Bernoulli **8**(5), 2002, 561–576

Nonparametric estimation of conditional quantiles using quantile regression trees

PROBAL CHAUDHURI[1] and WEI-YIN LOH[2]

[1]*Division of Theoretical Statistics and Mathematics, Indian Statistical Institute, 203 B.T. Road, Calcutta 700035, India. E-mail: probal@isical.ac.in*
[2]*Department of Statistics, University of Wisconsin, 1210 West Dayton Street, Madison WI 53706, USA. E-mail: loh@stat.wisc.edu*

A nonparametric regression method that blends key features of piecewise polynomial quantile regression and tree-structured regression based on adaptive recursive partitioning of the covariate space is investigated. Unlike least-squares regression trees, which concentrate on modelling the relationship between the response and the covariates at the centre of the response distribution, our quantile regression trees can provide insight into the nature of that relationship at the centre as well as the tails of the response distribution. Our nonparametric regression quantiles have piecewise polynomial forms, where each piece is obtained by fitting a polynomial quantile regression model to the data in a terminal node of a binary decision tree. The decision tree is constructed by recursively partitioning the data based on repeated analyses of the residuals obtained after model fitting with quantile regression. One advantage of the tree structure is that it provides a simple summary of the interactions among the covariates. The asymptotic behaviour of piecewise polynomial quantile regression estimates and the associated derivative estimates are studied under appropriate regularity conditions. The methodology is illustrated with an example on the incidence rates of mumps in the United States.

Keywords: derivative estimate; GUIDE algorithm; piecewise polynomial estimates; recursive partitioning; tree-structured regression; uniform asymptotic consistency; Vapnik–Chervonenkis class

1. Introduction

For $0 < \alpha < 1$, quantile regression analysis focuses on the conditional αth quantile of the response Y given the covariate vector $\mathbf{X} = (X_1, X_2, \ldots, X_k)$. Unlike usual regression analysis, which focuses only on the conditional mean (i.e., the 'centre' of the conditional distribution) of Y given $\mathbf{X}$, quantile regression is capable of providing insight into the centre as well as the lower and upper tails of the conditional distribution of the response with varying choices of α. As a result, quantile regression is quite effective as a tool for exploring and modelling the nature of dependence of a response on the covariates when the covariates have different effects on different parts of the conditional distribution of the response. Such situations occur in many econometric problems. For example, a covariate may have very different types of effect on high-, low- and middle-income groups. This is why quantile regression has become a popular methodology for the analysis of income data – see Hogg (1975) and Chaudhuri *et al*. (1997). Buchinsky (1994) used quantile regression to carry out an extensive analysis of changes in the US wage structure during 1963–87. In marketing

1350–7265

studies, where covariates may have different effects on high-, medium- and low-consumption groups, quantile regression can be useful in understanding the nature of the dependence between the response and the covariates. Hendricks and Koenker (1992) used quantile regression to study variations in electricity consumption over time.

Let $g_\alpha(\mathbf{x})$ denote the conditional αth quantile of Y given $\mathbf{X} = \mathbf{x}$. Many authors have considered various nonparametric methods for estimating a smooth quantile function from the data $(Y_1, \mathbf{X}_1)$, $(Y_2, \mathbf{X}_2)$, $\ldots$, $(Y_n, \mathbf{X}_n)$ – see Cheng (1983; 1984), Janssen and Veraverbeke (1987), Lejeune and Sarda (1998), Truong (1989), Dabrowska (1992), Fan *et al*. (1994), Koenker *et al*. (1994) and Welsh (1996). Chaudhuri (1991a; 1991b) studied in detail local polynomial estimates of a smooth conditional quantile function and discussed their asymptotic properties. Such estimates were subsequently used by Chaudhuri *et al*. (1997) in average derivative quantile regression, which is a useful methodology for nonparametric and semi-parametric modelling. They demonstrated how local polynomial estimates of a smooth regression quantile function can be used as an effective device for estimating the parametric components in semi-parametric models such as monotone transformation models, projection pursuit models and monotone single-index models that are quite popular in the econometric literature – see Han (1987), Härdle and Stoker (1989), Newey and Stoker (1993), Powell *et al*. (1989), Samarov (1993) and Sherman (1993).

Tree-structured methods and recursive partitioning algorithms for constructing piecewise polynomial estimates using local least-squares and local maximum likelihood techniques were studied by Chaudhuri *et al*. (1994; 1995), who gave some arguments in favour of the methodology – see also Breiman *et al*. (1984), who considered piecewise constant estimates of regression functions. Firstly, the decision tree produced by the data can describe the overall model complexity, including, for example, interactions among the covariates. This allows the polynomial model in each terminal partition to be kept simple for easy interpretation and analytic study. Secondly, the adaptive nature of the recursive partitioning algorithm allows for variation in the degree of smoothing across the covariate space so that the terminal partitions may have different sizes and contain different numbers of data points. This helps to cope with heteroscedasticity in the data and with the variable smoothness of the function being estimated in different regions of the covariate space.

Piecewise constant median regression trees constructed using least absolute deviations were considered by Breiman *et al*. (1984) as a robust alternative to least-squares regression trees. Our goal in this paper is to combine some fundamental ideas in piecewise polynomial quantile regression with recursive partitioning and tree-structured methods for constructing nonparametric estimates of conditional quantile functions and their derivatives. We also study the statistical performance of such estimates. Our *quantile regression tree* can be an effective exploratory data-analytic tool for empirical model building as well as for model checking and diagnostics.

Piecewise polynomial regression tree models have two advantages over piecewise constant regression tree models. Firstly, the latter trees tend to be very large and hence hard to interpret. The size of a piecewise polynomial regression tree, on the other hand, can be altered by changing the form of the polynomials fitted at the nodes. Secondly, the greater flexibility of polynomials over constants often translates to higher estimation accuracy of the piecewise polynomial tree models.

Another desirable feature of a piecewise polynomial estimate of an unknown function is that the coefficients of the locally fitted polynomials provide estimates of the derivatives of that function. This is useful for gaining insight into the shape and the geometry of the unknown function as well as for statistical estimation of parametric components in semi-parametric models, where those parametric components arise as some form of average multidimensional slope (gradient vector) or average Hessian matrix associated with the unknown function – see Härdle and Stoker (1989), Samarov (1993) and Chaudhuri *et al*. (1997).

The rest of this paper is organized as follows. Section 2 describes the piecewise polynomial estimate of a conditional quantile function and resulting derivative estimates. We establish uniform consistency of these estimates under appropriate regularity conditions. In the case of piecewise constant estimates of a conditional median function (constructed using least absolute deviations regression trees), asymptotic consistency was conjectured by Breiman *et al*. (1984, Section 8.11). Our result thus proves and generalizes their conjecture. Section 3 illustrates the ideas using a data set on mumps. Appendix A contains the proof of our theorem and Appendix B gives a brief discussion of the computational algorithm.

