Generalized Regression Trees
(*Statistica Sinica* 1995, v. 5, pp. 641–666)

Probal Chaudhuri Wen-Da Lo Wei-Yin Loh Ching-Ching Yang

Indian Statistical Institute, Calcutta, Chung Cheng Institute of Technology, Taiwan, University of Wisconsin, Madison, and Feng Chia University, Taiwan

Abstract

A method of generalized regression that blends tree-structured nonparametric regression and adaptive recursive partitioning with maximum likelihood estimation is studied. The function estimate is a piecewise polynomial, with the pieces determined by the terminal nodes of a binary decision tree. The decision tree is constructed by recursively partitioning the data according to the signs of the residuals from a model fitted by maximum likelihood to each node. Algorithms for tree-structured Poisson and logistic regression and examples to illustrate them are given. Large-sample properties of the estimates are derived under appropriate regularity conditions.

Key words and phrases: Anscombe residual, consistency, generalized linear model, maximum likelihood, pseudo residual, recursive partitioning, Vapnik-Chervonenkis class.

1 Introduction: motivation and main ideas

Consider a general regression setup in which a real-valued response Y is related to a real or a vector-valued regressor X through a probability model that characterizes the nature of the dependence of Y on X. Let $f\{y|g(x)\}$ denote the conditional density or mass function of Y given $X = x$, where the form of f is known but g is an unknown function to be estimated. Familiar examples include the logistic regression model (where Y is binary, and $g(x)$ is the "logit" of the conditional probability parameter given $X = x$), the Poisson regression model (where Y is a nonnegative integer-valued random variable with a Poisson distribution, and $g(x)$ is related to its unknown conditional mean given $X = x$), and generalized linear models (GLM) (Nelder and Wedderburn 1972, McCullagh and Nelder 1989), where g is related to the link function. On the other hand, $g(x)$ may be the unknown location parameter associated with the conditional distribution of Y given $X = x$. That is, Y may satisfy the equation $Y = g(X) + \epsilon$, where the conditional distribution of ϵ may be normal, Cauchy or exponential power (see, e.g., Box and Tiao 1973) with center at zero.

We focus on the situation where no finite-dimensional parametric model is imposed on g, and it is assumed to be a fairly smooth function. Nonparametric estimation of the functional parameter g has been explored by Chaudhuri and Dewanji (1995), Cox and O'Sullivan (1990), Gu (1990), Hastie and Tibshirani (1986, 1990), O'Sullivan, Yandell and Raynor (1986), Staniswalis (1989), Stone (1986, 1991a), and others, who considered nonparametric smoothers when the conditional distribution of the response given the regressor is assumed to have a known shape (e.g., the conditional distribution may possess a GLM-type exponential structure).

In the case of the usual regression setup, where $Y = g(X)+\epsilon$ with $E(\epsilon|X) = 0$, several attempts have been made to estimate g by recursively partitioning the regressor space and then constructing

a regression estimate in each partition using the method of least squares. Some examples are AID (Sonquist 1970, Sonquist, Baker and Morgan 1973), CART (Breiman, Friedman, Olshen and Stone 1984) and SUPPORT (Chaudhuri, Huang, Loh and Yao 1994). The purpose of this article is to explore recursive partitioning algorithms and related likelihood-based nonparametric function estimates in a generalized regression setting.

Tree-structured regression possesses three significant advantages over standard parametric and nonparametric regression:

1. By allowing the tree-structure to handle much of the overall model complexity, the models in each partition can be kept at a low order and hence be more easily interpreted.

2. Interactions among covariates are directly conveyed by the structure of the decision tree. As a result, interactions can be understood and interpreted more easily in qualitative terms.

3. The simple form of the fitted function in each terminal node permits the statistical properties of the method to be studied analytically.

The adaptive nature of recursive partitioning allows varying degrees of smoothing over the regressor space so that the terminal nodes may have variable sizes in terms of both numbers of observations and diameters of the sets in the regressor space to which they correspond. The main motivation behind such adaptive variable smoothing is to take care of heteroscedasticity as well as the possibility that the amount of smoothness in the functional parameter g may be different in different parts of the regressor space. This is an improvement over most of the earlier nonparametric estimation techniques in generalized regression, which concentrate either on adaptive but non-variable smoothing (i.e., using a smoothing parameter whose value is constant over the entire regressor space) or on deterministic smoothing.

The general recursive partitioning methodology explored in this paper consists of two recursive steps: (i) the function g is estimated from the data in each node by a low order polynomial using maximum likelihood and (ii) each node is split into two subnodes using a criterion based on the distributions of the covariate vectors according to the signs of the residuals. Recursive partitioning stops when the number of cases in each terminal node is smaller than a pre-assigned threshold. A cross-validation pruning procedure (Breiman et al. 1984) is applied to determine the final tree. Sections 2 and 3 give specific algorithms and illustrative examples for Poisson and logistic regression, respectively. One of the examples also shows how categorical (unordered) covariates can be included in the models.

Adaptive recursive partitioning algorithms construct random subsets of the regressor space to form the terminal nodes. A serious technical barrier in studying the analytic properties of the likelihood-based function estimates is the random nature of these subsets. A key tool in coping with this situation is a well-known combinatorial result of Vapnik and Chervonenkis (1971). In Section 4, we investigate the large-sample statistical properties of the estimates that are constructed via recursive partitioning of the regressor space followed by maximum likelihood estimation of g by piecewise polynomials.

The MARS (Friedman 1991) method combines spline fitting with recursive partitioning to produce a continuous regression function estimate. The complexity of the estimate makes interpretation difficult and theoretical analysis of its statistical properties extremely challenging. In the SUPPORT method of Chaudhuri et al. (1994), a weighted averaging technique is used to combine piecewise-polynomial fits into a smooth one. An identical technique can be used here to create a smooth estimate from a discontinuous piecewise-polynomial estimate without altering the asymptotic properties of the original estimate (see Chaudhuri, Lo, Loh and Yang (1993) for some

examples). Proposals for extending MARS to logistic regression and GLM-type problems are given in Friedman (1991), Buja, Duffy, Hastie and Tibshirani (1991) and Stone (1991b). Our approach is more general as it is applicable to other regression setups in addition to logistic regression.

2 Poisson regression trees

Our algorithm for fitting Poisson regression trees has three main components: (i) a method to select the variable and the splitting value to be used at a partition, (ii) a method to determine the size of the tree, and (iii) a method to fit a model to each terminal node. Although there are many reasonable solutions for each component (see Yang (1993) for some variations), the model fitting for the examples in this section is carried out recursively as follows.

1. The Poisson loglinear model, $\log(m) = \beta_0 + \sum_{k=1}^{K} \beta_k x_k$, is fitted to the data in node t. Here $m = EY$ and $x_1, \ldots, x_K$ are the K covariates.

2. Let $\hat{m}_i$ be the estimated value of m for the ith case and let y_i denote the observed value of Y_i. The adjusted Anscombe residual (Pierce and Schafer 1986)

$$r_i = \{y_i^{2/3} - (\hat{m}_i^{2/3} - (1/9)\hat{m}_i^{-1/3})\}/\{(2/3)\hat{m}_i^{1/6}\} \tag{1}$$

   is calculated for each y_i in t.

3. Observations with nonnegative r_i are classified as belonging to one group and the remainder to a second group.

4. Two-sample t-statistics to test for differences in means and variances between the two groups along each covariate axis are computed. The latter test is Levene's (1960) test.

5. The covariate selected to split the node is the one with the largest absolute t-statistic. The cut-point for the selected covariate is the average of the two group means along the covariate. Observations with covariate values less than or equal to the cut-point are channeled to the left subnode and the remainder to the right subnode.

6. After a large tree is constructed, a nested sequence of subtrees is obtained by progressively deleting branches according to the pruning method of Breiman et al. (1984), with residual deviance replacing apparent error in the cost-complexity function.

7. The subtree with the smallest cross-validation estimate of deviance is selected.

Remark 1. Our split selection strategy is motivated by the methods in Chaudhuri et al. (1994) for tree-structured least squares regression and Ahn and Loh (1994) for tree-structured proportional hazards regression. It differs fundamentally from the exhaustive search strategy used in the AID and CART algorithms. The latter strategy calls for all possible splits of the data in a node to be evaluated to find the one that most reduces some measure of node impurity (e.g., deviance). In the present problem, this requires Poisson loglinear models to be fitted to the subnodes induced by *every* split. Because loglinear fitting typically involves Newton-Raphson iteration, this strategy is not practical for routine application on present-day workstations. Our split selection strategy performs model fitting only once at each node. The task of finding the best split is reduced to a classification problem by grouping the covariate vectors into two classes according to the signs of the residuals. The t-tests, which were developed for tree-structured classification in Loh and

Table 1: Coefficients from two Poisson loglinear models fitted to NNM data

Model	Term	Coefficient	t-value
First-	Intercept	1.71862	23.64
degree	Dose	0.02556	26.75
GLM	Time	0.00122	7.74
Second-	Intercept	-1.529E+00	-4.71
degree	Dose	4.251E-02	4.43
GLM	Time	1.308E-02	8.73
	Dose-squared	-4.547E-04	-7.76
	Time-squared	-1.150E-05	-7.12
	Dose $\times$ Time	2.712E-04	10.14

Vanichsetakul (1988), essentially rank the covariates in terms of the degree of clustering of the signs. The highest ranking covariate is interpreted as the direction in which lack of model fit is greatest and is selected to split the node.

