefficients is frequently futile, because the estimates typically do not remain the same from one model to another. For example, the coefficient for L is 4.407, 4.424, 1.870, and 4.632 in models (29.2), (29.4), (29.5), and (29.6), respectively. This is due to multi-collinearity among the predictor variables.

Cook and *Weisberg* [29.1] try to solve the problem of interpretation by using a *partial one-dimensional* (POD) model, which employs a single linear combination of the noncategorical variables, $Z = \delta_1 \log(D) + \delta_2 L$, as predictor. For the tree data, they find that if 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 other species quite well. One advantage of this model is that the estimated p-functions may be displayed graphically, as shown in Fig. 29.3. The graph is not as natural as Fig. 29.1, however, because Z is a linear combination of two variables. In order to include the species BF and BS, [29.1] choose the larger model

$$\begin{aligned}\text{logit}(p) = \beta_0 + Z + \sum_{j=1}^{9} \gamma_j U_j \\ + (\theta_1 I_{\text{BF}} + \theta_2 I_{\text{BS}}) \log(D) \\ + (\phi_1 I_{\text{BF}} + \phi_2 I_{\text{BS}}) L \end{aligned} \quad (29.10)$$

which contains separate coefficients, θ_j and ϕ_j, for BF and BS. Here $I_{(\cdot)}$ denotes the indicator function, i. e., $I_A = 1$ if the species is A, and $I_A = 0$ otherwise. Of course, this model does not allow a graphical representation for BF and BS.

29.2 Logistic Regression Trees

The type of models and the method of selection described in the previous section are clearly not totally satisfactory. As the sample size or the number of predictor variables increases, so typically does model complexity. But a more complex model is always harder to interpret than a simple one. On the other hand, an overly simple model may have little predictive power.

A logistic regression tree model offers one way to retain simultaneously the graphical interpretability of simple models and the predictive accuracy of richer ones. Its underlying motivation is that of divide and conquer. That is, a complex set of data is divided into sufficiently many subsets such that a simple linear logistic regression model adequately fits the data in each subset. Data subsetting is performed recursively, with the sample split on one variable at a time. This results in the partitions being representable as a binary decision tree. The method is implemented by [29.4] in a computer program called LOTUS.

Figure 29.4 shows a LOTUS model fitted to the tree data. The data are divided into ten subsets, labeled 0–9. Balsam fir (BF), one of the two species excluded from the [29.1] model, is isolated in subsets 0 and 9, where $\log(D)$ is the best linear predictor. The estimated p-functions for these two subsets are shown in Fig. 29.5. The estimated p-functions for the trees that are not balsam firs can be displayed together in one graph, as shown in Fig. 29.6, because they all employ L as the best simple linear predictor.

From the graphs, we can draw the following conclusions:

1. The probability of blowdown consistently increases with D and L, although the value and rate of increase are species-dependent.

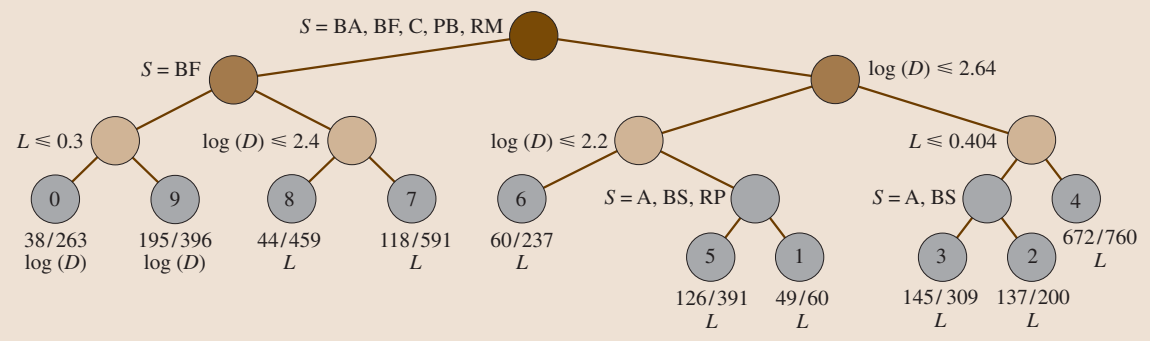

Fig. 29.4 A piecewise simple linear LOTUS model for estimating the probability that a tree is blown down. A splitting rule is given beside each intermediate node. If a case satisfies the rule, it goes to the left child node; otherwise the right child node. The second level split at $\log(D) = 2.64$ corresponds to the median value of D. Beneath each leaf node are the ratio of cases with $Y = 1$ to the node sample size and the name of the selected predictor variable