2. Description and large-sample performance of quantile regression and derivative estimates

We begin by introducing some notation. We assume that $(Y_1, \mathbf{X}_1)$, $(Y_2, \mathbf{X}_2)$, $\ldots$, $(Y_n, \mathbf{X}_n)$ are independent data points, where the response Y is real-valued and the regressor $\mathbf{X}$ is d-dimensional. Let the conditional αth quantile function of Y given $\mathbf{X} = \mathbf{x}$ be $g_\alpha(\mathbf{x})$, which is to be estimated on a subset C of the d-dimensional Euclidean space based on the data. We denote by T_n a random partition of C (i.e., $C = \cup_{t \in T_n} t$) generated by some adaptive recursive partitioning algorithm applied to the data. T_n is assumed to consist of only polyhedrons having at most M faces, where M is a fixed positive integer. We also assume that the diameter $\delta(t)$ of the set t (i.e., $\delta(t) = \sup\{|x - z| : x, z \in t\}$) is positive for each $t \in T_n$. Let $\overline{\mathbf{X}}_t$ denote the average of the $\mathbf{X}_i$ that belong to t. The conditional quantile function $g_\alpha(\mathbf{x})$ is assumed to be mth-order differentiable ($m \geqslant 0$), and we write its Taylor expansion around $\overline{\mathbf{X}}_t$ as

$$g_\alpha(\mathbf{x}) = \sum_{\mathbf{u} \in U} (\mathbf{u}!)^{-1} \{D^{\mathbf{u}} g_\alpha(\overline{\mathbf{X}}_t)\}(\mathbf{x} - \overline{\mathbf{X}}_t)^{\mathbf{u}} + r_t(\mathbf{x}, \overline{\mathbf{X}}_t).$$

Here U is the collection of all d-tuples of non-negative integers of the form $\mathbf{u} = (u_1, u_2, \ldots, u_d)$ such that $[\mathbf{u}] \leqslant m$, where we define $[\mathbf{u}] = u_1 + u_2 + \ldots + u_d$. For $\mathbf{u} \in U$, let $D^{\mathbf{u}}$ denote the mixed partial differential operator with index $\mathbf{u}$ and define $\mathbf{u}! = \prod_{i=1}^{d} u_i!$. For $\mathbf{x} = (x_1, x_2, \ldots, x_d)$, define $\mathbf{x}^{\mathbf{u}} = \prod_{i=1}^{d} x_i^{u_i}$. By convention, $0! = 0^0 = 1$. Let $s(U)$ denote the cardinality of the set U. For $\mathbf{X}_i \in t$, let $\mathbf{\Gamma}_i$ be the $s(U)$-dimensional column vector with components of the form $(\mathbf{u}!)^{-1}\{\delta(t)\}^{-[\mathbf{u}]}(\mathbf{X}_i - \overline{\mathbf{X}}_t)^{\mathbf{u}}$, where $\mathbf{u} \in U$. The $s(U) \times s(U)$ matrix $\sum_{\mathbf{X}_i \in t} \mathbf{\Gamma}_i \mathbf{\Gamma}_i^{\mathrm{T}}$ will be denoted by $\mathbf{D}_t$. From now on all vectors in this paper will be column

vectors unless otherwise specified, and the superscript T denotes the transpose of a vector or matrix.

For an $s(U)$-dimensional vector $\boldsymbol{\Theta} = (\theta_{\mathbf{u}})_{\mathbf{u} \in u}$, define the polynomial $P(\mathbf{x}, \boldsymbol{\Theta}, \overline{\mathbf{X}}_t)$ in $\mathbf{x}$ as

$$P(\mathbf{x}, \boldsymbol{\Theta}, \overline{\mathbf{X}}_t) = \sum_{\mathbf{u} \in u} \theta_{\mathbf{u}} (\mathbf{u}!)^{-1} \{\delta(t)\}^{-[\mathbf{u}]} (\mathbf{x} - \overline{\mathbf{X}}_t)^{\mathbf{u}}.$$

Let $\hat{\boldsymbol{\Theta}}_t^{(\alpha)}$ be the vector of coefficients of the polynomial fitted to the data points $(Y_i, \mathbf{X}_i)$ for which $\mathbf{X}_i \in t$. That is,

$$\hat{\boldsymbol{\Theta}}_t^{(\alpha)} = \arg\min_{\Theta} \sum_{\mathbf{X}_i \in t} \{|Y_i - P(\mathbf{X}_i, \boldsymbol{\Theta}, \overline{\mathbf{X}}_t)| + (2\alpha - 1)[Y_i - P(\mathbf{X}_i, \boldsymbol{\Theta}, \overline{\mathbf{X}}_t)]\}. \tag{1}$$

For $x \in t \in T_n$, our piecewise polynomial estimate of the conditional αth quantile function $g_\alpha(\mathbf{x})$ is $P(\mathbf{x}, \hat{\boldsymbol{\Theta}}_t^{(\alpha)}, \overline{\mathbf{X}}_t)$.

In a different context, asymptotic properties of kernel weighted local polynomial regression estimates are discussed in Wand and Jones (1995) and Fan and Gijbels (1996). Chaudhuri (1991a; 1991b) studied the asymptotics of local polynomial quantile regression estimates. A major technical barrier in studying the asymptotic properties of our piecewise polynomial quantile regression estimates is the complexity caused by the random nature of the partitions produced by the adaptive and recursive algorithm. In the proofs given in Appendix A, we use a well-known combinatorial result of Vapnik and Chervonenkis (1971) to cope with this problem.

The algorithm we use to analyse data in practice – see Appendix B and Loh (2002) – yields piecewise polynomial estimates that closely resemble rectangular kernel weighted local polynomial estimates. The support sets of these rectangular kernels are generated by our partitioning algorithm. The rectangular nature of the partition sets is a consequence of the splitting procedure used at each stage of our algorithm, which is based on a single 'best' variable. This makes the resulting tree and the partition sets easier to interpret and comprehend. Further, the rectangular partition sets facilitate numerical computation as well as asymptotic analysis. The derivation of the large-sample properties of our piecewise polynomial estimates requires that our partition sets be polyhedrons with a bounded number of faces, and clearly rectangles in a d-dimensional space satisfy this requirement.

We now state some conditions that are required to guarantee consistency of the piecewise polynomial estimates of $g_\alpha(\mathbf{x})$ and its derivatives as the sample size increases. These conditions are related to the asymptotic behaviour of the partition T_n and regressors $\mathbf{X}_i$, and they are similar to some of the conditions assumed in Chaudhuri *et al*. (1994; 1995).

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

Condition 2. *Let* N_t *be the number of* $\mathbf{X}_i$ *that lie in* t *and* $N_n = \min\{\{\delta(t)\}^{2m} N_t : t \in T_n\}$. *Then* $\log n / N_n \to 0$ *in probability as* $n \to \infty$.

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

Condition 1 ensures the asymptotic validity of the polynomial approximation of the conditional αth quantile function in each set of the partition T_n. When $\max\{\delta(t) : t \in T_n\} \to 0$ in probability as $n \to \infty$ (i.e., when the sets in the partition T_n shrink with increasing sample size), this condition is automatically satisfied if $g_\alpha(\mathbf{x})$ is continuously differentiable in C up to order m. Condition 2 guarantees that asymptotically there will be sufficiently many data points in each $t \in T_n$, while Condition 3 ensures that asymptotically the covariates $\mathbf{X}_i$ are properly distributed in each $t \in T_n$ so that the optimization problem that arises in piecewise polynomial quantile regression is sufficiently regular and does not suffer from singularities in the covariate distributions.