Remark 2. Empirical evidence (Yang 1993) suggests that the adjusted Anscombe residuals defined in (1) tend to yield superior splits compared to the unadjusted Anscombe residuals, especially when some of the Poisson means are small. The Pearson and deviance residuals are not employed because they have the same signs as the unadjusted Anscombe residuals.

We now give two examples to illustrate the Poisson regression tree method. The first example uses ordered covariates and the second example categorical (unordered) covariates.

2.1 Effect of N-nitrosomorpholine (NNM) on rats

The data come from an experiment (Moolgavkar, Luebeck, de Gunst, Port and Schwarz 1990) in which 173 female rats were exposed to a chemical, N-nitrosomorpholine, at various doses (0, 0.1, 1, 5, 10, 20, 40, 80ppm) in their drinking water starting at 14 weeks of age. The animals were killed at different ages and three sections from the identical lobe of the liver were examined for the number of ATPase-deficient transections. The response is the number of transections, which ranged from 0 to 160. These transections, sometimes called *foci*, are believed to represent clones of premalignant cells. The time to sacrifice ranged from 42 to 686 days.

Table 1 gives the results from fitting a first-degree and then a full second-degree Poisson loglinear model to the data. The residual deviances for the two models are 3,455 and 2,027 with 170 and 167 degrees of freedom, respectively. Clearly, the first-degree model is rejected in favor of the second-degree model. Notice that the coefficients for dose-squared and time-squared are negative and that the most significant term is the interaction between dose and time. This makes interpretation tricky.

The tree in Figure 1 shows the result of fitting piecewise Poisson loglinear models with the proposed method using only main effect terms. The presence of the dose-time interaction is obvious from the splits. The tree has five terminal nodes and its residual deviance is 1,432. The sample mean number of transections is given beside each terminal node. This increases from 0.9 when both dose and time are small to 29.8 when both covariates take large values. The regression coefficients and t-statistics for the models at the terminal nodes are given in Table 2. The coefficients for dose and time are all positive as expected. Except at node 8, both covariates are highly statistically

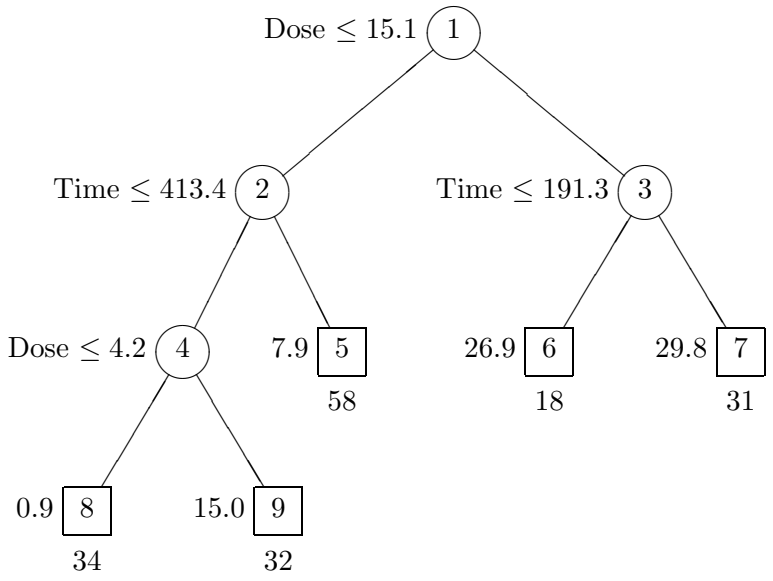

Figure 1: Poisson regression tree for NNM data using 10-fold cross-validation. A case goes to the left subnode if the condition at the split is true. The number beneath a terminal node is the learning sample size. The number on the left of a terminal is the sample mean number of transections.

Table 2: Estimated coefficients and t-values for models in terminal nodes in Figure 1.

Node	Intercept		Dose		Time		Residual	
no.	Coef.	t	Coef.	t	Coef.	t	deviance	Df
5	-1.970	-4.3	0.537	14.6	0.0062	8.5	479	55
6	-3.377	-8.1	0.039	13.5	0.0282	16.2	86	15
7	-0.231	-0.9	0.051	12.8	0.0093	14.2	629	28
8	-2.352	-2.9	1.519	3.8	0.0049	2.1	69	31
9	-2.439	-7.8	0.112	5.8	0.0146	20.4	168	29

significant. Since the nearest dose to 5ppm in the experiment was 1ppm, this implies that the number of transections is essentially random if the dose level is 1ppm or lower and sacrifice time is less than 414 days.

Figure 2 shows contour plots of the fitted log-means for the second-degree GLM and tree-structured models with the observed points superimposed. The contours for the tree-structured model are piecewise-linear and they track the shape of the contours from the GLM model. Observe that the data points are concentrated near the left and bottom sides of the plots and that the contours in the GLM plot increase rapidly in the upper-right corner. Notice also that the contour line for zero log-count in this plot has a U-shape. These are artifacts caused by the quadratic components in the GLM model. The tree-structured model does not have these problems because it models the data in a piecewise fashion. The trade-off is lack of smoothness of the fitted surface at the partition boundaries.

Qualitatively similar results are obtained when the above analysis is repeated with log(dose +

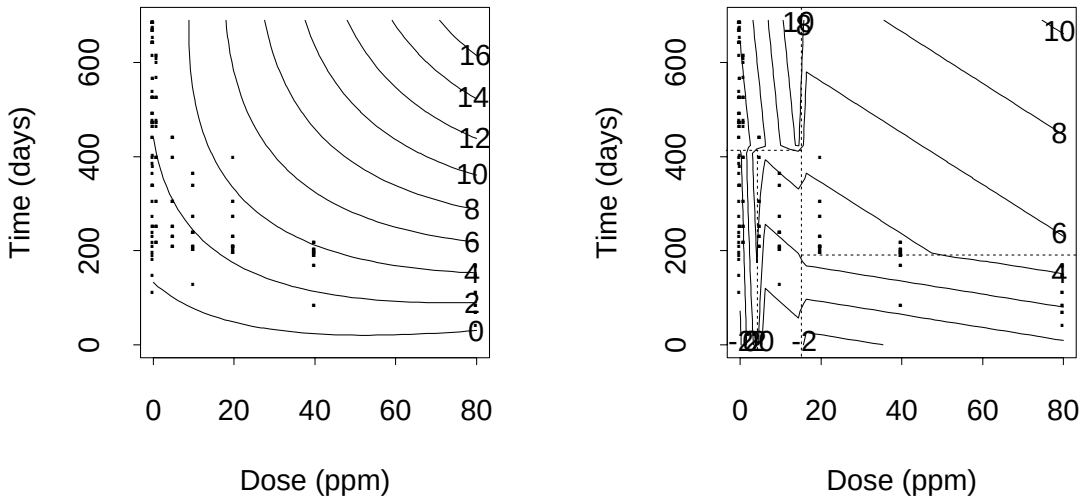

Figure 2: Contour plots of predicted log-means of number of transections. The left plot is for the second-degree GLM with interaction and the right one for the tree-structured model. Dotted lines in the lower plot mark the partitions and observations are indicated by dots.

0.01) instead of dose. All the coefficients except the one for $\{\log(\text{dose}+0.01)\}^2$ are highly significant when a second-degree GLM model is fitted to the entire data set. The corresponding Poisson regression tree has the same number of terminal nodes as before, but with different splits.

2.2 A factorial experiment with categorical covariates

The data come from an unreplicated $3 \times 2 \times 4 \times 10 \times 3$ experiment on wave-soldering of electronic components on printed circuit boards (Comizzoli, Landwehr and Sinclair 1990). There are 720 observations and the covariates are all categorical variables. The factor levels are:

1. *Opening:* amount of clearance around a mounting pad (levels 'small', 'medium', or 'large')
2. *Solder:* amount of solder (levels 'thin' and 'thick')
3. *Mask:* type and thickness of the material for the solder mask (levels A1.5, A3, B3, and B6)
4. *PadType:* geometry and size of the mounting pad (levels W4, D4, L4, D6, L6, D7, L7, L8, W9, and L9)
5. *Panel:* each board was divided into three panels (levels 1, 2, and 3)

The response is the number of solder skips which range from 0–48.

Table 3 gives the results from fitting a Poisson loglinear model to the data with all two-factor interactions. The three most significant two-factor interactions are between *Opening*, *Solder*, and *Mask*. These variables also have the most significant main effects. Chambers and Hastie (1992, p.

