

# Tree-Structured Logistic Model for Over-Dispersed Binomial Data with Application to Modeling Developmental Effects\*

**Hongshik Ahn**

Department of Applied Mathematics and Statistics  
State University of New York at Stony Brook  
Stony Brook, NY 11794-3600

and

**James J. Chen**

Division of Biometry and Risk Assessment  
National Center for Toxicological Research  
Food and Drug Administration  
Jefferson, Arkansas 72079

## Abstract

This article proposes tree-structured logistic regression modeling for over-dispersed binomial data. Recursive partitioning is performed using a combination of statistical tests and residual analysis. The splitting criterion in cross-validation is based on the deviance function. A nested grid algorithm to estimate the bootstrap parameters is developed. The regression tree procedure provides a new approach for exploring in detail the relationship between the binomial response and explanatory variables. The proposed procedure is applied to model the relationship between the incidence of malformation and dose and fetal weight using data from a developmental experiment conducted at the National Center for Toxicological Research. A conditional Gaussian chain model is used to account for the effect of fetal weight by dose.

## 1 Introduction

Recently, tree-based methods have been developed by many researchers. The tree-structured approaches are used for classification (Breiman et al., 1984; Loh and Vanichsetakul, 1988), least squares regression (Breiman et al., 1984; Chaudhuri et al., 1994) and analysis of censored survival data (Segal, 1988; Segal and Bloch, 1989; Ciampi and Thiffault, 1989; Davis and Anderson, 1989; Loh, 1991; LeBlanc and Crowley, 1992; Ahn and Loh, 1994). Tree-structured regression became possible due to rapid computer advances. The first tree-structured approach was the Automatic Interaction Detection (AID) program introduced by Morgan and Sonquist (1963). In the AID program, recursive partitioning was used as an alternative to the least squares regression for model fitting. Breiman et al. (1984) developed the Classification and Regression Trees (CART) method of selecting a tree of appropriate size for classification and piecewise constant regression. Loh and Vanichsetakul (1988) proposed a Fast Algorithm for Classification Trees (FACT) by recursive application of linear discriminant analysis. Their splitting rule is based on a separation of multivariate normal distributions. They used  $F$  ratios to determine when to split and when to stop splitting.

With regression trees, some of the restrictive classical assumptions about the relationship between the response and explanatory variables can be avoided. A tree-structured regression provides easier interpretation of the model than fitting a single regression equation to the whole sample because the tree identifies effects

---

\**Key words:* Bootstrap, Cross-validation, Dose-response, Over-dispersion, Regression tree.

of explanatory variables in each terminal node. Regression trees provide insight into the nature of the relationship between the response and explanatory variables within a node. Further, because the data in a node would be more homogeneous, they may be fitted with models having fewer covariates and thus ease the difficulties associated with collinearity.

Lately regression trees have been extended to generalized linear models. Chaudhuri et al. (1995) develop logistic regression trees for binary data and Poisson regression trees for count data. The splitting point of a node is based on an analysis of the pattern of residuals. Chaudhuri et al. (1995) used the Anscombe residual in the Poisson regression trees. They used the pseudo-residual from a smoothed response variable in the logistic regression trees since the response variable is binary.

The logistic regression tree procedure proposed by Chaudhuri et al. (1995) performs well for binary data. However, in many applications, binary responses often occur in clusters and the responses from the same cluster may be correlated. For example in teratology, a pregnant animal is treated with some compound of interest and responses are measured on the fetuses in the litter. In ophthalmology, the sampling unit is the person and responses are gathered on each of the eyes. The responses from the same cluster tend to be correlated. In these cases, the standard binomial model is not appropriate because the variation in the data is greater than expected under the binomial model. Stiratelli, Laird and Ware (1984), Prentice (1988), Rosner and Milton (1988) and many other authors have proposed various models for the analysis of this type of data. Other examples of the extra-binomial variation model are discussed in Morgan (1992, Chapter 6).