The next condition is about the conditional distribution of the response Y given the regressor $\mathbf{X}$.

Condition 4. *The conditional distribution of Y given $\mathbf{X} = \mathbf{x}$ has a density $f(y|\mathbf{x})$ which remains uniformly bounded and bounded away from zero as $\mathbf{x}$ varies in the set C and y varies in the interval $(g_\alpha(\mathbf{x}) - \epsilon,\ g_\alpha(\mathbf{x}) + \epsilon)$ for some fixed $\epsilon > 0$. In other words,*

$$0 < \inf_{\mathbf{x} \in C} \inf\{f(y|\mathbf{x}) : |y - g_\alpha(\mathbf{x})| < \epsilon\}$$

$$0 < \sup_{\mathbf{x} \in C} \sup\{f(y|\mathbf{x}) : |y - g_\alpha(\mathbf{x})| < \epsilon\}$$

$$< \infty.$$

With these conditions in hand, we can now state the main result on the uniform consistency of our piecewise polynomial estimate of the conditional quantile function and its derivatives on the set C. The proof is given in Appendix A.

Theorem 1. *The minimization problem defining $\hat{\boldsymbol{\Theta}}_t^{(\alpha)}$ has a solution for each $t \in T_n$. Further, under Conditions 1–4, there exist solutions $\hat{\boldsymbol{\Theta}}_t^{(\alpha)}$ for all $t \in T_n$ such that*

$$\max_{t \in T_n} \sup_{x \in t} |D^{\mathbf{u}} P(\mathbf{x}, \hat{\boldsymbol{\Theta}}_t^{(\alpha)}, \overline{\mathbf{X}}_t) - D^{\mathbf{u}} g_\alpha(\mathbf{x})| \to 0$$

in probability for any $\mathbf{u} \in u$ as $n \to \infty$.

In the special case of piecewise constant median regression trees, asymptotic consistency of the estimate of the conditional median function is established in Chaudhuri (2000) under appropriate regularity conditions. The piecewise polynomial estimates of $g_\alpha(\mathbf{x})$ and its derivatives may not be continuous at the boundaries of the sets in the partition. Although we have not implemented it here, smooth and asymptotically consistent estimates can be constructed by gluing the polynomial pieces with smooth weighted averaging as in Chaudhuri *et al*. (1994, Section 3).

It should be noted that the asymptotic result in Theorem 1 is very general in nature and is not specific to a particular recursive partitioning algorithm. Each algorithm will have its

own features with respect to splitting rule, pruning method, cross-validation strategy, etc. Nevertheless, as long as Conditions 1–3 are satisfied for the partitions generated by the algorithm, we will have asymptotic convergence of the piecewise polynomial quantile regression estimates provided that Condition 4 holds for the conditional distribution of the response given the covariates.

3. Example: Incidence rates of mumps

We illustrate our method with some data on the incidence of mumps in the 48 contiguous states of the United States (excluding the District of Columbia) from 1953 to 1989. The data were the focus of a special poster session sponsored by the Statistical Graphics Section of the American Statistical Association at its 1991 annual meeting. There are 1523 observations and three predictor variables (some states did not report data for some years). The dependent variable (y) is the natural logarithm of the number of mumps cases reported per million population in each state (the population figures are based on the 1970 Census). The predictor variables are the year (t, coded as actual year minus 1900) and the longitude (x) and the latitude (z) of each state's centre. Longitudes are measured in negative degrees west of the International Date Line. These data were important to public health officials in 1991 because there were large outbreaks of the disease between 1986 and 1989, especially in those states that did not require mumps vaccination. Chaudhuri *et al*. (1994) applied a least-squares regression tree algorithm to a subset of the data, fitting the observations in each node with a linear function in t, x and z. Their tree was very large with 19 terminal nodes, indicating the presence of complex spatio-temporal interactions.

To demonstrate the advantages of quantile regression, we will fit 0.1-, 0.5- and 0.9-quantile regression trees to the whole data set. The type of polynomial to be fitted in the nodes is determined by two factors: the complexity of the tree structure and its mean quantile prediction error as given in equation (1). Because a highly complex tree is difficult to interpret, simpler trees are usually preferred. Tree complexity, however, can often be reduced by increasing the degree of the piecewise polynomials. We will use cross-validation to estimate the prediction error of a tree model. As in Breiman *et al*. (1984), our algorithm first grows an overly large tree and then employs cross-validation to prune it to the smallest possible size (in terms of number of terminal nodes) such that its cross-validation error estimate is within one estimated standard deviation of the minimum.

Using fivefold cross-validation, we found that the estimated error rates are all very similar for a wide variety of piecewise polynomial tree models. The tree sizes are, however, very large for polynomials that are piecewise linear in x, z and t (the 0.1-, 0.5- and 0.9-quantile trees have 15, 39 and 14 terminal nodes). After some experimentation, we found that reasonably simple tree structures are given by the polynomial

$$\beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \beta_4 x + \beta_5 x^2 + \beta_6 z + \beta_7 z^2 + \beta_8 tx + \beta_9 tz + \beta_{10} xz, \qquad (2)$$

which is of second order in x, z and t, but with an additional cubic term in t. The cubic term is needed to capture the rapid rise and fall in incidence rates in the late 1980s. (We did not fit

a full third-degree polynomial in x, z and t as this would require too many parameters to be estimated at each node.)

Figure 1 shows the three trees. The 0.1- and 0.5-quantile trees first split the USA into two geographical regions according to longitude: a western region consisting of states from Minnesota westward (longitude $\leqslant -94.6$) and an eastern region for states to its east. The 0.9-quantile tree splits first on year: if the year is greater than 1981, the branch has only one node, suggesting that the polynomial model (2) is sufficient for the entire country during 1982–1989. On the other hand, for years from 1953–1981, the 0.9-quantile tree