2. Balsam fir (BF) has the highest chance of blowdown, given any values of D and L.
3. The eight species excluding the balsam fir can be divided into two groups. Group I consists of black ash (BA), cedar (C), paper birch (PB), and red maple (RM). They belong to subsets 7 and 8, and are most likely to survive. This is consistent with the POD model of [29.1]. Group II contains aspen (A), black spruce (BS), jack pine (JP), and red pine (RP).
4. The closeness of the estimated p-functions for subsets 6 and 7 show that the smaller group II trees and the larger group I trees have similar blowdown probabilities for any given value of L.
5. Although aspen (A) and black spruce (BS) are always grouped together, namely, in subsets 3–6, less than 15% of the aspen trees are in subsets 5 and 6. Similarly, only 2% of the red pines (RP) are in these two subsets. Hence the p-function of aspen (A) is mainly described by the curves for subsets 3 and 4, and that for red pine (RP) by the curves for

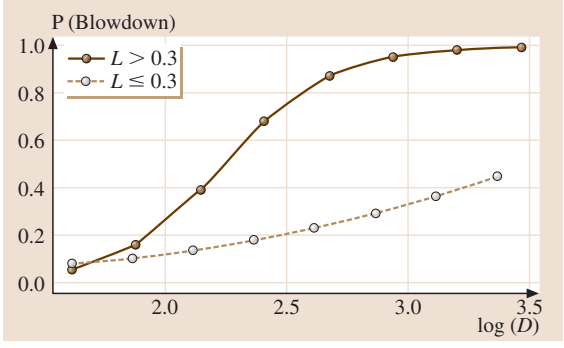

Fig. 29.5 Estimated probability of blowdown for balsam fir (BF), according to the LOTUS model in Fig. 29.4

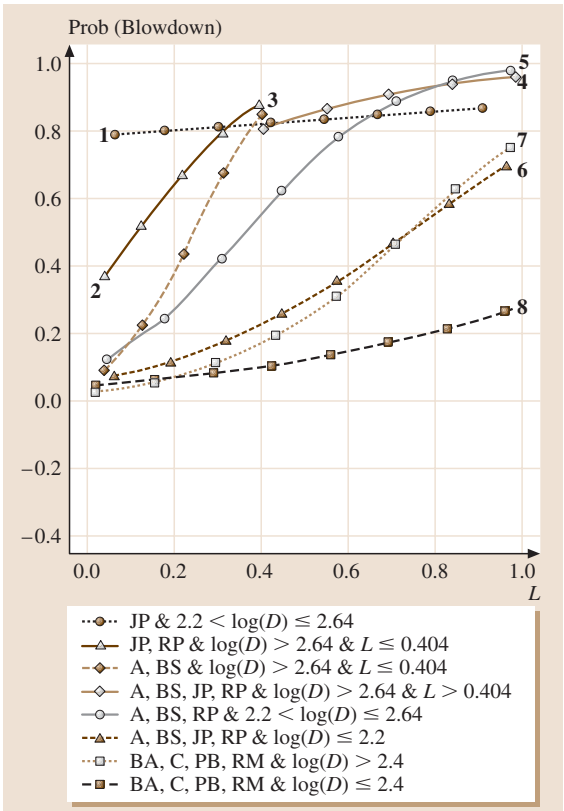

Fig. 29.6 Estimated probability of blowdown for all species except balsam firs, according to the LOTUS model in Fig. 29.4