In teratology studies, a dose-response model is often fit to bioassay data to provide a relationship between the probability of a developmental defect and the level of exposure. A well-known problem in dose-response modeling is that larger doses are generally used in animals than in humans in order to elicit potential toxic effects at levels that are measurable in a limited number of animals. Furthermore, fetal weight reduction has routinely been examined and used as an indication of developmental toxicity. Ryan et al. (1991) found a tendency for malformed fetuses to have a lower weight at term than nonmalformed fetuses. To utilize the interrelationship between the incidence of malformation and fetal weight, Catalano and Ryan (1992) assumed the malformation outcomes have some corresponding unobserved latent variable, and both the fetal weight and the latent variable share a joint multivariate normal distribution. Chen (1993) proposed a conditional regression chain model where fetal weight is modeled as a function of dose conditional on other developmental endpoints (e.g., litter size) and then the malformation incidence is modeled as a function of dose and the residual from the fetal weight model.

The primary purpose of this paper is to present a logistic regression tree algorithm for analysis of over-dispersed binomial data or binomial data as a special case. The regression tree algorithm is applied to dose-response modeling of a developmental effect. Separate dose-response models are fitted in each terminal node. The focus of the application is to identify a low-dose region such that a single regression equation can be fitted well. The regression tree procedure is then extended to incorporate fetal weight as an explanatory variable. This application is important and useful since the fetal weight variable itself is a developmental endpoint and is affected by dose.

The present paper extends and improves the Chaudhuri et al. (1995) logistic regression tree procedure by incorporating the following approaches: (1) using the robust variance-covariance estimator used by Liang and Zeger (1986) for the quasi-likelihood estimation; (2) using the deviance function in  $V$ -fold cross-validation for determining whether to split as opposed to the stopping rules or pruning procedures presented by Chaudhuri et al. (1995); (3) using the bootstrap procedures of Ahn and Loh (1994) for estimating the parameters necessary for the  $V$ -fold cross-validation procedure; and (4) extending the logistic approach of Chaudhuri et al. (1995) to accommodate the conditional Gaussian regression chain model of Chen (1993).

Data from a developmental toxicology study of exposure to the herbicide 2,4,5-trichlorophenoxyacetic acid conducted at the National Center for Toxicological Research are used for illustration. One outbred (CD-1) and four inbred (C57BL/6, C3H/He, BALB/C and A/JAX) strains of mice were tested with six or seven dose levels. Further details of the study are given in Holson et al. (1992). Only the data from the A/JAX strain are used in this illustration. The data contain seven doses (0, 15, 20, 25, 30, 45, and 60 mg/kg/day), and the numbers of litters for each dose are 89, 86, 56, 40, 76, 33, and 9, respectively. Table 1 lists the number of malformation (cleft palate,  $x_i$ ), the number of live fetuses ( $n_i$ ), and the average fetal weight ( $w_i$ ) for each litter. Figure 1 shows a scatterplot of  $\text{logit}(\text{malformation})$  versus dose. A smooth curve generated by the "lowess" function using the default parameters in S (Becker, Chambers, and Wilks, 1988) was superimposed on the figure. A possible modeling approach is adding a quadratic dose parameter.

However, it is seen that the smooth curve is quite different from a parabola. If the data are divided at some point of dose between 20 and 25, then a linear fit might be adequate in each group.