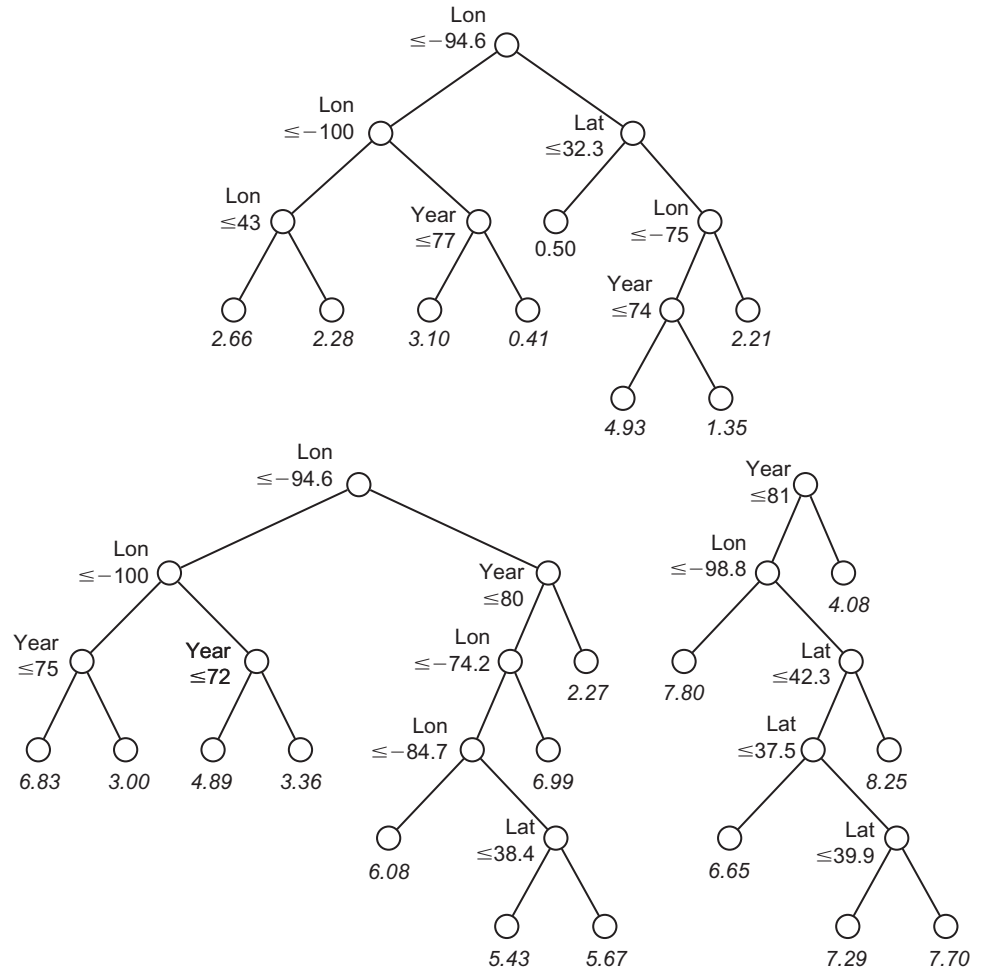

Figure 1. Quantile regression trees for the mumps data using 0.1 (top), 0.5 (bottom left), and 0.9 (bottom right) quantiles. Each node is fitted with the polynomial (2). The number beneath each terminal node is the sample Y-quantile.

splits the country longitudinally in the middle at North Dakota. The numbers beneath the terminal nodes of the trees give the respective sample quantiles of log-rates. They show clearly that the incidence of the disease decreased substantially during the whole time period.

A display of the spatial distribution of the incidence rates as the years increase is given in Figure 2, which shows bubble plots of the fitted median incidence rates for four equally spaced years (states without bubbles did not report for that year). The rates held constant at least until 1966 but were significantly reduced by 1977. The state of Wisconsin was among the hardest hit up to 1977.

The tree diagrams and the bubble plots do not reveal the sharp rise and fall in incidence rates in the late 1980s. To see this, we plot the data and fitted quantile values as functions of year for nine representative states in Figure 3. The rise and fall in rates is now evident, especially for the fitted 0.9-quantile curves. Further, the shapes of the fitted 0.1- and 0.9-quantile curves indicate that there is a fairly substantial degree of heterogeneity in the data (the incidence rates are plotted on a log scale), both between states and over time.

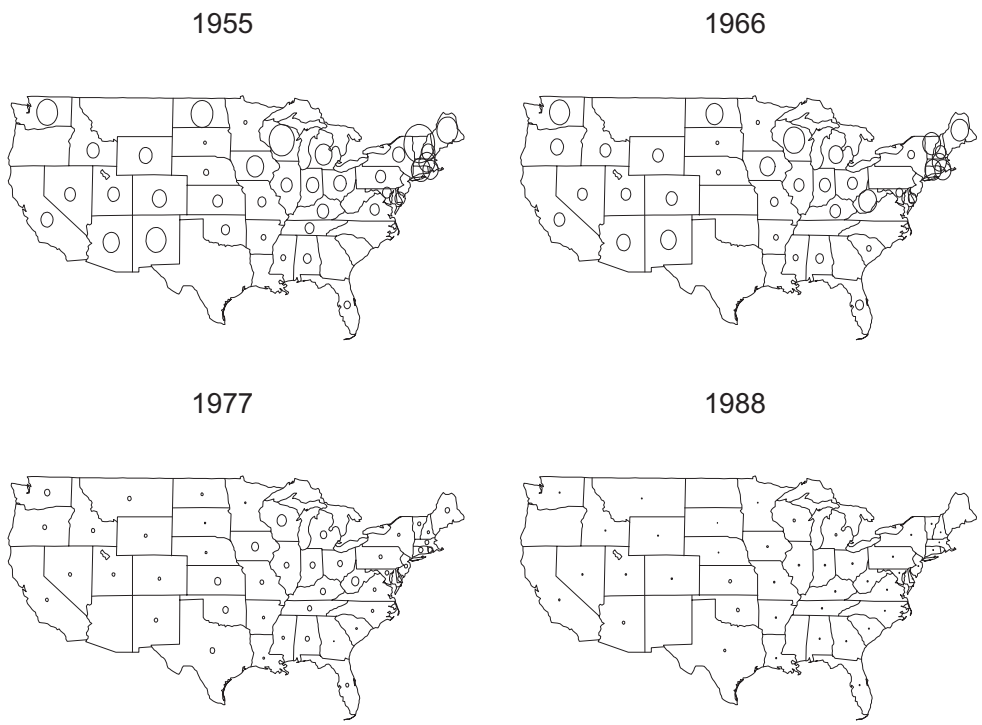

Figure 2. Bubble plots of mumps incidence rates for four equally spaced years. The area of a bubble is proportional to the fitted incidence rate for that state and year. States that did not report mumps incidence have no bubbles.