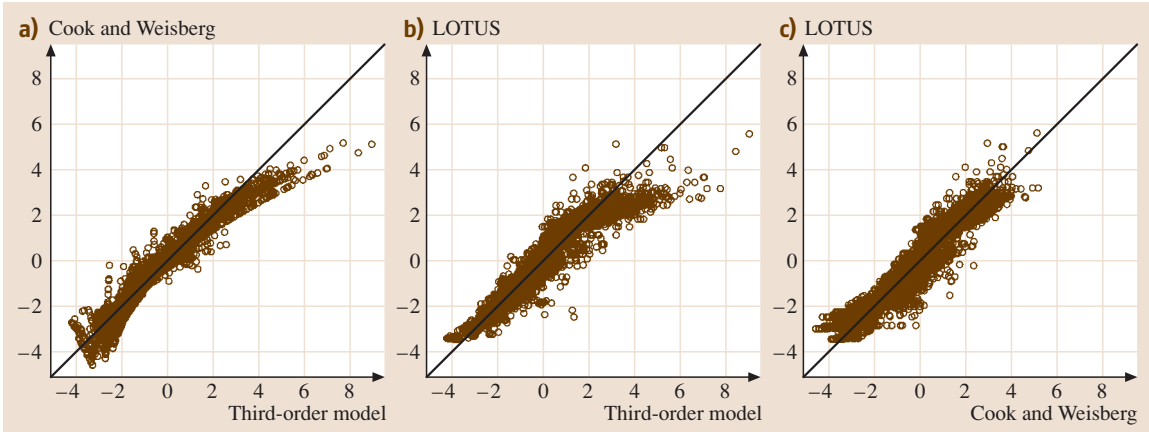

Fig. 29.7a–c Comparison of fitted logit values among three models. **(a)** Cook & Weisenberg versus third-order model **(b)** LOTUS versus third-order model **(c)** Cook & Weisenberg versus LOTUS

subsets 2 and 4. We conclude that, after balsam fir (BF), the three species most at risk of blowdown are the jack pine (JP), red pine (RP), and aspen (A), in that order. This ordering of JP, RP, and A is the same as the POD model of [29.1], as can be seen in Fig. 29.3.

6. Recall that the black spruce (BS) was the other species that [29.1] could not include in their POD model. The reason for this is quite clear from Fig. 29.6, where we use solid lines to draw the estimated p-function for black spruce. Four curves are required, corresponding to subsets 3, 4, 5, and 6. The spread of these curves suggests that the p-function of black spruce is highly sensitive to changes in D. This explains why the species cannot be included in the POD model.

How does the LOTUS model compare with the others? The former is clearly superior in terms of interpretability. But does it predict future observations as well as the other models? Unfortunately, this question cannot be answered completely, because we do not have an independent set of data to test the models. The best we can do is to compare the fitted values from the different models. This is done in Fig. 29.7, which plots the fitted logit values of the LOTUS model against those of the POD and the linear logistic regression model with all interactions up to third order. The plots show that there is not much to choose among them.

29.3 LOTUS Algorithm

As already mentioned, the idea behind LOTUS is to partition the sample space into one or more pieces such that a simple model can be fitted to each piece. This raises two issues: (i) how to carry out the partitioning, and (ii) how to decide when a partition is good enough. We discuss each question in turn.

29.3.1 Recursive Partitioning

Like all other regression tree algorithms, LOTUS splits the dataset recursively, each time choosing a single variable X for the split. If X is an ordered variable, the split has the form $s = \{X \le c\}$, where c is a constant. On the other hand, 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. The way s is chosen is critically important if the tree structure is to be used for inference about the variables.

For least-squares regression trees, many algorithms, such as automatic interaction detector (AID) [29.5], CART [29.6] and M5 [29.7], choose the split s that minimizes the total sum of squared residuals of the regression models fitted to the two data subsets created by s. Although this approach can be directly extended to logistic regression by replacing the sum of squared residuals with the deviance, it is fundamentally flawed, because it is biased toward choosing X

variables that allow more splits. To see this, suppose that X is an ordered variable taking n unique values. Then there are $n-1$ ways to split the data along the X axis, with each split $s = \{X \le c\}$ being such that c is the midpoint between two consecutively ordered values. This creates a selection bias toward X variables for which n is large. For example, in our tree dataset, variable L has 709 unique values but variable $\log(D)$ has only 87. Hence if all other things are equal, L is eight times more likely to be selected than $\log(D)$.

The situation can be worse if there are one or more categorical X variables with many values. If X takes n categorical values, there are $2^{n-1}-1$ splits of the form $s = \{X \in \omega\}$. Thus the number of splits grows exponentially with the number of categorical values. In our example, the species variable S generates $2^{9-1}-1 = 255$ splits, almost three times as many splits as $\log(D)$.

Doyle [29.8] was the first to warn that this bias can yield incorrect inferences about the effects of the variables. The GUIDE [29.9] least-squares regression tree algorithm avoids the bias by employing a two-step approach to split selection. First, it uses statistical significance tests to select X. Then it searches for c or ω. The default behavior of GUIDE is to use categorical variables for split selection only; they are not converted into indicator variables for regression modeling in the nodes. LOTUS extends this approach to logistic regression. The details are given in [29.4], but the essential steps in the recursive partitioning algorithm can be described as follows.