Table 3: Results from a full second-degree Poisson loglinear model fitted to solder data

Term	Df	Sum of Sq	Mean Sq	F-value	Pr(F)
Opening	2	1587.563	793.7813	568.65	0.00000
Solder	1	515.763	515.7627	369.48	0.00000
Mask	3	1250.526	416.8420	298.62	0.00000
PadType	9	454.624	50.5138	36.19	0.00000
Panel	2	62.918	31.4589	22.54	0.00000
Opening:Solder	2	22.325	11.1625	8.00	0.00037
Opening:Mask	6	66.230	11.0383	7.91	0.00000
Opening:PadType	18	45.769	2.5427	1.82	0.01997
Opening:Panel	4	10.592	2.6479	1.90	0.10940
Solder:Mask	3	50.573	16.8578	12.08	0.00000
Solder:PadType	9	43.646	4.8495	3.47	0.00034
Solder:Panel	2	5.945	2.9726	2.13	0.11978
Mask:PadType	27	59.638	2.2088	1.58	0.03196
Mask:Panel	6	20.758	3.4596	2.48	0.02238
PadType:Panel	18	13.615	0.7564	0.54	0.93814
Residuals	607	847.313	1.3959		

10) (see also Hastie and Pregibon (1992, p. 217)) analyze these data and conclude that a parsimonious model is one containing all main effect terms and these three two-factor interactions. The residual deviance for the latter model is 972 with 691 degrees of freedom (the null deviance is 6,856 with 719 degrees of freedom). Estimates of the individual terms in this model are given in Table 4. The model is very complicated and is not easy to interpret.

To confirm the inadequacy of a main-effects model, we fit a Poisson regression tree to these data using only main effects models in each node. Because the covariates are categorical, we need to convert them to ordered variables before using the algorithm. Instead of arbitrarily assigning scores, we use a loglinear model to determine the scores as follows. Let X be a categorical variable with values in the set $\{1, 2, \ldots, c\}$.

1. Define dummy variables $Z_1, \ldots, Z_{c-1}$ such that

$$Z_k = \begin{cases} 1 & \text{if } X = k \\ 0 & \text{if } X \neq k \end{cases} \qquad k = 1, \ldots, c-1.$$

2. Fit the Poisson loglinear model, $\log(m) = \gamma_0 + \gamma_1 Z_1 + \cdots + \gamma_{c-1} Z_{c-1}$, and let $\hat{\gamma}_i$ $(i = 0, \ldots, c-1)$ denote the estimated coefficients.

3. Transform X to the ordered variable V, where

$$V = \begin{cases} \hat{\gamma}_0 + \hat{\gamma}_k & \text{if } X = k \text{ and } k \neq c \\ \hat{\gamma}_0 & \text{if } X = c. \end{cases}$$

4. Use the variable V in place of X in the main algorithm.

Table 4: Estimates from a Poisson loglinear model fitted to solder data. The model contains all main effects and all two-factor interactions involving *Opening*, *Solder*, and *Mask*. The letters 'L' and 'Q' below refer to the linear and quadratic components of the *Opening* factor.

Term	Value	t
Intercept	0.5219	12.28
Opening.L	-1.6244	-24.68
Opening.Q	0.4573	6.93
Solder	-1.0894	-20.83
Mask1	0.3110	4.47
Mask2	0.3834	13.07
Mask3	0.4192	28.11
PadType1	0.0550	1.66
PadType2	0.1058	6.10
PadType3	-0.1049	-6.92
PadType4	-0.1229	-9.03
PadType5	0.0131	1.48
PadType6	-0.0466	-5.28
PadType7	-0.0076	-1.09
PadType8	-0.1355	-12.79
PadType9	-0.0283	-4.31
Panel1	0.1668	7.93
Panel2	0.0292	2.49
Opening.LSolder	-0.3808	-5.19
Opening.QSolder	-0.2607	-3.63
Opening.LMask1	0.0308	0.32
Opening.QMask1	-0.3510	-3.45
Opening.LMask2	0.0524	1.28
Opening.QMask2	0.2024	4.26
Opening.LMask3	0.0871	4.07
Opening.QMask3	-0.0187	-0.80
SolderMask1	0.0120	0.16
SolderMask2	0.1858	6.10
SolderMask3	0.1008	6.25

Table 5: Covariate V-scores for solder data

Covariate	Opening					Mask				
Category	Small	Medium	Large			A1.5	A3	B3	B6	
V-score	11.071	2.158	1.667			1.611	2.472	5.361	10.417	
Covariate	Panel					Solder				
Category	1	2	3			Thin	Thick			
V-score	4.042	5.642	5.213			7.450	2.481			
Covariate	Pad type									
Category	W4	D4	L4	D6	L6	D7	L7	L8	W9	L9
V-score	5.972	6.667	8.667	4.611	3.417	6.042	4.083	5.083	1.583	3.528

Table 6: Estimated coefficients for loglinear models in terminal nodes of tree in Figure 3

Covariate	Node 4		Node 5		Node 6		Node 8		Node 9	
	Coef.	t	Coef.	t	Coef.	t	Coef.	t	Coef.	t
Intercept	-4.674	-4.93	-3.036	-9.10	-3.910	-8.67	-0.997	-2.21	0.753	3.23
Opening	0.139	6.38	0.226	23.33	0.210	1.38	-	-	-	-
Mask	0.542	2.38	0.136	9.11	0.223	20.38	0.358	3.33	0.090	8.54
Pad type	0.257	5.15	0.212	11.42	0.226	11.65	0.209	8.76	0.166	12.29
Panel	0.152	1.05	0.122	2.25	0.389	6.42	0.241	3.38	0.169	4.27

This method of scoring is similar in concept to the method of dealing with categorical covariates in the FACT method (Loh and Vanichsetakul 1988) for tree-structured classification, although in the latter the scoring is done at each node instead of merely at the root node; a future version of the present algorithm will perform scoring at every node.

The V-scores for each covariate are given in Table 5. Notice that for the variable *Opening*, the score assigned to the 'small' category is much larger than those assigned to the 'medium' and 'large' categories. This suggests that the response is likely to be quite a bit larger when *Opening* is small than when it is medium or large. Similarly, the scores for *Mask* when it takes values B3 or B6 are much larger than for other values.

Figure 3 shows the Poisson regression tree. It has a residual deviance of 1,025. The splits are on *Solder*, *Mask* and *Opening*, indicating substantial interactions among these covariates. The sample mean response is given beside each terminal node of the tree. These numbers show that the response is least when the *Solder* amount is thick and the *Mask* is A1.5 or A3. It is largest when the *Solder* amount is thin, *Opening* is small, and the *Mask* is B3 or B6. These conclusions are not readily apparent from Tables 3 and 4. Table 6 gives the estimated coefficients for the loglinear models in each terminal node. Because the effect of interactions is modeled in the splits, no interaction terms are needed in the piecewise models.

Figure 4 shows plots of the observed values versus the fitted values from the tree-structured model and from the generalized linear model with all main effects and all two-factor interactions involving *Solder*, *Mask*, and *Opening*. The agreement between two sets of fitted values is quite good

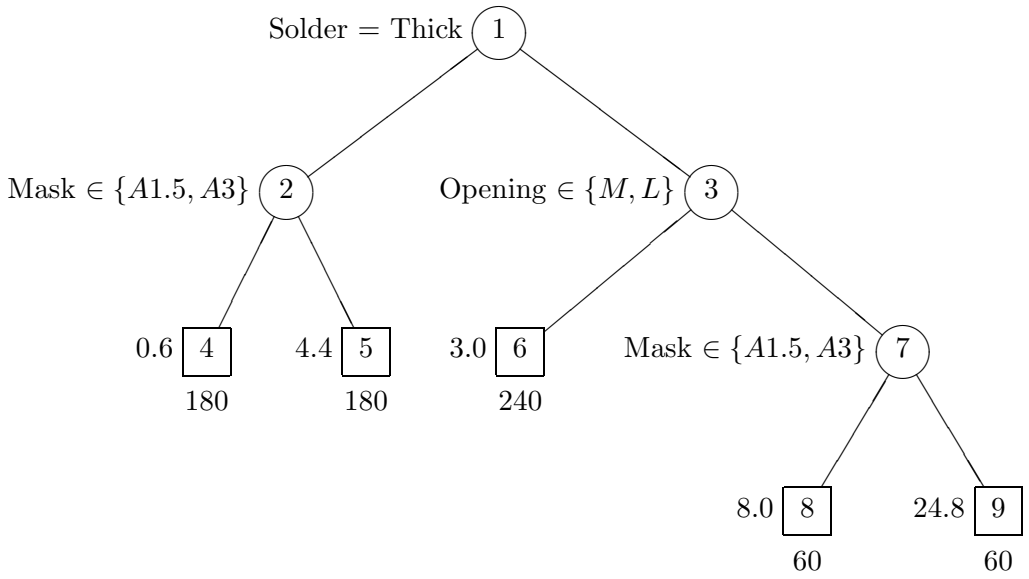

Figure 3: Poisson regression tree for solder data using 10-fold cross-validation. A case goes to the left subnode if the condition at a split is true. The number beneath a terminal node is the learning sample size. The number on the left of a terminal node is the sample mean number of solder skips.

and lends support to our method of scoring categorical variables.