## 2 The Over-Dispersed Binomial Distribution and Logistic Regression Model

Let  $Z_{i1}, \dots, Z_{in_i}$  be binary variables from a cluster of size  $n_i, i = 1, \dots, n$ . Assume that  $E(Z_{ij}) = \mu_i$ ,  $\text{var}(Z_{ij}) = \mu_i(1 - \mu_i)$  and  $\text{Corr}(Z_{ij}, Z_{ik}) = \rho$  for  $j \neq k$ . Let  $Y_i = n_i^{-1} \sum_{j=1}^{n_i} Z_{ij}$ . Then  $E(Y_i) = \mu_i$  and  $\text{var}(Y_i) = \mu_i(1 - \mu_i)[1 + (n_i - 1)\rho]/n_i$ . The parameters  $\mu_i$  and  $\rho$  are the mean and over-dispersion parameters of the distribution, respectively. If  $\rho > 0$ , then  $n_i Y_i$  is an over-dispersed binomial, and if  $\rho = 0$ , then  $n_i Y_i$  reduces to a binomial. Assume  $\text{Corr}(Z_{ij}, Z_{i'k}) = 0$  for  $i \neq i'$ , and consequently,  $Y_1, \dots, Y_n$  are independent. Under the generalized linear model (GLM), the logistic regression model for an over-dispersed binomial variable assumes  $g(\mu_i) = \text{logit}(\mu_i) = \mathbf{x}_i' \boldsymbol{\beta}$ ,  $\text{var}(Y_i) = v(\mu_i) = \mu_i(1 - \mu_i)\phi_i/n_i$ , where  $X = (\mathbf{x}_1, \dots, \mathbf{x}_n)'$  is an  $n \times p$  matrix of covariates,  $\phi_i = 1 + (n_i - 1)\rho$ , and the functions  $g$  and  $v$  are the link and variance functions, respectively.

Wedderburn (1974) defined the log quasi-likelihood  $Q$  for an observation  $y$  with mean  $\mu$  and variance  $v(\mu)$  by

$$Q(y; \mu) = \int_y^\mu \frac{y - t}{v(t)} dt + (\text{function of } y),$$

or equivalently

$$\frac{\partial Q(y; \mu)}{\partial \mu} = \frac{y - \mu}{v(\mu)}.$$

Since  $Y_1, \dots, Y_n$  are independent, the log quasi-likelihood for the complete data can be written as

$$Q(\mathbf{y}; \boldsymbol{\mu}) = \sum_{i=1}^n Q_i(y_i; \mu_i).$$

The quasi-likelihood estimating equations for  $\boldsymbol{\beta}$  are of the form

$$\frac{\partial Q(\mathbf{y}; \boldsymbol{\mu})}{\partial \boldsymbol{\beta}} \Big|_{\hat{\boldsymbol{\beta}}} = U(\hat{\boldsymbol{\beta}}) = \hat{D}' \hat{V}^{-1} (Y - \hat{\boldsymbol{\mu}}) = \mathbf{0}, \quad (1)$$

where  $D = [d_{ij}]$  is an  $n \times p$  matrix such that  $d_{ij} = \partial \mu_i / \partial \beta_j$ , and  $\hat{V} = \text{diag}\{v(\hat{\mu}_1), \dots, v(\hat{\mu}_n)\}$ . The regression parameters  $\boldsymbol{\beta}$  can be estimated by using the Newton-Raphson method (see McCullagh and Nelder, 1989, Chapter 9). The over-dispersion parameter  $\rho$  is estimated by the moment method in each iteration of the Newton-Raphson method,

$$\hat{\rho} = \sum_{i=1}^n \frac{(y_i - \hat{\mu}_i)^2 - v(\hat{\mu}_i)}{(n_i - 1)v(\hat{\mu}_i)}.$$

The quasi-likelihood estimator  $\hat{\boldsymbol{\beta}}$  of  $\boldsymbol{\beta}$  is asymptotically normal with the (model-based) covariance matrix  $\text{Cov}(\hat{\boldsymbol{\beta}}) = [D' V^{-1} D]^{-1}$ . Liang and Zeger (1986) showed that the robust variance estimator  $M_0^{-1} M_1 M_0^{-1}$  is consistent even when  $\text{var}(Y_i)$  is misspecified, where  $M_0 = \hat{D}' \hat{V}^{-1} \hat{D}$  and  $M_1 = \hat{D}' \hat{V}^{-1} (Y - \hat{\boldsymbol{\mu}}) (Y - \hat{\boldsymbol{\mu}})' \hat{V}^{-1} \hat{D}$ . Hence, the robust variance estimate is close to the model-based variance estimate if the variance structure is adequately modeled. If  $\rho = 0$ , then the estimator  $\hat{\boldsymbol{\beta}}$  in Equation (1) is the maximum likelihood estimator under the binomial model or under the binary model where each individual response within a cluster is considered to be an independent experimental unit. The logistic regression tree model presented in this paper includes both the standard binomial ( $\rho = 0$ ) and the over-dispersed binomial ( $\rho > 0$ ) models.