1. Fit a logistic regression model to the data using only the noncategorical variables.
2. For each ordered X variable, discretize its values into five groups at the sample quintiles. Form a 2×5 contingency table with the Y values as rows and the five X groups as columns. Compute the significance probability of a trend-adjusted chi-square test for nonlinearity in the data.
3. For each categorical X variable, since they are not used as predictors in the logistic regression models, compute the significance probability of the chi-square test of association between Y and X.
4. Select the variable with the smallest significance probability to partition the data.

By using tests of statistical significance, the selection-bias problem due to some X variables taking more values than others disappears. Simulation results to support the claim are given in [29.4].

After the X variable is selected, the split value c or split subset ω can be found in many ways. At the time of this writing, LOTUS examines only five candidates. If X is an ordered variable, LOTUS evaluates the splits at c equal to the 0.3, 0.4, 0.5, 0.6, and 0.7 quantiles of X. If X is categorical, it evaluates the five splits around the subset ω that minimizes a weighted sum of the binomial variance in Y in the two partitions induced by the split. The full details are given in [29.4]. For each candidate split, LOTUS computes the sum of the deviances in the logistic regression models fitted to the data subsets. The split with the smallest sum of deviances is selected.

29.3.2 Tree Selection

Instead of trying to decide when to stop the partitioning, GUIDE and LOTUS follow the CART method of first growing a very big tree and then progressively pruning it back to the root node. This yields a nested sequence of trees from which one is chosen. If an independent test dataset is available, the choice is easy: just apply each tree in the sequence to the test set and choose the tree with the lowest prediction deviance.

If a test set is not available, as is the case in our example, the choice is made by ten-fold crossvalidation. The original dataset is divided randomly into ten subsets. Leaving out one subset at a time, the entire tree-growing process is applied to the data in the remaining nine subsets to obtain another nested sequence of trees. The subset that is left out is then used as a test set to evaluate this sequence. After the process is repeated ten times, by leaving out one subset in turn each time, the combined results are used to choose a tree from the original tree sequence grown from all the data. The reader is referred to [29.6, Chapt. 3] for details on pruning and tree selection. The only difference between CART and LOTUS here is that LOTUS uses deviance instead of the sum of squared residuals.

29.4 Example with Missing Values

We now show how LOTUS works when the dataset has missing values. We use a large dataset from the National Highway Transportation Safety Administration (ftp://www.nhtsa.dot.gov/ges) on crash tests of vehicles

Table 29.2 Predictor variables in the crash-test dataset. Angular variables `crbang`, `pdof`, and `impang` are measured in degrees clockwise (from -179 to 180) with 0 being straight ahead

Name	Description	Variable type
make	Vehicle manufacturer	63 categories
model	Vehicle model	466 categories
year	Vehicle model year	continuous
body	Vehicle body type	18 categories
engine	Engine type	15 categories
engdsp	Engine displacement	continuous
transm	Transmission type	7 categories
vehtwt	Vehicle test weight	continuous
vehwid	Vehicle width	continuous
colmec	Steering column collapse mechanism	10 categories
modind	Vehicle modification indicator	4 categories
vehspd	Resultant speed of vehicle before impact	continuous
crbang	Crabbed angle	continuous
pdof	Principal direction of force	continuous
tksurf	Test track surface	5 categories
tkcond	Test track condition	6 categories
impang	Impact angle	continuous
occloc	Occupant location	6 categories
occtyp	Occupant type	12 categories
dumsiz	Dummy size percentile	8 categories
seposn	Seat position	6 categories
rsttyp	Restraint type	26 categories
barrig	Rigid or deformable barrier	2 categories
barshp	Barrier shape	15 categories