3 Logistic regression trees

The basic algorithm for Poisson regression trees is applicable to logistic regression trees. The only difference is that a more careful definition of residual is needed. This is because the 0-1 nature of the response variable Y makes the signs of the Pearson and deviance residuals too variable (Lo (1993) gives some empirical evidence). To reduce the amount of variability, the following additional steps are taken at each node to smooth the observed Y-values prior to computation of the residuals.

1. Compute $\hat{p}_i$, the estimate of $p_i = P(Y_i = 1)$ from a logistic regression model.

2. Smooth the Y-values using the following nearest-neighbor average method (Fowlkes 1987). Let $d(x_s, x_i)$ be the Euclidean distance between the standardized (i.e., sample variance one) values of x_s and x_i and let A_s be the set of $[hn]$-nearest neighbors of x_s, where $h \in (0, 1)$ is a fixed smoothing parameter. Define $d_s = \max_{x_i \in A_s} d(x_s, x_i)$. The smoothed estimate (called a 'pseudo-observation') is given by

$$p_s^* = \sum_{x_i \in A_s} w_{i,s} y_i \Big/ \sum_{x_i \in A_s} w_{i,s}$$

   where $w_{i,s} = \{1 - (d(x_i, x_s)/d_s)^3\}$ is the tricube weight function. This method of smoothing is similar to the LOWESS method of Cleveland (1979) except that a weighted average instead of a weighted regression is employed.

3. Compute the 'pseudo-residual,' $r_i^* = (p_i^* - \hat{p}_i)$, for each observation. The pseudo-residual replaces the adjusted Anscombe residual in the Poisson regression tree algorithm.

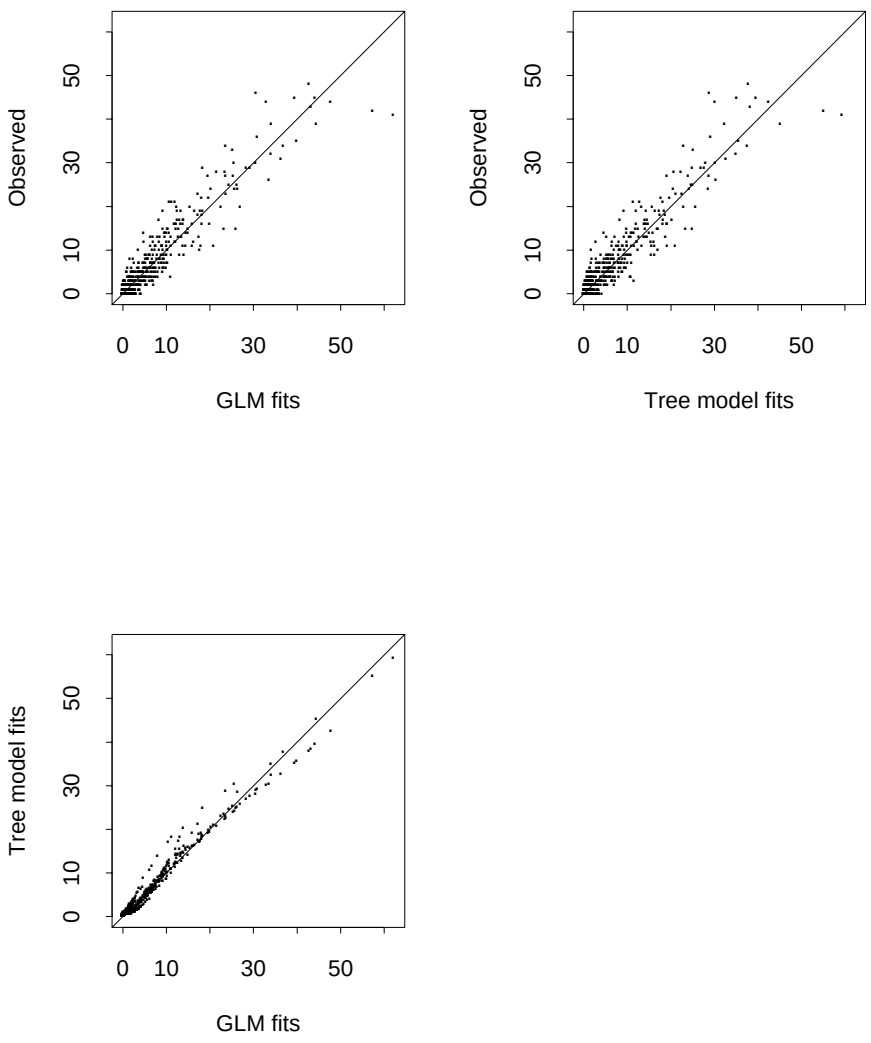

Figure 4: Observed versus fitted values for solder example. The GLM model contains all main effects and all two-factor interactions involving *Opening*, *Solder* and *Mask*.

Table 7: Range of values of covariates for breast cancer data

Covariate	Minimum	Maximum
Age	30	83
Year	58	69
Nodes	0	52

Table 8: Estimated coefficients for two linear logistic models fitted to the breast cancer data

Model 1 (deviance = 328 with 302 df)			Model 2 (deviance = 316 with 302 df)		
Term	Coefficient	t-value	Term	Coefficient	t-value
Constant	1.862	0.70	Constant	2.617	0.97
Age	-0.020	-1.57	Age	-0.024	-1.89
Year	0.010	0.23	Year	0.008	0.19
Nodes	-0.088	-4.47	log(Nodes + 1)	-0.733	-5.76

The value of the smoothing parameter h may be chosen by cross-validation if necessary. Our experience shows, however, that a fixed value between 0.3 and 0.4 is often satisfactory. This is because the pseudo-observations are used here to provide only a preliminary estimate of p that does not have to be very precise.

3.1 Survival following breast cancer surgery

The data in this example come from a study conducted between 1958 and 1970 at the University of Chicago Billings Hospital on the survival of patients who had undergone surgery for breast cancer. There are 306 observations on each of three covariates: *Age* of patient at time of surgery, *Year* (year of surgery minus 1900), and *Nodes* (number of positive axillary nodes detected in the patient). The response variable Y is equal to 1 if the patient survived 5 years or more, and is equal to 0 otherwise. Two hundred twenty-five of the cases had $Y = 1$. Table 7 shows the ranges of values taken by the covariates.

Table 8 gives the coefficient estimates for two linear logistic models fitted to the data. The only difference between the models is that the first uses *Nodes* as a covariate while the second uses $\log(Nodes + 1)$. Only the covariate involving *Nodes* is significant in either model. The residual deviances of the models are 328 and 316 respectively, each with 302 degrees of freedom.

Haberman (1976) finds that the model

$$\log(p/(1-p)) = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Year} + \beta_3 \text{Nodes} + \beta_4 (\text{Nodes})^2 + \beta_5 \text{Age} \times \text{Year} \qquad (2)$$

fits better than the linear logistic Model 1 of Table 8. The residual deviance of (2) is 314 with 300 degrees of freedom. The regression coefficients and t-statistics are given on the left half of Table 9. Except for *Nodes* which is highly significant, all the other covariates are marginally significant (the Bonferroni two-sided t-statistic at the 0.05 simultaneous significance level is 2.58).

Landwehr, Pregibon and Shoemaker (1984) re-analyze these data with the help of graphical diagnostics. Their model replaces the linear and squared terms in *Nodes* in Haberman's model

Table 9: Estimated coefficients for models of Haberman and Landwehr et al.

Haberman (deviance = 314 with 300 df)			Landwehr et al. (deviance = 302 with 299 df)		
Term	Coefficient	t-value	Term	Coefficient	t-value
Constant	35.931	2.62	Constant	77.831	3.29
Age	-0.661	-2.62	Age	-2.868	-2.91
Year	-0.528	-2.43	Year	-0.596	-2.44
Age × Year	0.010	2.54	Age × Year	0.011	2.51
Nodes	-0.175	-4.57	log(Nodes + 1)	-0.756	-5.73
$(\text{Nodes})^2$	0.003	2.61	$(\text{Age})^2$	0.039	2.35
			$(\text{Age})^3$	-0.000	-2.31

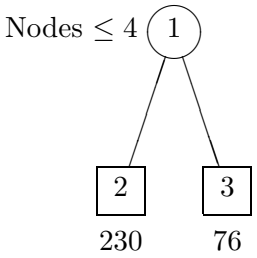

Figure 5: Logistic regression tree for breast cancer data using 10-fold cross-validation. A case goes to the left subnode if the condition at a split is true. The number beneath a terminal node is the learning sample size.

with the terms $\log(1+\text{Nodes})$, $(\text{Age})^2$, and $(\text{Age})^3$. This model has a residual deviance of 302 with 299 degrees of freedom. The estimated coefficients are given on the right half of Table 9. Again the term involving *Nodes* is highly significant and the other terms are marginally significant.

Using *Age, Year,* and *Nodes* as covariates and the smoothing parameter value $h = 0.3$, our method yields the logistic regression tree in Figure 5. It has only one split, on the covariate *Nodes*. The estimated logistic regression coefficients are given in Table 10. None of the t-statistics is significant, although that for *Nodes* is marginally significant in the left subnode. The reason is that much of the significance of *Nodes* is captured in the split. The estimated coefficient for *Nodes* changes, however, from the marginally significant value of -0.289 to the non-significant value of -0.012 as we move from the left subnode to the right subnode. This implies that the survival probability decreases as the value of *Nodes* increases from 0 to 4; for values of *Nodes* greater than 4, survival probability is essentially independent of the covariate.