Measures of goodness of fit may be formed in various ways, but the deviance function is a standard measure of discrepancy in quasi-likelihood models. The deviance is defined as

$$\text{dev}(\mathbf{y}; \hat{\boldsymbol{\mu}}) = \sum_{i=1}^n [-2Q_i(y_i; \hat{\mu}_i)].$$

For the normal distribution, the deviance is just the residual sum of squares. In generalized linear models, minimizing the deviance is equivalent to maximizing the quasi-likelihood. The deviance will be used as the splitting criterion in cross-validation of the proposed logistic regression tree.

The observations  $y_i$  have mean  $\mu_i$  and can be expressed as  $y_i = \mu_i + \epsilon_i$ . The residual components  $\epsilon_i = y_i - \mu_i$  have zero mean, and the  $n_i\epsilon_i$  have a shifted binomial distribution in the case that  $\rho = 0$  (see Collet, 1991, p57). The patterns of these residuals are analyzed for partitioning the sample data to construct a regression tree. Because the residual under the binomial model is cluster-based and the residual under the Bernoulli model is individual-based, the regression trees of these two models may be different. In the case of modeling within-litter covariates such as pup-level fetal weight or different malformation types, binary residuals should be used. However, in this paper, we consider the responses from clusters, which is important in many applications as we discussed in Section 1. We assume within-cluster covariates are constant or approximately equal.

### 3 Tree-Structured Models

Methods for recursive partitioning of the data leading to logistic regression trees are described in this section. Any covariate that takes categorical values is transformed into a dummy vector of 0/1 indicator variables for the purpose of fitting logistic regression models.

#### 3.1 Splitting

Binary regression trees are constructed by repeated splitting of nodes into two subnodes. At each node, a covariate vector is considered to be in class 1 if its associated residual of the logistic regression is positive, and to be in class 2 otherwise. If the fitted model is not appropriate (a node should be split), the distributions of the covariate values in the two classes should be quite different. The two-sample  $t$  tests and Levene's (1960) tests for differences in variances of each covariate are performed to detect the heterogeneity of the two classes. This method has proven to be effective for tree-structured classification (Loh and Vanichsetakul, 1988), piecewise-polynomial regression (Chaudhuri et al., 1994) and regression with censored data (Ahn and Loh, 1994). It takes less computation time than the exhaustive search method in CART. The algorithm is given in Appendix A.1.

#### 3.2 Stopping

To determine if a node should be split, the deviance function for the logistic model is used as a measure of goodness of fit in  $V$ -fold cross-validation. Before cross-validation, the values of the fractional reduction  $f$  and splitting threshold  $\eta$  need to be estimated by the bootstrap. The  $f$  and  $\eta$  values determine the size of the tree. A cross-validated multi-step look-ahead stopping rule given by Chaudhuri et al. (1994) is used to decide whether or not to split a node. The deviance function is used as a measure of goodness of fit in cross-validation. The procedure is stated as follows. The data in the node are randomly divided into  $V$  subsets each containing nearly the same number of observations. A nested sequence of trees is constructed from the data consisting of  $(V - 1)$  subsets and the remaining subset is used as a test sample to decide if the data should be split. This procedure is applied  $V$  times, each time leaving out a different subset as a test sample. If a cross-validation tree has an estimate of deviance that is less than or equal to  $(1 - f)$  times that for the trivial tree, then it is considered better than the trivial tree (a tree without any split). If the number of better cross-validation trees is larger than  $\eta V$ , then the node is split. The cross-validation algorithm that is similarly described in Ahn and Loh (1994) is given in Appendix A.2.