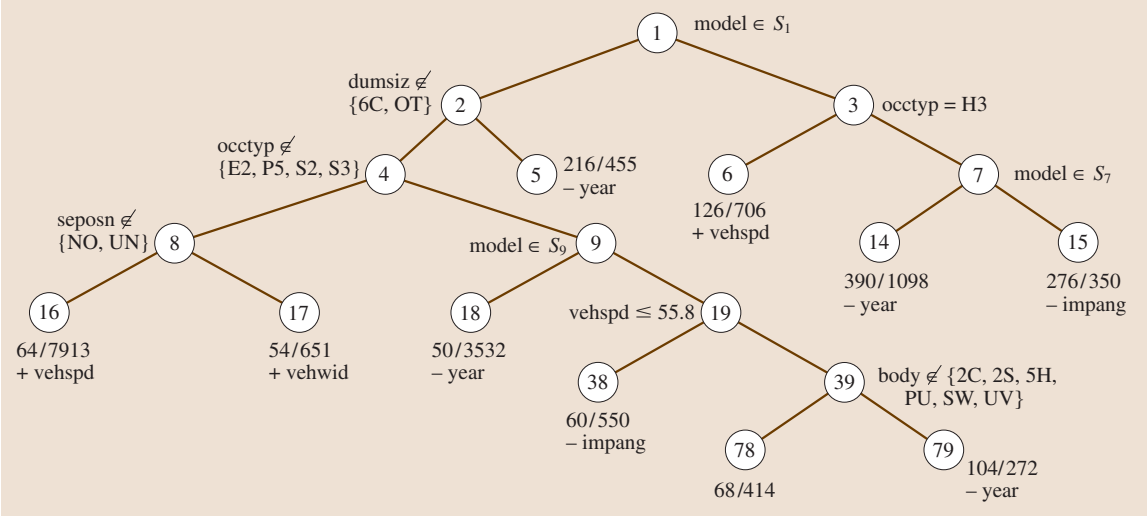

Fig. 29.8 LOTUS model for the crash-test data. Next to each leaf node is a fraction showing the number of cases with $Y = 1$ divided by the sample size, and the name of the best predictor variable, provided it is statistically significant. If the latter has a positive regression coefficient, a plus sign is attached to its name; otherwise a minus sign is shown. The constituents of the sets S_1, S_7, and S_9 may be found from Tables 29.3 and 29.4

involving test dummies. The dataset gives the results of 15 941 crash tests conducted between 1972 and 2004. Each record consists of measurements from the crash of a vehicle into a fixed barrier. The head injury criterion (`hic`), which is the amount of head injury sustained by a test dummy seated in the vehicle, is the main variable of interest. Also reported are eight continuous variables and 16 categorical variables; their names and descriptions are given in Table 29.2. For our purposes, we define $Y = 1$ if `hic` exceeds 1000, and $Y = 0$ otherwise. Thus Y indicates when severe head injury occurs.

One thousand two hundred and eleven of the records are missing one or more data values. Therefore a linear logistic regression using all the variables can be fitted only to the subset of 14 730 records that have complete values. After transforming each categorical variable into a set of indicator variables, the model has 561 regression coefficients, including the constant term. All but six variables (`engine`, `vehwid`, `tkcond`, `impang`, `rsttyp`, and `barrig`) are statistically significant. As mentioned in Sect. 29.1, however, the regression coefficients in the model cannot be relied upon to explain how each variable affects $p = P(Y = 1)$. For example, although `vehspd` is highly significant in this model, it is not significant in a simple linear logistic model that employs it as the only predictor. This phenomenon is known as Simpson's paradox. It occurs when a variable has an effect in the same direction within subsets of the data, but when the subsets are combined, the effect vanishes or reverses in direction.

Being composed of piecewise simple linear logistic models, LOTUS is quite resistant to Simpson's paradox. Further, by partitioning the dataset one variable at a time, LOTUS can use all the information in the dataset, instead of only the complete data records. Specifically, when LOTUS fits a simple linear logistic model to a data subset, it uses all the records that have complete information in Y and the X variable used in the model. Similarly, when X is being evaluated for split selection, the chi-square test is applied to all the records in the subset that have complete information in X and Y.

Figure 29.8 gives the LOTUS tree fitted to the crash-test data. The splits together with the p-functions fitted to the leaf nodes in Fig. 29.9 yield the following conclusions:

1. The tree splits first on `model`, showing that there are significant differences, with respect to p, among vehicle models. The variable is also selected for splitting in nodes 7 and 9. Tables 29.3 and 29.4 give the precise nature of the splits.
2. Immediately below the root node, the tree splits on `dumsiz` and `occtyp`, two characteristics of the test dummy. This shows that some types of dummies are more susceptible to severe injury than others. In