Figure 6 shows plots of the predicted logit values from the three models. The Haberman and tree models appear to be most similar. Note the outlying point marked by an 'X' in each plot. It represents an 83 year-old patient who was operated in 1958, had 2 positive axillary nodes and died within 5 years. The cubic term $(\text{Age})^3$ in the Landwehr et al. (1984) model causes it to predict a much lower logit value for this case than the other models.

Following Landwehr et al. (1984, p. 69), we plot the estimated survival probability as a function of *Age* for the situations when *Year* = 63 and *Nodes* = 0 or 20. The results are shown in Figure 7. The non-monotonic shapes of the curves for the Landwehr et al. model are due to the cubic term

Table 10: Estimated coefficients for logistic regression tree model in Figure 5

Left subnode (Nodes ≤ 4)			Right subnode (Nodes > 4)		
Term	Coefficient	t-value	Term	Coefficient	t-value
Constant	2.4509	0.734	Constant	0.5117	0.106
Age	-0.0199	-1.267	Age	-0.0378	-1.528
Year	0.0063	0.120	Year	0.0243	0.326
Nodes	-0.2890	-2.255	Nodes	-0.0124	-0.472

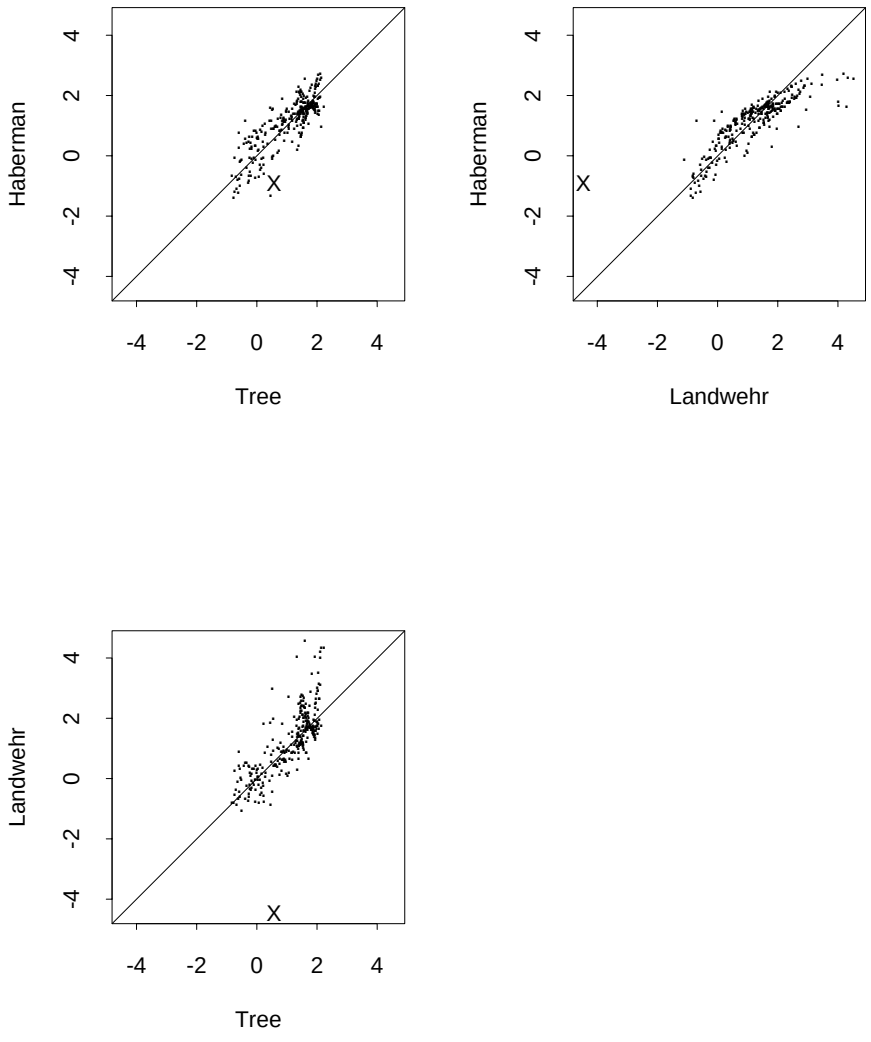

Figure 6: Plots of predicted logit values for breast cancer data according to various models.

in *Age*. Figure 8 shows corresponding plots against *Nodes* for the cases *Year* = 63 and *Age* = 40 or 70. The presence of the quadratic term in *Nodes* is now obvious in the plot for the Haberman model. The plots for the tree-structured show that survival probability decreases monotonically with *Age* and *Nodes*, as might be expected.

When the covariate *Nodes* is replaced by its log-transformed version log(Nodes + 1), the logistic tree method yields a trivial tree with no splits. This suggests that if the log-transformation is used, then the simple linear logistic Model 2 given in Table 8 is adequate. This conclusion is consistent with the earlier observation that the t-statistics corresponding to the nonlinear terms in the Landwehr et al. model are marginally significant at best.

4 Consistency of function estimates

We now give conditions for the consistency of the function estimates in a very general setup. Assume that $(Y_1, X_1), (Y_2, X_2), \ldots, (Y_n, X_n)$ are independent data points, where the response Y_i is real-valued and the regressor X_i is d-dimensional. As before, let $f\{y_i|g(x_i)\}$ be the conditional pdf/pmf of Y_i given $X_i = x_i$. We wish to estimate the function g over a compact set $C \subset R^d$.

Let T_n be a random partition of C (i.e., $C = \cup_{t \in T_n} t$), which is generated by some adaptive recursive partitioning algorithm applied to the data, and it is assumed to consist of polyhedrons having at most M faces, where M is a fixed positive integer. Denote the diameter of a set $t \in T_n$ by $\delta(t)$ (i.e., $\delta(t) = \sup_{x,y \in t} |x - y|$), which is assumed to be positive for each set $t \in T_n$. For $t \in T_n$, let $\bar{X}_t$ denote the average of the X_i's that belong to t. Also, assuming that the function g is m-th order differentiable ($m \geq 0$), write its Taylor expansion around $\bar{X}_t$ as

$$g(x) = \sum_{u \in U} (u!)^{-1} D^u g(\bar{X}_t)(x - \bar{X}_t)^u \ + \ r_t(x, \bar{X}_t) \, .$$

Here $U = \{u | u = (v_1, v_2, \ldots, v_d), [u] \leq m\}$, where $[u] = v_1 + v_2 + \ldots + v_d$ and the v_i's are nonnegative integers. For $u \in U$, D^u is the mixed partial differential operator with index u, $u! = \prod_{i=1}^d v_i!$, and for $x = (z_1, z_2, \ldots, z_d)$, $x^u = \prod_{i=1}^d z_i^{v_i}$ (with the convention that $0! = 1$ and $0^0 = 1$). Let $s(U)$ be the cardinality of the set U. For $X_i \in t$, let Γ_i be the $s(U)$-dimensional column vector with components given by $(u!)^{-1}\{\delta(t)\}^{-[u]}(X_i - \bar{X}_t)^u$, where $u \in U$. Finally, denote by D_t the $s(U) \times s(U)$ matrix defined as $\sum_{X_i \in t} \Gamma_i \Gamma_i^T$, where T indicates transpose. We impose the following conditions which are similar to conditions (a) through (c) in Chaudhuri et al. (1994). A detailed discussion of these conditions is given in Chaudhuri et al. (1993).

Condition 1 $\max_{t \in T_n} \sup_{x \in t} \{\delta(t)\}^{-m} |r_t(x, \bar{X}_t)| \xrightarrow{P} 0$ *as* $n \to \infty$.

Condition 2 *Let* N_t *be the number of* X_i*'s that lie in* t*, and* $N_n = \min_{t \in T_n} \{\delta(t)\}^{2m} N_t$*. Then* $N_n / \log n \xrightarrow{P} \infty$ *as* $n \to \infty$.

Condition 3 *Let* λ_t *be the smallest eigenvalue of* $N_t^{-1} D_t$ *and let* $\lambda_n = \min_{t \in T_n} \lambda_t$*. Then* λ_n *remains bounded away from zero in probability as* $n \to \infty$.