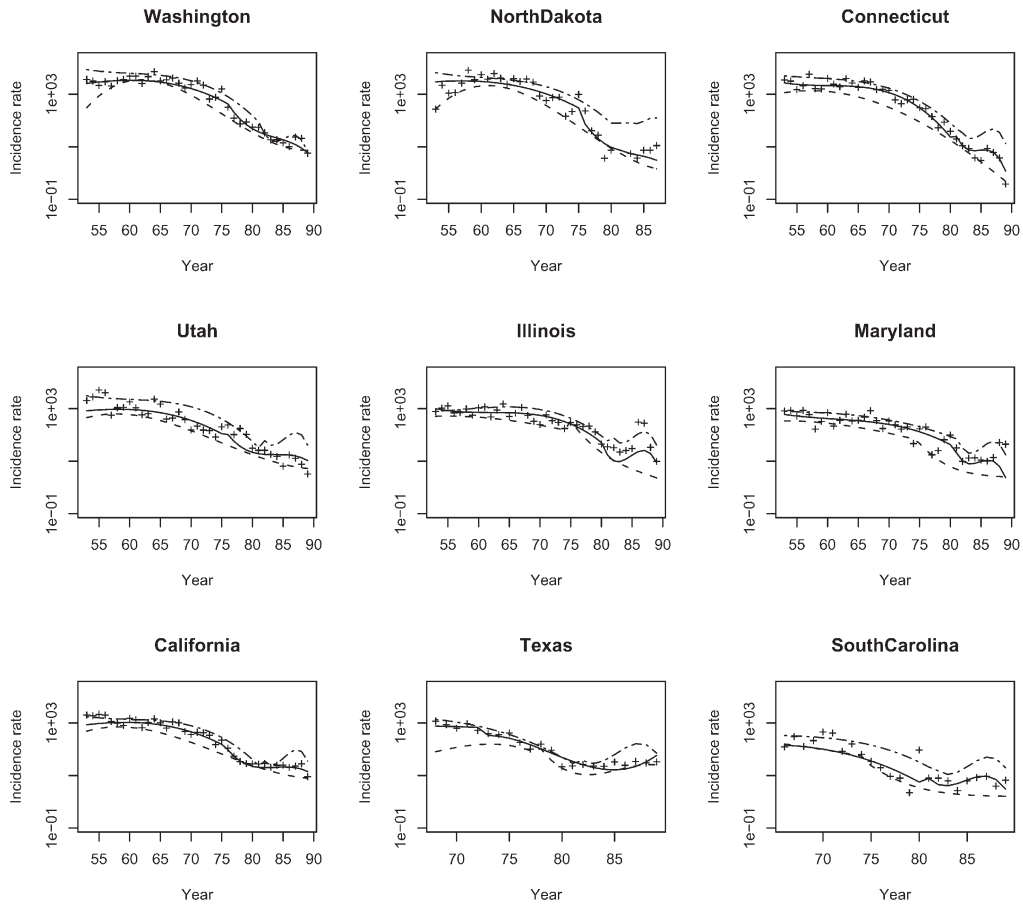

Figure 3. Observed and fitted values from the 0.1 (dashed line), 0.5 (solid line), and 0.9 (dash-dotted line) quantile regression tree models for nine states. The incidence rate axis is on a log scale.

Appendix A. Proof of Theorem 1

Before we present the proof of Theorem 1, we state and prove a few preliminary results and introduce some notation. Recall that $s(A)$ denotes the number of elements of a finite set A. For any subset H of the set of indices $\{1, 2, \ldots, n\}$ such that $s(H) = s(U) \leqslant n$, we write $\mathbf{Y}_H$ to denote the $s(U)$-dimensional vector $(Y_i)_{i \in H}$ and $\mathbf{\Lambda}_h$ to denote the $s(U) \times s(U)$ matrix with rows $\mathbf{\Gamma}_i^{\mathrm{T}}$, $i \in H$, where the $\mathbf{\Gamma}_i$ are defined at the beginning of Section 2.

Proposition 1. *For any fixed* $0 < \alpha < 1$*, the minimization problem*

$$\min_{\mathbf{\Theta}} \sum_{X_i \in t} \{|Y_i - P(\mathbf{X}_i, \mathbf{\Theta}, \overline{\mathbf{X}}_t)| + (2\alpha - 1)[Y_i - P(\mathbf{X}_i, \mathbf{\Theta}, \overline{\mathbf{X}}_t)]\}$$

always has a solution. If the matrix $\mathbf{D}_t = \sum_{\mathbf{x}_i \in t} \mathbf{\Gamma}_i \mathbf{\Gamma}_i^{\mathrm{T}}$ *is non-singular, then there exists at least one set* H *of* $s(U)$ *indices such that* $\mathbf{X}_i \in t$ *for all* $i \in H$, *and* $\hat{\mathbf{\Theta}}_H = \mathbf{\Lambda}_H^{-1} \mathbf{Y}_H$ *is a solution to this minimization problem. Further, for any such solution, the* $s(U)$*-dimensional vector*

$$\mathbf{\Phi}_{H,t} = \frac{1}{2} \sum_{\mathbf{X}_i \in t, i \notin H} \{1 - \mathrm{sgn}[Y_i - P(\mathbf{X}_i, \hat{\mathbf{\Theta}}_H, \overline{\mathbf{X}}_t)] - 2\alpha\} \mathbf{\Gamma}_i$$

lies in the $s(U)$*-dimensional hyperrectangle* $[\alpha - 1, \alpha]^{s(U)}$. *In other words, each real-valued coordinate of* $\mathbf{\Phi}_{H,t}$ *will be bounded above by* α *and bounded below by* $\alpha - 1$.

Proof. First observe that $P(\mathbf{X}_i, \mathbf{\Theta}, \overline{\mathbf{X}}_t) = \mathbf{\Gamma}_i^{\mathrm{T}} \mathbf{\Theta}$. Therefore we can rewrite the minimization problem as

$$\min_{\mathbf{\Theta}} \sum_{\mathbf{x}_i \in t} \{|Y_i - \mathbf{\Gamma}_i^{\mathrm{T}} \mathbf{\Theta}| + (2\alpha - 1)(Y_i - \mathbf{\Gamma}_i^{\mathrm{T}} \mathbf{\Theta})\}.$$

Clearly, any solution to this minimization problem will correspond to an element in the column space of the matrix whose rows are $\mathbf{\Gamma}_i^{\mathrm{T}}$ with $\mathbf{X}_i \in t$. Next notice that, for any fixed $0 < \alpha < 1$, the function $|x| + (2\alpha - 1)x$ is a continuous function in x, and it tends to ∞ as $|x| \to \infty$. This implies that the minimization problem has a solution which corresponds to a point in a compact subset of the linear space spanned by the columns of that matrix with rows $\mathbf{\Gamma}_i^{\mathrm{T}}$. The rest of the proof now follows straightforwardly from Theorems 3.1 and 3.3 of Koenker and Bassett (1978). □

We note that if Condition 3 is satisfied, the assumption in Proposition 1 that $\mathbf{D}_t$ is non-singular holds with large probability for each $t \in T_n$ asymptotically.

Proposition 2. *Let* $|.|$ *denote the usual Euclidean norm of vectors and matrices and let* $F(y|\mathbf{x})$ *be the conditional distribution of* Y *given* $\mathbf{X} = \mathbf{x}$. *Given any* $t \in T_n$ *and* $s(U)$*-dimensional vector* $\mathbf{\Delta}$, *define the* $s(U)$*-dimensional vector*

$$\mathbf{\Psi}_t(\mathbf{\Delta}) = N_t^{-1} \{\delta(t)\}^{-m} \sum_{\mathbf{x}_i \in t} [F\{\mathbf{\Gamma}_i^{\mathrm{T}} \mathbf{\Delta} + r_t(\mathbf{X}_i, \overline{\mathbf{X}}_t) + g_\alpha(\mathbf{X}_i) | \mathbf{X}_i\} - \alpha] \mathbf{\Gamma}_i.$$

Then under Conditions 1–4, $\min_{t \in T_n} \inf\{|\mathbf{\Psi}_t(\mathbf{\Delta})| : |\mathbf{\Delta}| > \xi\{\delta(t)\}^m\}$ *is bounded away from zero in probability as* $n \to \infty$ *for any* $\xi > 0$.