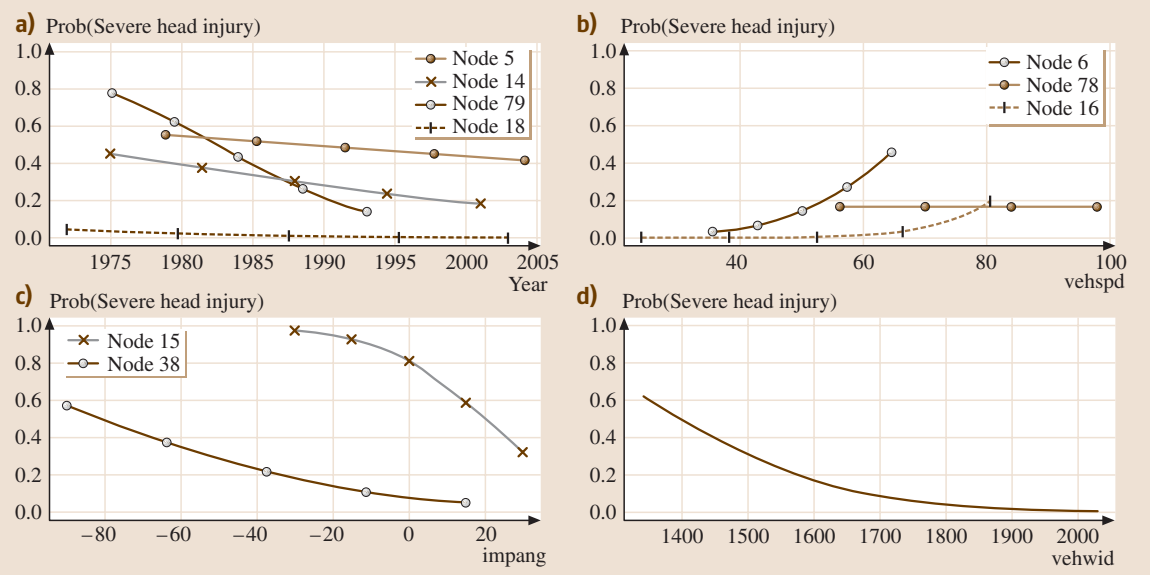

Fig. 29.9a–d Fitted probabilities of severe head injury in the leaf nodes of Fig. 29.8. **(a)** Nodes 5, 14, 18 and 79 **(b)** Nodes 6, 16, and 78 **(c)** Nodes 15 and 38 **(d)** Node 17

Table 29.3 Split at node 7 of the tree in Fig. 29.8

Make	Node 14	Node 15
American	Concord	
Audi	4000, 5000	
Buick	Electra	
Champion	Motorhome	
Chevrolet	K20 Pickup, Monza, Nova, S10 Blazer, Spectrum, Sportvan	Astro, Malibu, Sprint
Chrysler	Imperial, Lebaron	Intrepid
Comuta-Car	Electric	
Dodge	Aries, Challenger, Colt, Lancer, Magnum	Colt Pickup, St. Regis
Ford	Clubwagon MPV, Courier, E100 Van, EXP, Fairmont, Fiesta, Granada, Merkur	Torino
GMC	Sportvan	
Hyundai	Excel GLS	
Isuzu	Impulse, Spacecab	I-Mark, Trooper II
Jeep	Comanche	
Kia	Sorento	
Lectric	Leopard	
Mazda	GLC	B2000
Mercury		Bobcat
Mitsubishi	Montero, Tredia	Pickup
Nissan	2000, 210, Kingcab Pickup, Murano	
Oldsmobile		98
Peugeot		504, 505
Plymouth	Champ, Fury, Horizon	Breeze, Volare
Pontiac	T1000	
Renault	18, Alliance, LeCar, Medallion	Fuego, Sportswagon
Saab	38235	
Saturn	L200	
Subaru	GF, GLF, Wagon	
Suzuki	Sidekick	
Toyota	Celica, Starlet	
Volkswagen	Fox, Scirocco	Beetle, EuroVan
Volvo	244, XC90	
Yugo	GV	

particular, the cases in node 5 contain mainly dummies that correspond to a six-year-old child. The fitted p-function for this node can be seen in the upper left panel of Fig. 29.9. Compared with the fitted p-functions of the other nodes, this node appears to have among the highest values of p. This suggests that six-year-old children are most at risk of injury. They may be too big for child car seats and too small for adult seat belts.