For $\Theta = (\theta_u)_{u \in U}$, define the polynomial $P(x, \Theta, \bar{X}_t)$ in x as

$$P(x, \Theta, \bar{X}_t) \ = \ \sum_{u \in U} \theta_u (u!)^{-1} \{\delta(t)\}^{-[u]} (x - \bar{X}_t)^u \, .$$

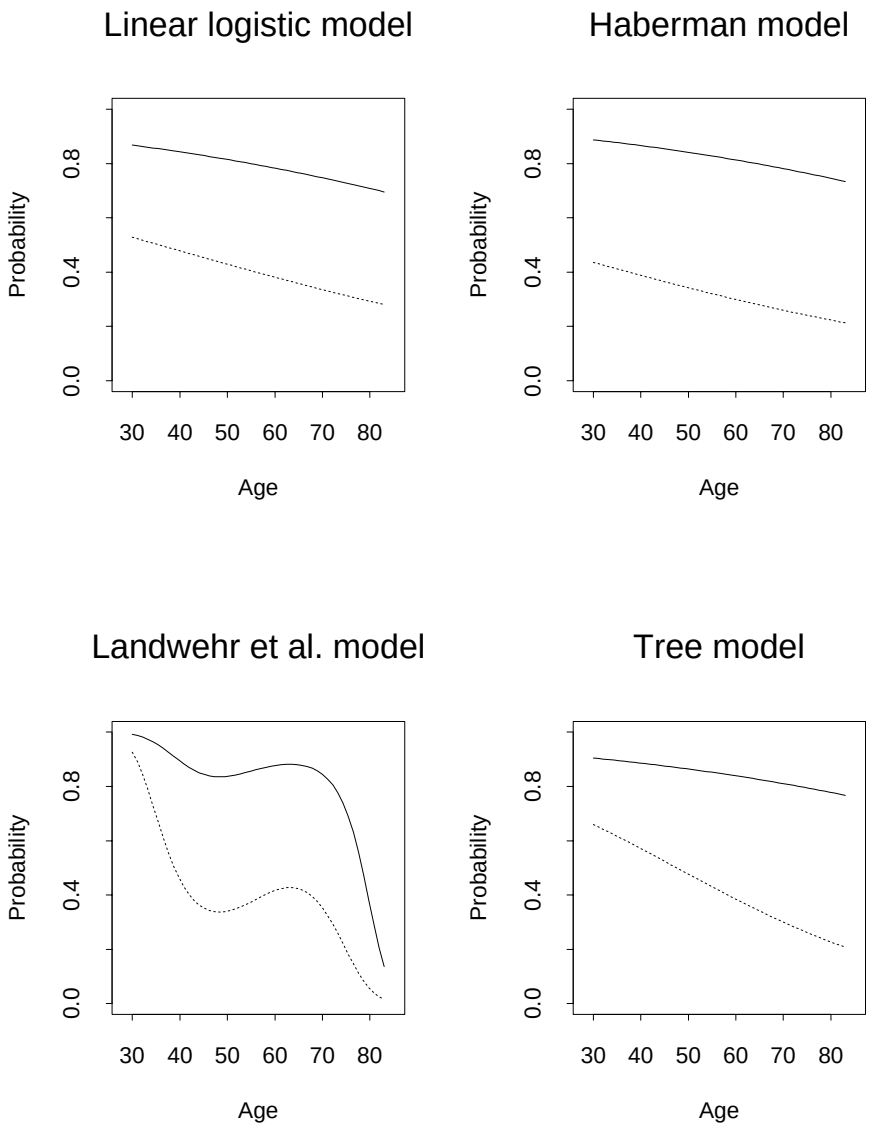

Figure 7: Plots of estimated survival probability as a function of *Age* when *Year* = 63. The solid line corresponds to the cases for which *Nodes* = 0 and the dotted line to *Nodes* = 20.

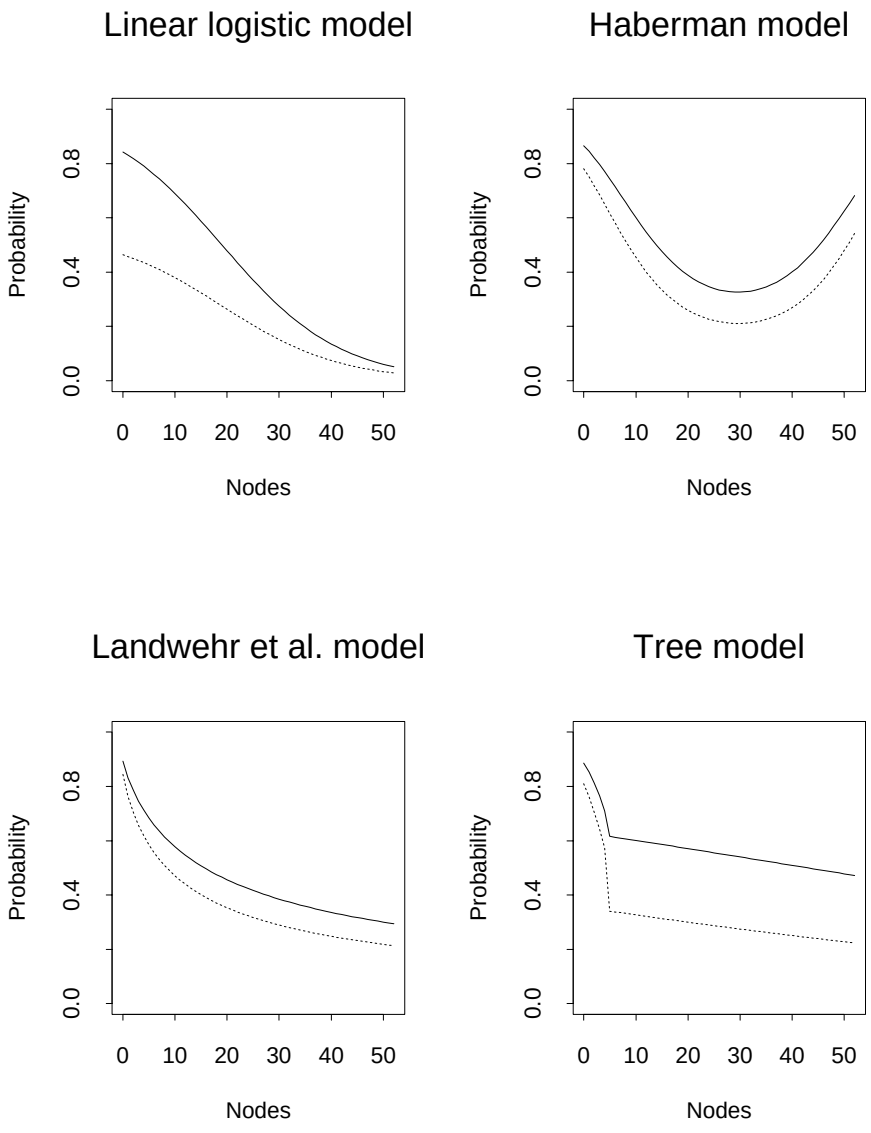

Figure 8: Plots of estimated survival probability as a function of *Nodes* when $Year = 63$. The solid line corresponds to the cases for which $Age = 40$ and the dotted line to $Age = 70$.

Following the estimation procedure described in the previous sections, let $\hat{\Theta}_t$ be the estimate obtained by applying the maximum likelihood technique to the data points (Y_i, X_i) for which $X_i \in t$. In other words,

$$\hat{\Theta}_t \;=\; \arg\max_{\Theta} \prod_{X_i \in t} f\{Y_i | P(X_i, \Theta, \bar{X}_t)\} \; .$$

Condition 3 guarantees that for large sample size, each of the matrices D_t's will be nonsingular and nicely behaved with high probability (cf. Condition (c) in Chaudhuri et al. (1994)). It ensures regularity in the behavior of the Fisher information matrix associated with the finite-dimensional model fitted to the conditional distribution within each set in T_n. Note that we fit a polynomial of a fixed degree with a finite number of coefficients to the data points in any set in T_n.

Finally, we need a Cramér-type regularity condition on the conditional distribution of the response given the regressor. This condition is crucial in establishing desirable asymptotic behavior of our estimates, which are constructed using maximum likelihood.

Condition 4 *Consider the pdf/pmf $f(y|s)$ as a function of two variables so that s is a real-valued parameter varying in a bounded open interval J. Here J is such that as x varies over some open set containing C, $g(x)$ takes its values in J. The support of $f(y|s)$ for any given $s \in J$ is the same, independent of s. The function $\log\{f(y|s)\}$ is three times continuously differentiable w.r.t. s for any given value of y. Let $A(y|s)$, $B(y|s)$ and $H(y|s)$ be the first, second and third derivatives respectively of $\log\{f(y|s)\}$ w.r.t. s. Let Y have pdf/pmf $f(y|s)$. The random variable $A(Y|s)$ has zero mean, and the mean of $B(Y|s)$ is negative and stays away from zero as s varies in J. There exists a nonnegative function $K(y)$ which dominates each of $A(y|s)$, $B(y|s)$ and $H(y|s)$ for all values of $s \in J$ (i.e., $|A(y|s)| \le K(y)$, $|B(y|s)| \le K(y)$ and $|H(y|s)| \le K(y)$). The moment generating function of $K(Y)$, $M(w, s) = E\left[\exp\{wK(Y)\}\right]$, remains bounded as w varies over an open interval around the origin and s varies over J.*

Note that Condition 4 is trivially satisfied when the response Y is binary or, more generally, when its conditional distribution given the regressor is binomial, and s is the logit of the probability parameter such that the probability remains bounded away from 0 and 1. This condition holds whenever the conditional distribution of the response belongs to a standard exponential family (e.g., binomial, Poisson, exponential, gamma, normal, etc.), and s is the natural parameter taking values in a bounded interval. If $f(y|s)$ is a location model with s behaving like a location parameter varying over a bounded parameter space, Condition 4 remains true for several important cases such as the Cauchy or exponential power distribution (see e.g., Box and Tiao (1973)). This condition can be viewed as an extension of Condition (d) in Chaudhuri et al. (1994).