Proof. Let $c_1 > 0$ be a constant depending on $s(U)$ such that $|\mathbf{\Gamma}_i| \leqslant c_1$ for all $1 \leqslant i \leqslant n$. Then, for any non-zero $\mathbf{\Delta}$ and any $t \in T_n$, we have

$$\begin{aligned}\lambda_n &\leqslant N_t^{-1}|\boldsymbol{\Delta}|^{-2}\boldsymbol{\Delta}^{\mathrm{T}}\mathbf{D}_t\boldsymbol{\Delta}\\ &= |\boldsymbol{\Delta}|^{-2}N_t^{-1}\sum_{\mathbf{X}_i\in t}\boldsymbol{\Delta}^{\mathrm{T}}\boldsymbol{\Gamma}_i\boldsymbol{\Gamma}_i^{\mathrm{T}}\boldsymbol{\Delta}\\ &\leqslant \frac{\lambda_n}{2}N_t^{-1}\left[s\left\{i:\ \mathbf{X}_i\in t,\ |\boldsymbol{\Delta}|^{-1}|\boldsymbol{\Gamma}_i^{\mathrm{T}}\boldsymbol{\Delta}|\leqslant\left(\frac{\lambda_n}{2}\right)^{1/2}\right\}\right]\\ &\quad + c_1^2N_t^{-1}\left[s\left\{i:\ \mathbf{X}_i\in t,\ |\boldsymbol{\Delta}|^{-1}|\boldsymbol{\Gamma}_i^{\mathrm{T}}\boldsymbol{\Delta}|>\left(\frac{\lambda_n}{2}\right)^{1/2}\right\}\right]\\ &\leqslant \frac{\lambda_n}{2}+c_1^2p_{n,t},\end{aligned}$$

where $p_{n,t}=N_t^{-1}[s\{i:\ \mathbf{X}_i\in t,\ |\boldsymbol{\Delta}|^{-1}|\boldsymbol{\Gamma}_i^{\mathrm{T}}\boldsymbol{\Delta}|>(\lambda_n/2)^{1/2}\}]$. This implies that $\min_{t\in T_n}p_{n,t}\geqslant\lambda_n/2c_1^2$.

By Condition 4, we can choose a constant $c_2>0$ such that $c_2\leqslant f(y|\mathbf{x})$ for all $\mathbf{x}\in C$ and all $y\in(g_\alpha(\mathbf{x})-\epsilon,\ g_\alpha(\mathbf{x})+\epsilon)$. Let

$$\eta_n=c_2\left(\frac{\lambda_n}{2}\right)^{1/2}\min_{t\in T_n}\min\left(\epsilon/2,\ [\xi\{\delta(t)\}^m]\left(\frac{\lambda_n}{2}\right)^{1/2}\right)$$

$$\begin{aligned}G(t,\mathbf{X}_i,\boldsymbol{\Delta})&=[F\{\boldsymbol{\Gamma}_i^{\mathrm{T}}\boldsymbol{\Delta}+r_t(\mathbf{X}_i,\overline{\mathbf{X}}_t)+g_\alpha(\mathbf{X}_i)|\mathbf{X}_i\}\\ &\quad -F\{r_t(\mathbf{X}_i,\overline{\mathbf{X}}_t)+g_\alpha(\mathbf{X}_i)|\mathbf{X}_i\}]|\boldsymbol{\Delta}|^{-1}(\boldsymbol{\Gamma}_i^{\mathrm{T}}\boldsymbol{\Delta}).\end{aligned}$$

Then Conditions 1 and 3 imply that the event

$$\min_{t\in T_n}\min_{\mathbf{X}_i\in t}\inf\left\{\frac{G(t,\mathbf{X}_i,\boldsymbol{\Delta})}{\{\delta(t)\}^m}:\ |\boldsymbol{\Delta}|>\xi\{\delta(t)\}^m,\ |\boldsymbol{\Delta}|^{-1}|\boldsymbol{\Gamma}_i^{\mathrm{T}}\boldsymbol{\Delta}|>\left(\frac{\lambda_n}{2}\right)^{1/2}\right\}\geqslant\eta_n$$

occurs with probability tending to one as $n\to\infty$. Also, it is obvious that $G(t,\mathbf{X}_i,\boldsymbol{\Delta})\geqslant 0$ for all $s(U)$-dimensional vectors $\boldsymbol{\Delta}$, $t\in T_n$ and $\mathbf{X}_i\in t$.

Let us now use Condition 4 again to choose a constant $c_3>0$ such that $f(y|\mathbf{x})\leqslant c_3$ for all $\mathbf{x}\in C$ and all $y\in(g_\alpha(\mathbf{x})-\epsilon,\ g_\alpha(\mathbf{x})+\epsilon)$. Then by Condition 1, the event

$$\max_{t\in T_n}\max_{\mathbf{X}_i\in t}\sup_{|\boldsymbol{\Delta}|}\left|F\{r_t(\mathbf{X}_i,\overline{\mathbf{X}}_t)+g_\alpha(\mathbf{X}_i)|\mathbf{X}_i\}-\alpha\right||\boldsymbol{\Delta}|^{-1}|\boldsymbol{\Gamma}_i^{\mathrm{T}}\boldsymbol{\Delta}|\leqslant\max_{t\in T_n}\max_{\mathbf{X}_i\in t}c_1c_3r_t(\mathbf{X}_i,\overline{\mathbf{X}}_t)$$

occurs with probability tending to one as $n\to\infty$.

Observe that

$$\begin{aligned}|\boldsymbol{\Psi}_t(\boldsymbol{\Delta})|&\geqslant\{\boldsymbol{\Delta}^{\mathrm{T}}\boldsymbol{\Psi}_t(\boldsymbol{\Delta})\}|\boldsymbol{\Delta}|^{-1}\\ &= N_t^{-1}\{\delta(t)\}^{-m}\{S_1(t,\boldsymbol{\Delta})+S_2(t,\boldsymbol{\Delta})+S_3(t,\boldsymbol{\Delta})\},\end{aligned}$$

where

$$S_1(t, \boldsymbol{\Delta}) = \sum_{\mathbf{X}_i \in t, |\boldsymbol{\Delta}|^{-1}|\boldsymbol{\Gamma}_i^{\mathrm{T}}\boldsymbol{\Delta}| > (\lambda_n/2)^{1/2}} G(t, \mathbf{X}_i, \boldsymbol{\Delta}),$$

$$S_2(t, \boldsymbol{\Delta}) = \sum_{\mathbf{X}_i \in t, |\boldsymbol{\Delta}|^{-1}|\boldsymbol{\Gamma}_i^{\mathrm{T}}\boldsymbol{\Delta}| \leqslant (\lambda_n/2)^{1/2}} G(t, \mathbf{X}_i, \boldsymbol{\Delta}),$$

$$S_3(t, \boldsymbol{\Delta}) = \sum_{\mathbf{X}_i \in t} [F\{r_t(\mathbf{X}_i, \overline{\mathbf{X}}_t) + g_\alpha(\mathbf{X}_i)|\mathbf{X}_i\} - \alpha]|\boldsymbol{\Delta}|^{-1}(\boldsymbol{\Gamma}_i^{\mathrm{T}}\boldsymbol{\Delta}).$$