In order to determine the pruning parameters  $f$  and  $\eta$ , bootstrap resampling is used. The hypothesis that a nontrivial tree results when in fact a single logistic model suffices for all the data is tested. The probability of a Type I error is

$$\alpha = P(\text{Split the root node} | H_0 : \text{The root node should not be split}). \quad (2)$$

The probability (2) is evaluated using different values of  $f$  and  $\eta$ . The  $f$  and  $\eta$  are chosen to be the values for which (2) is closest to the preselected  $\alpha$ . Among the three methods of estimating  $f$  and  $\eta$  proposed by Ahn and Loh (1994), the following two are recommended for the proposed regression trees:

B1 Fixing  $f = \eta$ , select the value of  $f$  for which (2) is closest to  $\alpha$ .

B2 Fixing  $f = 0$ , select the value of  $\eta$  for which (2) is closest to  $\alpha$ .

In simulations, the B1 method performed well and the B2 method performed adequately, but the latter gave less power than the B1 method (see Section 5). However, a third method (fixing  $\eta = .5$  and choosing the value of  $f$  for which (2) is closest to  $\alpha$ ) was not satisfactory and hence not reported. A finite grid with an increment of .1 (Ahn and Loh, 1994) does not give accurate estimates of  $f$  and  $\eta$  for the logistic regression trees, because even a small change in  $f$  and  $\eta$  often results in a different sized regression tree. In this paper, therefore, a nested grid method with an increment of .01 is developed for selecting the values of  $f$  and  $\eta$ . By using the  $f$  (or  $\eta$ ) value chosen from the grid with an increment of .1, increase the value of  $f$  (or  $\eta$ ) by .01 until the best value is found. Figure 2 shows how to estimate  $f$  and  $\eta$ . In this figure, Ahn and Loh’s (1994) approach does not have the smaller (nested) dots. Therefore, the nested grid approach has better precision than Ahn and Loh’s grid approach. For the proposed logistic regression trees, the nested grid approach makes the probability of a Type I error of (2) closer to the preselected  $\alpha$  than the approach in Ahn and Loh (1994). Further details of the estimation procedure are given in Appendix A.3.

To generate bootstrap samples, the following steps are performed.

- Randomly select the bootstrapped version of the vector of covariates for the  $i$ th case,  $\mathbf{x}_i^*$ , and the corresponding number of trials  $n_i^*$  with replacement from the entire sample.
- Because  $y_i$  given  $\mathbf{x}_i$  is distributed as an over-dispersed binomial with parameters  $n_i, \mu_i$  and  $\rho$ , generate the bootstrap estimate  $y_i^*$  from the over-dispersed binomial distribution with parameters  $n_i^*, \hat{\mu}_i^* = \exp(\mathbf{x}_i^* \hat{\boldsymbol{\beta}}) / [1 + \exp(\mathbf{x}_i^* \hat{\boldsymbol{\beta}})]$  and  $\hat{\rho}$ . The method of generating multivariate binary data proposed by Emrich and Piedmonte (1991) is used to generate the over-dispersed binomial data. If  $\rho = 0$ , then binomial samples are generated.
- Repeat the above procedures  $n$  times to get a bootstrap sample with  $n$  observations.

The above parametric bootstrap approach is used in our logistic regression tree algorithm. A possible alternative approach is with the quasi-likelihood estimation procedure by Moulton and Zeger (1989). They provide a precedent in the clustered binary data context for a bootstrap technique based upon sampling from original data as opposed to simulating random data.

## 4 Analysis of Herbicide 2,4,5-T Data

The proposed logistic regression tree procedure is applied to the data discussed in Section 1. This procedure can analyze data in complex situations, rich in covariates. We apply the proposed tree procedure to identify a low-dose region with a single regression equation, which is a very important application in quantitative risk assessment. Two logistic regression functions are considered. The first function is the simple linear logistic dose-response model. The second function also includes dose and weight but uses the conditional Gaussian regression chain model proposed by Chen (1993). The tree-structured logistic regression was performed using both the B1 and B2 methods. Both the model-based and robust standard error estimates were computed. The robust standard error estimate was used for tests of the significance of the model parameters because the robust estimate is consistent even when the correlation structure is misspecified. Before presenting the analysis, let  $\text{node}(i, j)$  denote the  $j$ th node from the left (including the empty nodes) at the  $i$ th level of the tree. The root node is  $\text{node}(0, 1)$ .