Theorem 1 *Suppose that Conditions 1 through 4 hold. There is a choice of the maximum likelihood estimate $\hat{\Theta}_t$ (possibly a local maximizer of the likelihood) for every $t \in T_n$ such that given any $u \in U$,*

$$\max_{t \in T_n} \sup_{x \in t} |D^u P(x, \hat{\Theta}_t, \bar{X}_t) - D^u g(x)| \xrightarrow{P} 0 \quad \text{as } n \to \infty.$$

This theorem guarantees that there is a choice of the maximum likelihood estimate $\hat{\Theta}_t$ for each $t \in T_n$ so that the resulting piecewise polynomial estimates of the function g and its derivatives are consistent. It may happen that the estimate $\hat{\Theta}_t$ is only a local maximizer of the likelihood instead of being a global maximizer. For instance, the likelihood based on the data points in a set in T_n may have multiple maxima. However, when the conditional distribution of the response given the regressor belongs to a standard exponential family, strict concavity of the loglikelihood

guarantees uniqueness of the maximum likelihood estimate in large samples. In the special case where a constant (i.e., a polynomial of degree zero) is fitted to the data points in each set in T_n using the maximum likelihood approach, Theorem 1 generalizes the consistency result for piecewise constant tree-structured regression estimates discussed in Breiman et al. (1984).

The piecewise polynomial estimates of g and its derivatives are not continuous everywhere in the regressor space. Smooth estimates, which can be constructed by combining the polynomial pieces by smooth weighted averaging, will be consistent provided the weight functions are chosen properly. Theorem 2 in Chaudhuri et al. (1994) describes a way of constructing families of smooth weight functions that give smooth and consistent estimates of g and its derivatives. Some examples for smoothing estimates from Poisson and logistic regression trees are given in Chaudhuri et al. (1993).

Remark 3. The results in this section are very general and hence are not specific to any particular partitioning algorithm. The main difficulty with applying them to a given algorithm lies in the complexity of algorithmic details such as the choice of splitting rule, pruning method, etc. This is a problem associated with any nontrivial adaptive recursive partitioning algorithm that has its own particular set of features and tuning parameters. (See page 327 of Breiman et al. (1984) for a discussion of similar issues in the context of tree-structured classification.)

Acknowledgements

The authors are grateful to Dr. Michael Schwarz of the German Cancer Research Center for his permission to use the NNM data set used in Example 2.1.1. Thanks are also due to Dr. Anup Dewanji of Indian Statistical Institute and Dr. Suresh Moolgavkar of Fred Hutchinson Cancer Research Center for their help in making this data set available to the authors.

Chaudhuri's research was partially supported by a grant from the Indian Statistical Institute. Loh's research was supported in part by U. S. Army Research Office grants DAAL03-91-G-0111 and DAAH04-94-G-0042 and National Science Foundation grant DMS-9304378.

Appendix: proofs

We give a brief sketch of the proof of Theorem 1. More details can be found in Chaudhuri et al. (1993). We begin by giving some preliminary results. Unless stated otherwise, all vectors are assumed to be column vectors. Let Θ_t^* denote the $s(U)$-dimensional vector with typical entry $\{\delta(t)\}^{[u]} D^u g(\bar{X}_t)$ where $u \in U$. Then $P(x, \Theta_t^*, \bar{X}_t)$ is the Taylor polynomial of $g(x)$ expanded around $\bar{X}_t$.

Lemma 1 *Under Conditions 1, 2 and 4, we have*

$$\max_{t \in T_n} N_t^{-1} \{\delta(t)\}^{-m} \left| \sum_{X_i \in t} \left[A\{Y_i | P(X_i, \Theta_t^*, \bar{X}_t)\} \right] \Gamma_i \right| \xrightarrow{P} 0 \text{ as } n \to \infty \, .$$

Proof: First observe that a straightforward application of the mean value theorem of differential calculus yields

$$N_t^{-1} \{\delta(t)\}^{-m} \sum_{X_i \in t} \left[A\{Y_i | P(X_i, \Theta_t^*, \bar{X}_t)\} \right] \Gamma_i$$

$$
\begin{aligned}
&= N_t^{-1}\{\delta(t)\}^{-m} \sum_{X_i \in t} [A\{Y_i|g(X_i)\}]\,\Gamma_i \\
&\quad - N_t^{-1}\{\delta(t)\}^{-m} \sum_{X_i \in t} \{r_t(X_i, \bar{X}_t) B(Y_i|Z_i)\}\Gamma_i \qquad (3)
\end{aligned}
$$

where Z_i is a random variable that lies between $P(X_i, \Theta_t^*, \bar{X}_t)$ and $g(X_i)$. Because of Condition 4, the conditional mean of $A\{Y|g(X)\}$ given $X = x$ is zero, and if we denote its conditional moment generating function by $M_1(w|x)$, there exist constants $k_1 > 0$ and $\rho_1 > 0$ such that $M_1(w|x) \leq 2\exp(k_1 w^2/2)$ for all $x \in C$ and $0 \leq w \leq \rho_1$ (see the arguments at the beginning of Lemma 12.27 in Breiman et al. (1984)). Recall that each set in T_n is a polyhedron in R^d having at most M faces. The fundamental combinatorial result of Vapnik and Chervonenkis (1971) (Dudley 1978, Section 7) implies that there exists a collection $\mathcal{C}$ of subsets of the set $\{X_1, X_2, \ldots, X_n\}$ such that $\#(\mathcal{C}) \leq (2n)^{M(d+2)}$, and for any polyhedron t with at most M faces, there is a set $t^* \in \mathcal{C}$ with the property that $X_i \in t$ if and only if $X_i \in t^*$. By Condition 2 and the arguments used in handling the "variance term" in the proof of Theorem 1 in Chaudhuri et al. (1994), it can be shown that

$$
\max_{t \in T_n} \{\delta(t)\}^{-m} N_t^{-1} \left| \sum_{X_i \in t} [A\{Y_i|g(X_i)\}]\,\Gamma_i \right| \xrightarrow{P} 0 \text{ as } n \to \infty\,.
$$

Further, using Conditions 1, 2 and 4,

$$
\begin{aligned}
&\max_{t \in T_n} N_t^{-1}\{\delta(t)\}^{-m} \left| \sum_{X_i \in t} \{r_t(X_i, \bar{X}_t) B(Y_i|Z_i)\}\,\Gamma_i \right| \\
&\quad \leq \left[\max_{t \in T_n} \{\delta(t)\}^{-m} \sup_{x \in t} |r_t(x, \bar{X}_t)| \right] \max_{t \in T_n} N_t^{-1} \sum_{X_i \in t} K(Y_i)|\Gamma_i| \\
&\quad \xrightarrow{P} 0 \text{ as } n \to \infty.
\end{aligned}
$$

This proves the lemma.

Lemma 2 *Let $\gamma(t)$ denote the smallest eigenvalue of the $s(U) \times s(U)$ matrix*

$$
- N_t^{-1} \sum_{X_i \in t} \left[B\{Y_i|P(X_i, \Theta_t^*, \bar{X}_t)\}\right] \Gamma_i \Gamma_i^T.
$$

Define $\gamma_n = \min_{t \in T_n} \gamma(t)$. Then, under Conditions 1 through 4, γ_n remains positive and bounded away from zero in probability as $n \to \infty$.

Proof: The mean value theorem of differential calculus yields

$$
\begin{aligned}
&N_t^{-1} \sum_{X_i \in t} \left[B\{Y_i|P(X_i, \Theta_t^*, \bar{X}_t)\}\right] \Gamma_i \Gamma_i^T \\
&\quad = N_t^{-1} \sum_{X_i \in t} \left[B\{Y_i|g(X_i)\} - \psi(X_i)\right] \Gamma_i \Gamma_i^T + N_t^{-1} \sum_{X_i \in t} \psi(X_i) \Gamma_i \Gamma_i^T \\
&\qquad - N_t^{-1} \sum_{X_i \in t} \{r_t(X_i, \bar{X}_t) H(Y_i|V_i)\} \Gamma_i \Gamma_i^T, \qquad (4)
\end{aligned}
$$

where $\psi(x)$ is the conditional mean of $B\{Y|g(X)\}$ given $X = x$, and V_i is a random variable that falls between $g(X_i)$ and $P(X_i, \Theta_t^*, \bar{X}_t)$. It follows from Conditions 3 and 4 that if $\eta_n = \min_{t \in T_n} \eta(t)$,

where $\eta(t)$ is the smallest eigenvalue of the matrix $-N_t^{-1}\sum_{X_i\in t}\psi(X_i)\Gamma_i\Gamma_i^T$, then η_n remains positive and bounded away from zero in probability as $n\to\infty$. On the other hand, the first term on the right of (4) can be handled in the same way as the first term on the right of (3) in the proof of Lemma 1 to yield

$$\max_{t\in T_n} N_t^{-1}\left|\sum_{X_i\in t}\left[B\{Y_i|g(X_i)\}\ -\ \psi(X_i)\right]\Gamma_i\Gamma_i^T\right| \ \stackrel{P}{\to}\ 0 \text{ as } n\to\infty\ .$$

Finally, using Conditions 1, 2 and 4, and arguments similar to those employed to treat the second term on the right of (3) in the proof of Lemma 1, we obtain the following result for the third term on the right of (4):

$$\begin{aligned}
&\max_{t\in T_n} N_t^{-1}\left|\sum_{X_i\in t}\{r_t(X_i,\bar{X}_t)H(Y_i|V_i)\}\Gamma_i\Gamma_i^T\right| \\
&\quad\leq\quad \left\{\max_{t\in T_n}\sup_{x\in t}|r_t(x,\bar{X}_t)|\right\}\max_{t\in T_n} N_t^{-1}\sum_{X_i\in t}K(Y_i)|\Gamma_i\Gamma_i^T| \\
&\quad\stackrel{P}{\to}\quad 0 \text{ as } n\to\infty.
\end{aligned}$$

This completes the proof of the lemma.