Our previous analysis implies that

$$\max_{t \in T_n} \sup_{\boldsymbol{\Delta}} N_t^{-1}\{\delta(t)\}^{-m} S_3(t, \boldsymbol{\Delta}) \to 0$$

in probability as $n \to \infty$. Also, $S_2(t, \boldsymbol{\Delta})$ is non-negative for any $t \in T_n$ and any $\boldsymbol{\Delta}$, and the probability of the event

$$\min_{t \in T_n} \inf_{|\boldsymbol{\Delta}| > \xi\{\delta(t)\}^m} N_t^{-1}\{\delta(t)\}^{-m} S_1(t, \boldsymbol{\Delta}) \geqslant \frac{\eta_n \lambda_n}{2c_1^2}$$

tends to one as $n \to \infty$. Combining these results, we conclude that the event

$$\min_{t \in T_n} \inf\{|\boldsymbol{\Psi}_t(\boldsymbol{\Delta})| : |\boldsymbol{\Delta}| > \xi\{\delta(t)\}^m\} \geqslant \frac{\eta_n \lambda_n}{4c_1^2}$$

occurs with probability tending to one as $n \to \infty$. Since η_n and λ_n are positive and bounded away from zero in probability as $n \to \infty$, this completes the proof. □

For any $t \in T_n$, let $\mathcal{S}(t)$ denote the collection of sets H such that $H \subseteq \{i : \mathbf{X}_i \in t\}$ and $s(H) = s(U)$. Note that, by Condition 2, $\mathcal{S}(t)$ is a non-empty collection for each $t \in T_n$ with probability tending to one as $n \to \infty$. Also, for any such H, let $\hat{\boldsymbol{\Theta}}_H$ and $\boldsymbol{\Phi}_{H,t}$ be as defined in Proposition 1. Define $\boldsymbol{\Theta}_t^{(\alpha)}$ to be the $s(U)$-dimensional vector with typical component $\{\delta(t)\}^{[\mathbf{u}]}(\mathbf{u}!)^{-1} D^{\mathbf{u}} g_\alpha(\overline{\mathbf{X}}_t)$ for $\mathbf{u} \in U$. In other words, $g_\alpha(\mathbf{X}_i) = \boldsymbol{\Gamma}_i^{\mathrm{T}} \boldsymbol{\Theta}_t^{(\alpha)} + r_t(\mathbf{X}_i, \overline{\mathbf{X}}_t)$. Also, for $H \in \mathcal{S}(t)$, define

$$\boldsymbol{\Omega}_{H,t}(\boldsymbol{\Delta}) = \sum_{\mathbf{X}_i \in t, i \notin H} [F\{\boldsymbol{\Gamma}_i^{\mathrm{T}}\boldsymbol{\Delta} + r_t(\mathbf{X}_i, \overline{\mathbf{X}}_t) + g_\alpha(\mathbf{X}_i)|\mathbf{X}_i\} - \alpha]\boldsymbol{\Gamma}_i.$$

Proposition 3. *As* $n \to \infty$,

$$\max_{t \in T_n} \max_{H \in \mathcal{S}(t)} \{N_t - s(U)\}^{-1}\{\delta(t)\}^{-m} |\boldsymbol{\Phi}_{H,t} - \boldsymbol{\Omega}_{H,t}(\boldsymbol{\Theta}_t^{(\alpha)} - \hat{\boldsymbol{\Theta}}_H)| \xrightarrow{P} 0.$$

Proof. Recall that each set in T_n is a polyhedron in d-dimensional Euclidean space having at most M faces. A combinatorial result of Vapnik and Chervonenkis (1971) – see Dudley (1978, Section 7) – implies that there exists a collection $\mathcal{V}$ of subsets of the set $\{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n\}$ such that $s(\mathcal{V}) \leqslant (2n)^{M(d+2)}$, and for any polyhedron t with at most M

faces there is a set $t^* \in \mathcal{V}$ with the property that $\mathbf{X}_i \in t$ if and only if $\mathbf{X}_i \in t^*$. For any $\omega > 0$, let $p(\omega, \mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)$ denote the conditional probability of the event

$$\max_{t \in T_n} \max_{H \in \mathcal{S}(t)} \{N_t - s(U)\}^{-1} \{\delta t)\}^{-M} |\mathbf{\Phi}_{H,t} - \mathbf{\Omega}_{H,t}(\mathbf{\Theta}_t^{(\alpha)} - \hat{\mathbf{\Theta}}_h)| > \omega$$

given $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$. Observe that for any $t^* \in \mathcal{V}$ and $H \in \mathcal{S}(t^*)$, the difference $\mathbf{\Phi}_{H,t^*} \mathbf{\Omega}_{H,t^*}(\mathbf{\Theta}_t^{(\alpha)} - \hat{\mathbf{\Theta}}_H)$ is a sum of $s(U)$-dimensional random vectors that are conditionally independently distributed and each of them has conditional mean zero given the $\mathbf{X}_i$ in t^* and the Y_i for which $i \in H$. It follows from Bernstein's inequality (Shorack and Wellner 1986) that there exist constants $c_4 > 0$ and $c_5 > 0$ such that by Condition 2, the event

$$p(\omega, \mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n) \leqslant c_4 (2n)^{M(d+2)} n^{s(U)} \exp(-c_5 N_n \omega^2)$$

occurs with probability tending to one as n tends to ∞. Since $N_n / \log n \to \infty$ in probability as $n \to \infty$, this completes the proof. □

Proof of Theorem 1. The first assertion made in the statement of the theorem follows immediately from Proposition 1. The second assertion will follow if we can show that $\max_{t \in} T_n \{\delta(t)\}^{-m} |\hat{\mathbf{\Theta}}_t^{(\alpha)} - \mathbf{\Theta}_t^{(\alpha)}|$ tends to zero in probability as $n \to \infty$. Now Proposition 1 implies that for any $\xi > 0$, the event

$$\max_{t \in T_n} \{\delta(t)\}^{-m} |\hat{\mathbf{\Theta}}_t^{(\alpha)} - \mathbf{\Theta}_t^{(\alpha)}| > \xi$$

is contained in the event

$$\bigcup_{t \in T_n} \bigcup_{H \in \mathcal{S}(t)} \{|\hat{\mathbf{\Theta}}_H - \mathbf{\Theta}_t^{(\alpha)}| > \xi \{\delta(t)\}^m \text{ and } \mathbf{\Phi}_{H,t} \in [\alpha - 1, \alpha]^{s(U)}\}.$$

The proof now follows from Propositions 2 and 3. □

Appendix B. Algorithmic and computational details

The method used in Section 3 to obtain the quantile regression trees is an extension of the GUIDE algorithm for piecewise linear least-squares regression trees described in Loh (2002). GUIDE differs from regression tree algorithms such as CART (Breiman *et al*. 1984) in many significant ways. The most important difference is that GUIDE does not use greedy search to split each node. Greedy search has two undesirable features. Firstly, it is computationally intensive – for each candidate split of a node into two subnodes, a quantile regression model is fitted to the data in each subnode. Since the number of candidate splits increases with the sample size and with the number of predictor variables, this procedure can be time-consuming to carry out. The second disadvantage of greedy search is that it is biased towards selecting variables that have more candidate splits. This problem was recognized long ago for classification trees (Doyle 1973; Loh and Shih 1997) and was confirmed for regression trees by Loh (2002) in simulation experiments.