### 4.1 Dose-Response Model

The logistic dose-response model for malformation is

$$\text{logit}(\mu_i) = \beta_0 + \beta_1 d_i,$$

where  $d_i \in \{0, 15, 20, 25, 30, 45, 60\}$ . This model is referred to as Model 1. First, we fitted the model without over-dispersion. For this model, the B1 method gave a tree with one split, at dose 20.06 (Figure 3). The

upper half of Table 2 gives the parameter estimates. The robust estimates of the standard error are larger than the model-based estimates for the whole sample and the subsamples at the terminal nodes. For the lower dose groups ( $d \leq 20$ ), the differences between the two estimates are smaller than those for the whole sample. However, the robust estimates are more than twice that of the model-based estimates at the higher dose groups ( $d \geq 25$ ). This indicates that the model without over-dispersion is not adequate for the data. Furthermore, some efficiency may be gained by explicitly modeling the correlation structure.

Applying the tree-structured regression with over-dispersion, the B1 method gave the same tree (the same split point) as that in Figure 3. The lower half of Table 2 presents the parameter estimates for the model. The model-based and robust standard error estimates are much closer for the whole sample and both terminal nodes than for the model without over-dispersion. The similarity of the two standard error estimates provides evidence of a good model fit. The B2 method gave a trivial tree (the root node was not split) for both procedures.

The table shows a significant dose effect ( $p < .01$ ) for the whole sample and for each subsample. Both  $Z$ -values from the subsamples are less than the  $Z$ -value from the whole sample ( $Z = \hat{\beta} / \text{S.E.}$ ). The slope of the dose-response function of the lower dose groups (0, 15, and 20) in node(1, 1) is smaller than the slope of the higher dose groups (25, 30, 45, and 60) in node(1, 2), as are the estimates of the intralitter correlation parameter  $\rho$ . That is, separate intralitter correlation parameters were estimated at each terminal node. The estimates of  $\rho$  are .2184 for the lower dose groups and .4078 for the higher dose groups. Using different intralitter correlation estimates for different dose groups rather than a constant correlation estimate across groups may reduce biases in the estimation of dose-response coefficients using a full likelihood procedure (Kupper et al., 1986) but does increase the number of estimated parameters. In the quasi-likelihood approach, the GEE (generalized estimating equation) mean parameter estimators should be asymptotically unbiased (i.e., consistent) regardless of the specification of the correlation structure. It is true, however, that the efficiency of the estimators increases with more accurate specification of the correlation structure. The tree-structured regression can also provide a simple indication for the effect of dose on the intralitter correlation.

## 4.2 Conditional Gaussian Regression Chain Model

To account for the effect of dose on fetal weight, Chen (1993) proposed a two-stage conditional Gaussian regression chain model. The first stage modeled the fetal weight as a linear regression on dose,

$$w_i = \alpha_0 + \alpha_1 d_i + \epsilon_i, \quad i = 1, \dots, n,$$

and the second stage modeled the malformation as a logistic regression on dose and the residual from the fetal weight model,

$$\begin{aligned} \text{logit}(\mu_i) &= \beta_0 + \beta_1 d_i + \beta_2 (w_i - \hat{w}_i) \\ &= \beta_0 + \beta_1 d_i + \beta_2 w_i^*. \end{aligned}$$

This model is referred as Model 2.

The B2 method produced a trivial tree. The B1 method produced the tree in Figure 4. The first split was at  $d = 19.69$ , the second split occurred at  $d = 8.21$ , and the third split was at  $w^* = .025$  at node(2, 1). The dose variable was not included in the regression equation in node(2, 2), node(3, 1) and node(3, 2) because each (sub)sample contained only one dose level. Table 3 presents the parameter estimates of fitting Model 2 to the whole sample and to the subsamples at the terminal nodes of the regression trees from the B1 method.