Proof of Theorem 1: First note that the assertion in the Theorem will follow if we show that there exist choices for the maximum likelihood estimates $\hat{\Theta}_t$'s such that

$$\max_{t\in T_n}\{\delta(t)\}^{-m}|\hat{\Theta}_t\ -\ \Theta_t^*|\ \stackrel{P}{\to}\ 0 \text{ as } n\to\infty\ .$$

For $t\in T_n$, let $l_t(\Theta)$ denote the loglikelihood based on the observations (Y_i,X_i) such that $X_i\in t$. That is, $l_t(\Theta)=\sum_{X_i\in t}\log\left[f\{Y_i|P(X_i,\Theta,\bar{X}_t)\}\right]$. Given $\rho>0$, define $E_t(\rho)$ to be the event:

> $l_t(\Theta)$ is concave in a neighborhood of Θ_t^* with radius $\{\delta(t)\}^m\rho$ (i.e., for Θ satisfying $\{\delta(t)\}^{-m}|\Theta-\Theta_t^*|\leq\rho$), and it has a (possibly local) maximum in the interior of this neighborhood.

The occurrence of this event implies that the maximum likelihood equation obtained by differentiating $l_t(\Theta)$ w.r.t. Θ will have a root $\hat{\Theta}_t$ such that $\{\delta(t)\}^{-m}|\hat{\Theta}_t-\Theta_t^*|<\rho$. A Taylor expansion of $l_t(\Theta)$ around Θ_t^* yields

$$\begin{aligned}
l_t(\Theta) \quad=\quad & l_t(\Theta_t^*)\ +\ \sum_{X_i\in t}(\Theta-\Theta_t^*)^T\Gamma_i A\{Y_i|P(X_i,\Theta_t^*,\bar{X}_t)\} \\
&+(1/2)\sum_{X_i\in t}(\Theta-\Theta_t^*)^T\Gamma_i\Gamma_i^T B\{Y_i|P(X_i,\Theta_t^*,\bar{X}_t)\}(\Theta-\Theta_t^*) \\
&+(1/6)\sum_{X_i\in t}\{(\Theta-\Theta_t^*)^T\Gamma_i\}^3 H(Y_i|W_i), \qquad (5)
\end{aligned}$$

where W_i is a random variable lying between $P(X_i,\Theta_t^*,\bar{X}_t)$ and $P(X_i,\Theta,\bar{X}_t)$. For the third term on the right of (5), note that the Γ_i's are bounded vectors. Also, for Θ in a sufficiently small neighborhood of Θ_t^*, we have $\sum_{X_i\in t}|H(Y_i|W_i)|\leq\sum_{X_i\in t}K(Y_i)$ in view of Condition 4. It follows from Lemmas 1 and 2 and the arguments used in their proofs that there exists $\rho_3>0$ such that whenever $\rho\leq\rho_3$, we must have $\Pr(\cap_{t\in T_n}E_t(\rho))\to 1$ as $n\to\infty$. This proves the theorem.

References

Ahn, H. and Loh, W.-Y. (1994). Tree-structured proportional hazards regression modeling, *Biometrics* **50**: 471–485.

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*, Addison-Wesley, Reading.

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

Buja, A., Duffy, D., Hastie, T. and Tibshirani, R. (1991). Comment on "Multivariate adaptive regression splines", *Annals of Statistics* **19**: 93–99.

Chambers, J. M. and Hastie, T. J. (1992). An appetizer, *in* J. M. Chambers and T. J. Hastie (eds), *Statistical Models in S*, Wadsworth & Brooks/Cole, Pacific Grove, pp. 1–12.

Chaudhuri, P. and Dewanji, A. (1995). On a likelihood based approach in nonparametric smoothing and cross-validation, *Statistics and Probability Letters* **22**: 7–15.

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

Chaudhuri, P., Lo, W.-D., Loh, W.-Y. and Yang, C.-C. (1993). Generalized regression trees: Function estimation via recursive partitioning and maximum likelihood, *Technical Report 903*, University of Wisconsin, Madison, Department of Statistics.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatter plots, *Journal of the American Statistical Association* **74**: 829–836.

Comizzoli, R. B., Landwehr, J. M. and Sinclair, J. D. (1990). Robust materials and processes: Key to reliability, *AT&T Technical Journal* **69**: 113–128.

Cox, D. D. and O'Sullivan, F. (1990). Asymptotic analysis of penalized likelihood and related estimators, *Annals of Statistics* **18**: 1676–1695.

Dudley, R. M. (1978). Central limit theorems for empirical measures, *Annals of Probability* **6**: 899–929. Corr: **7**, 909–911.

Fowlkes, E. B. (1987). Some diagnostics for binary logistic regression via smoothing, *Biometrika* **74**: 503–515.

Friedman, J. H. (1991). Multivariate adaptive regression splines (with discussion), *Annals of Statistics* **19**: 1–67.

Gu, C. (1990). Adaptive spline smoothing in non-Gaussian regression models, *Journal of the American Statistical Association* **85**: 801–807.

Haberman, S. J. (1976). Generalized residuals for log-linear models, *Proceedings of the 9th International Biometrics Conference*, Biometric Society, Boston, pp. 104–122.

Hastie, T. J. and Pregibon, D. (1992). Generalized linear models, *in* J. M. Chambers and T. J. Hastie (eds), *Statistical Models in S*, Wadsworth & Brooks/Cole, Pacific Grove, pp. 195–248.

Hastie, T. J. and Tibshirani, R. J. (1986). Generalized additive models (with discussion), *Statistical Science* **1**: 297–310.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*, Chapman and Hall, London.

Landwehr, J. M., Pregibon, D. and Shoemaker, A. C. (1984). Graphical methods for assessing logistic models, *Journal of the American Statistical Association* **79**: 61–83.

Levene, H. (1960). Robust tests for equality of variances, *in* I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow and H. B. Mann (eds), *Contributions to Probability and Statistics*, Stanford University Press, Stanford, pp. 278–292.

Lo, W.-D. (1993). *Logistic Regression Trees*, PhD thesis, University of Wisconsin, Madison.

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

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, Chapman and Hall, London.

Moolgavkar, S. H., Luebeck, E. G., de Gunst, M., Port, R. E. and Schwarz, M. (1990). Quantitative analysis of enzyme-altered foci in rat hepatocarcinogenesis experiments—I. Single agent regimen, *Carcinogenesis* **11**: 1271–1278.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models, *Journal of the Royal Statistical Society, Series A* **135**: 370–384.

O'Sullivan, F., Yandell, B. S. and Raynor, W. J. J. (1986). Automatic smoothing of regression functions in generalized linear models, *Journal of the American Statistical Association* **81**: 96–103.

Pierce, D. A. and Schafer, D. W. (1986). Residuals in generalized linear models, *Journal of the American Statistical Association* **81**: 977–986.

Sonquist, J. N. (1970). Multivariate model building, *Technical report*, Institute for Social Research, University of Michigan.

Sonquist, J. N., Baker, E. L. and Morgan, J. A. (1973). Searching for structure, *Technical report*, Institute for Social Research, University of Michigan.

Staniswalis, J. G. (1989). The kernel estimate of a regression function in likelihood-based models, *Journal of the American Statistical Association* **84**: 276–283. Corr: **85**, 1182.

Stone, C. J. (1986). The dimensionality reduction principle for generalized additive models, *Annals of Statistics* **14**: 590–606.

Stone, C. J. (1991a). Asymptotics for doubly flexible log-spline response models, *Annals of Statistics* **19**: 1832–1854.

Stone, C. J. (1991b). Comment on "Multivariate adaptive regression splines", *Annals of Statistics* **19**: 113–115.

Vapnik, V. N. and Chervonenkis, A. Y. (1971). On the uniform convergence of relative frequencies of events to their probabilities, *Theory of Probability and Its Applications* **16**: 264–280.

Yang, C.-C. (1993). *Tree-structured Poisson Regression*, PhD thesis, University of Wisconsin, Madison.