3. The split on `seposn` at node 8 shows that passengers in vehicles with adjustable seats are ten times (average p of 0.008 versus 0.08) less likely to suffer severe head injury than those with nonadjustable seats. This could be due to the former type of vehicle being more expensive and hence able to withstand collisions better.
4. Similarly, the split on `body` at node 39 shows that passengers in two-door cars, pick-ups, station wagons, and sports utility vehicles (SUVs) are twice as likely (average p of 0.38 versus 0.16) to suffer severe head injury than other vehicles.
5. The linear predictor variables selected in each leaf node tell us the behavior of the p-function within each partition of the dataset. Four nodes have `year` as their best linear predictor. Their fitted p-functions are shown in the upper left panel of Fig. 29.9. The

Table 29.4 Split at node 9 of the tree in Fig. 29.8

Make	Node 18	Node 19
Acura	Integra, Legend, Vigor	2.5TL, 3.2TL, 3.5RL, MDX, RSX
American	Gremlin, Matador, Spirit	
Audi	100, 200, 80	A4, A6, A8
Battronics	Van	
BMW	325I, 525I	318, 328I, X5, Z4 Roadster
Buick	Century, LeSabre, Regal, Riviera, Skyhawk, Skylark, Somerset	ParkAvenue, Rendezvous, Roadmaster
Cadillac	Deville, Seville	Brougham, Catera, Concourse, CTS, Eldorado, Fleetwood
Chevrolet	Beretta, Camaro, Cavalier, Celebrity, Chevette, Citation, Corsica, Corvette, Elcamino, Impala, Lumina, LUV, MonteCarlo, Pickup, S-10, Vega	Avalanche, Beauville, Blazer, C-1500, K2500 Pickup, Silverado, Suburban, Tahoe, Tracker, Trailblazer, Venture,
Chinook	Motorhome	
Chrysler	Cirrus, Conquest, FifthAvenue, Newport, NewYorker	LHS, Pacifica, PT Cruiser, Sebring Convertible
Daewoo		Leganza, Nubira
Daihatsu	Charade	
Delorean	Coupe	
Dodge	400, 600, Caravan, D-150, Dakota, Daytona, Diplomat, Dynasty, Mirada, Neon, Rampage, Ramwagonvan, Sportsman	Avenger, Durango, Grand Caravan, Intrepid, Omni, Ram150, Ram1500, Ram, Ram250 Van, Shadow, Spirit, Stratus
Eagle	Medallion, MPV, Premier	Summit, Vision
Eva	Evcort	
Fiat	131, Strada	
Ford	Bronco, Bronco II, Crown Victoria, Escort, F150 Pickup, F250 Pickup, F350 Pickup, Festiva, LTD, Mustang, Pickup, Probe, Ranger, Taurus, Thunderbird, Van, Windstar	Aerostar, Aspire, Contour, E150 Van, Escape, Escort ZX2, EV Ranger, Expedition, Explorer, Focus, Freestar, Other, Tempo
Geo	Metro, Prizm	Storm, Tracker
GMC	Astro Truck, Vandura	EV1
Holden		Commodore Acclaim
Honda	Accord	Civic, CRV, Element, Insight, Odyssey, Pilot, Prelude, S2000
Hyundai	Elantra, Scoupe, Sonata	Accent, Pony Excel, Santa Fe, Tiburon
IH	Scout MPV	
Infinity	G20, M30	J30
Isuzu	Amigo, Pup	Axiom, Pickup, Rodeo, Stylus
Jaguar		X-Type
Jeep	CJ, Wrangler	Cherokee, Cherokee Laredo, Grand Cherokee, Liberty
Jet	Courier, Electrica, Electrica 007	
Kia	Sephia	Rio, Sedona, Spectra, Sportage

Table 29.5 Split at node 9 of the tree in Fig. 29.8 (cont.)