To avoid the computational cost and selection bias of greedy search, GUIDE breaks the

split selection procedure into two steps – first it chooses the variable to split the node, and then it chooses the split point (if the variable takes ordered values) or split set (if the variable takes categorical, i.e. unordered, values). The entire algorithm is described in detail for least-squares regression in Loh (2002). We briefly summarize the steps in the context of quantile regression here:

1. Fit a quantile regression model to the data in the node using the algorithm in Koenker and D'Orey (1987) and compute the residuals.
2. For each predictor variable, cross-tabulate the signs of the residuals (positive versus non-positive) against the grouped values of the variable and compute a chi-square p-value.
3. If there are categorical predictor variables, adjust the chi-square p-values with a bootstrap bias correction.
4. Select the variable with the smallest adjusted p-value to split the node.
5. If the selected variable takes ordered values, search for the best split point for the variable over a grid of 100 empirical q-quantiles, with $q = i/101$, $i = 1, \ldots, 100$.
6. If the selected variable is categorical, search for the subset of categorical values that best separates the two groups of signed residuals in terms of binomial variance.

The bootstrap adjustment is needed to overcome the tendency for the regressor variables (which are used for split selection as well as for fitting the quantile regression model in the node) to have larger p-values than the categorical variables (which are used for split selection only). These steps are performed recursively to produce an overly large tree, which is pruned to a smaller size using the cost-complexity pruning algorithm of Breiman *et al*. (1984) with fivefold cross-validation.

Much of the computation saved is due to fitting only one quantile regression model at each node. Further, the use of residuals permits all kinds of quantile regression models to be fitted. Thus we can fit piecewise-constant (as in CART), piecewise-linear or piecewise-polynomial (as in the mumps example) models.

Acknowledgements

Chaudhuri's research was partially supported by a grant from the Indian Statistical Institute. Loh's research was partially supported by US Army Research Office grants DAAH04-94-G-0042, DAAG55-98-1-0333, and DAAD19-01-1-0586, a grant from Pfizer, Inc., and a University of Wisconsin Vilas Associateship. The authors thank an associate editor and two referees for their helpful comments.

References

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

Buchinsky, M. (1994) Changes in the U.S. wage structure 1963–1987: Application of quantile regression. *Econometrica*, **62**, 405–458.

Chaudhuri, P. (1991a) Global nonparametric estimation of conditional quantile functions and their derivatives. *J. Multivariate Anal.*, **39**, 246–269.

Chaudhuri, P. (1991b) Nonparametric estimates of regression quantiles and their local Bahadur representation. *Ann. Statist.*, **19**, 760–777.

Chaudhuri, P. (2000) Asymptotic consistency of median regression trees. *J. Statist. Plann. Inference*, **91**, 229–238.

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

Chaudhuri, P., Lo, W.-D., Loh, W.-Y. and Yang, C.-C. (1995) Generalized regression trees. *Statist. Sinica*, **5**, 641–666.

Chaudhuri, P., Doksum, K. and Samarov, A. (1997) On average derivative quantile regression. *Ann Statist.*, **25**, 715–744.

Cheng, K.F. (1983) Nonparametric estimators for percentile regression functions. *Comm. Statist. Theory Methods*, **12**, 681–692.

Cheng, K.F. (1984) Nonparametric estimation of regression function using linear combinations of sample quantile regression function. *Sankhyā Ser. A*, **46**, 287–302.

Dabrowska, D. (1992) Nonparametric quantile regression with censored data. *Sankhyā Ser. A*, **54**, 252–259.

Doyle, P. (1973) The use of Automatic Interaction Detector and similar search procedures. *Oper. Res. Quart.*, **24**, 465–467.

Dudley, R.M. (1978) Central limit theorems for empirical measures. *Ann. Probab.*, **6**, 899–929. Correction (1979): **7**, 909–911.

Fan, J. and Gijbels, I. (1996) *Local Polynomial Modeling and Its Applications*. London: Chapman & Hall.

Fan, J., Hu, T.C. and Truong, Y.K. (1994) Robust nonparamentric function estimation. *Scand. J. Statist.*, **21**, 433–446.

Han, A. (1987) A nonparametric analysis of transformations. *J. Econometrics*, **35**, 191–209.

Härdle, W. and Stoker, T. (1989) Investigating smooth multiple regression by the method of average derivatives. *J. Amer. Statist. Assoc.*, **84**, 986–995.

Hendricks, W. and Koenker, R. (1992) Hierarchical spline model for conditional quantiles and the demand for electricity. *J. Amer. Statist. Assoc.*, **87**, 58–68.

Hogg, R.V. (1975) Estimates of percentile regression line using salary data. *J. Amer. Statist. Assoc.*, **70**, 56–59.

Janssen, P. and Veraverbeke, N. (1987) On nonparametric regression estimators based on regression quantiles. *Comm. Statist. Theory Methods*, **16**, 383–396.

Koenker, R. and Bassett, G. (1978) Regression quantiles. *Econometrica*, **46**, 33–50.

Koenker, R. and D'Orey, V. (1987) Computing regression quantiles. *Appl. Statist.*, **36**, 383–393.

Koenker, R., Ng, P. and Portnoy, S. (1994) Quantile smoothing splines. *Biometrika*, **81**, 673–680.

Lejeune, M.G. and Sarda, P. (1988) Quantile regression: a nonparametric approach. *Comput. Statist. Data Anal.*, **6**, 229–239.

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

Loh, W.-Y. and Shih, Y.-S. (1997) Split selection methods for classification trees. *Statist. Sinica*, **7**, 815–840.

Newey, W.K. and Stoker, T.M. (1993) Efficiency of weighted average derivative estimators and index models. *Econometrica*, **61**, 1199–1223.

Powell, J., Stock, J. and Stoker, T. (1989) Semiparametric estimation of index coefficients. *Econometrica*, **57**, 1403–1430.

Samarov, A. (1993) Exploring regression structure using nonparametric functional estimation. *J. Amer. Statist. Assoc.*, **88**, 836–849.

Sherman, R. (1993) The limiting distribution of the maximum rank correlation estimator. *Econometrica*, **61**, 123–137.

Shorack, G.R. and Wellner, J.A. (1986) *Empirical Processes with Applications to Statistics*. New York: Wiley.

Truong, Y.K.N. (1989) Asymptotic properties of kernel estimators based on local medians. *Ann. Statist.*, **17**, 606–617.

Vapnik, V.N. and Chervonenkis, A.Ya. (1971) On the uniform convergence of relative frequencies of events to their probabilities. *Theory Probab. Appl.*, **16**, 264–280.

Wand, M.P. and Jones, M.C. (1995) *Kernel Smoothing*. London: Chapman & Hall.

Welsh, A.H. (1996) Robust estimation of smooth regression and spread functions and their derivatives. *Statist. Sinica*, **6**, 347–366.

Received February 2000 and revised February 2002