Make	Node 18	Node 19
Landrover		Discovery, Discovery II
Lectra	400, Centauri	
Lexus	ES250	ES300, GS300, GS400, IS300, RX300, RX330
Lincoln	Continental, Town Car	LS, Mark, Navigator
Mazda	323, 323-Protege, 929, Miata, Millenia, MPV, MX3, MX6, Pickup, RX	626, Mazda6, MX5
Mercedes	190, 240, 300	C220, C230, C240, E320, ML320
Mercury	Capri, Cougar, Lynx, Marquis, Monarch, Sable, Topaz, Tracer, Villager, Zephyr	Mystique
Mini		Cooper
Mitsubishi	Diamante, Eclipse, Galant, Mightymax, Mirage, Precis, Starion, Van	3000GT, Cordia, Endeavor, Lancer, Montero Sport, Outlander
Nissan	240SX, 810, Altima, Axxess, Pathfinder, Pulsar, Quest, Sentra, Van	200SX, 300ZX, 350Z, Frontier, Maxima, Pickup, Stanza, Xterra
Odyssey	Motorhome	
Oldsmobile	Calais, Cutlass, Delta 88, Omega, Toronado	Achieva, Aurora, Intrigue, Royale
Other	Other	
Peugeot	604	
Plymouth	Acclaim, Caravelle, Laser, Reliant, Sundance, Voyager	Colt Vista, Conquest, Neon
Pontiac	Bonneville, Fiero, Firebird, Grand AM, Lemans, Parisienne, Sunbird	Aztek, Grand Prix, Sunfire, Trans Sport
Renaissance		Tropica
Renault	Encore	
Saab	900	38233, 9000
Saturn	SL1	Ion, LS, LS2, SC1, SL2, Vue
Sebring		ZEV
Solectria		E-10, Force
Subaru	DL, Impreza, Justy, XT	Forestee, GL, Legacy
Suzuki	Samurai	Swift, Vitara
Toyota	Camry, Corolla, Corona, Cosmo, Landcruiser, MR2, Paseo, T100, Tercel, Van	4Runner, Avalon, Camry Solara, Cressida, Echo, Highlander, Matrix, Pickup, Previa, Prius, Rav4, Sequoia, Sienna, Tacoma, Tundra
UM	Electrek	
Volkswagen	Cabrio, Corrado, Golf, Passat, Quantum, Rabbit	Jetta, Polo, Vanagon
Volvo	240, 740GL, 850, 940, DL, GLE	960, S60, S70, S80
Winnebago	Trekker	

decreasing trends show that crash safety is improving over time.

6. Three nodes have `vehspd` as their best linear predictor, although the variable is not statistically significant in one (node 78). The fitted p-functions are shown in the upper right panel of Fig. 29.9. As expected, p is nondecreasing in `vehspd`.
7. Two nodes employ `impang` as their best linear predictor. The fitted p-functions shown in the bottom left panel of Fig. 29.9 suggest that side impacts are more serious than frontal impacts.

8. One node has `vehwid` as its best linear predictor. The decreasing fitted p-function shown in the lower right panel of Fig. 29.9 shows that vehicles that are smaller are less safe.

29.5 Conclusion

Logistic regression is a statistical technique for modeling the probability p of an event in terms of the values of one or more predictor variables. The traditional approach expresses the logit of p as a linear function of these variables. Although the model can be effective for predicting p, it is notoriously hard to interpret. In particular, multi-collinearity can cause the regression coefficients to be misinterpreted.

A logistic regression tree model offers a practical alternative. The model has two components, namely, a binary tree structure showing the data partitions and a set of simple linear logistic models, fitted one to each partition. It is this division of model complexity that makes the model intuitive to interpret. By dividing the dataset into several pieces, the sample space is effectively split into different strata such that the p-function is adequately explained by a single predictor variable in each stratum. This property is powerful because: (i) the partitions can be understood through the binary tree, and (ii) each p-function can be visualized through its own graph. Further, stratification renders each of the individual p-functions resistant to the ravages of multi-collinearity among the predictor variables and to Simpson's paradox. Despite these advantages, it is crucial for the partitioning algorithm to be free of selection bias. Otherwise, it is very easy to draw misleading inferences from the tree structure. At the time of writing, LOTUS is the only logistic regression tree algorithm designed to control such bias.

Finally, as a disclaimer, it is important to remember that, in real applications, there is no *best* model for a given dataset. This situation is not unique to logistic regression problems; it is prevalent in least-squares and other forms of regression as well. Often there are two or more models that give predictions of comparable average accuracy. Thus a LOTUS model should be regarded as merely one of several possibly different ways of explaining the data. Its main virtue is that, unlike many other methods, it provides an *interpretable* explanation.

References

29.1 R. D. Cook, S. Weisberg: Partial one-dimensional regression models, Am. Stat. **58**, 110–116 (2004)

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

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

29.4 K.-Y. Chan, W.-Y. Loh: LOTUS: An algorithm for building accurate and comprehensible logistic regression trees, J. Comp. Graph. Stat. **13**, 826–852 (2004)

29.5 J. N. Morgan, J. A. Sonquist: Problems in the analysis of survey data, and a proposal, J. Am. Stat. Assoc. **58**, 415–434 (1963)

29.6 L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone: *Classification and Regression Trees* (Wadsworth, Belmont 1984)

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

29.8 P. Doyle: The use of automatic interaction detector and similar search procedures, Oper. Res. Q. **24**, 465–467 (1973)

29.9 W.-Y. Loh: Regression trees with unbiased variable selection and interaction detection, Stat. Sin. **12**, 361–386 (2002)