Both  $d$  and  $w^*$  are highly significant ( $p < .01$ ) for the whole sample. With the tree-structured regression, both variables are also significant for the higher dose group ( $d \geq 20$ ) (node(1, 2)). The variable  $w^*$  is significant for  $d = 15$  (node(2, 2)), but it is not significant within the two terminal nodes for  $d = 0$  (node(3, 1) and node(3, 2)). Note that all the splitting tends toward the left side of Figure 4. Various patterns were detected at the lower dose and lower fetal weight.

## 4.3 Dose-Response Risk Estimation

In health risk assessment, a main purpose of fitting a dose-response function is to predict the probability of developmental effect at a given low dose level. Fetal weight is generally obtained after the completion

of the experiment. The estimate obtained from Model 2 is viewed as the dose-response function evaluated at weight  $w = \hat{w}$  ( $w^* = 0$ ). The estimate from Model 2 is an adjusted mean in the context of the analysis of covariance discussed by Urquhart (1982) when a covariate (fetal weight) was affected by the treatment (dose).

Model 2 can be regarded as a conditional dose-response function, conditional on fetal weight. The unconditional dose-response function can be obtained by taking the expectation with respect to  $w^* = (w - \hat{w})$ . However, the exact closed form for the expectation of a cumulative logistic function is not available. By applying the Gaussian approximation to a logistic function as proposed by Zeger, Liang, and Albert (1988), the unconditional dose-response function for Model 2 is

$$\text{logit}(\mu_i) \simeq \frac{\beta_0 + \beta_1 d_i}{[1 + \text{var}(w_i - \hat{w}_i)(\beta_2 c)^2]^{1/2}},$$

where  $c = 16\sqrt{3}/(15\pi)$ . Using the above approximation, the (predicted) dose-response function for the whole sample is

$$\text{logit}(\mu_i) = -2.510 + .082d_i.$$

The dose-response function based on the tree-structured regression procedure is

$$\text{logit}(\mu_i) = \begin{cases} -2.207, & d = 0 \\ -1.367, & d = 15 \\ -3.709 + .122d_i, & d_i \geq 20. \end{cases}$$

Note that for this example, the two terminal nodes at dose 0 were combined for prediction. This has been done after the final tree was obtained. The observed and predicted probabilities of malformation for Model 2 are given in Table 4. For the purpose of comparison, the predicted probabilities for Model 1 are also given. It can be seen that the regression trees gave better predicted values at the lower dose region than the standard logistic regression using the whole sample in both Model 1 and Model 2, with Model 2 performing best in that region. The reduction in the observed proportion at dose 20 is not reflected in the smoothing shown in Figure 1 which is monotonic. The tree in Figure 4 reflects the reduction, however. Table 4 also shows that the trees provide better predictions than the standard logistic regression models at dose 20.

Toxicologists have argued that mechanisms of the action at the high dose region may be different from those at the low dose (Gold, Manley and Ames, 1992). For example, for some chemicals (e.g., formaldehyde and saccharin) mitogenesis occurs only at high doses. For others (e.g., butadiene), carcinogenic effects have been found considerably below the maximum tolerated dose. Krewski, Murdoch and Dewanji (1986) proposed a model-free procedure for low dose risk estimation based on a secant approximation to the slope of the dose-response curve in the low dose region. The range of the low dose region could be chosen between the control and the largest dose below the first dose at which the observed response rate was significantly greater than the response in control. In the tree-structured regression, the low-dose region is determined by a goodness-of-split criterion based on a combination of statistical tests and residual analysis. The tree-structured regression approach provides a refined regression function consistent with the data at lower dose.

## 5 Simulations

A Monte Carlo simulation was conducted to evaluate the performance of the proposed regression tree algorithm. The simulation consisted of five experiments. In Experiments 1 and 2, data were generated from one logistic model. In Experiments 3 and 4, data were generated from two different logistic models. Also, data in Experiments 1 and 3 were generated from a binomial distribution, and data in Experiments 2 and 4 were generated from an over-dispersed binomial distribution. The random binomial data under zero and positive correlations were generated using the method given in Ahn and Chen (1995). Experiments 1 and 2 are designed to study the probability of a Type I error and Experiments 3 and 4 are designed to study the power of the procedure. In addition, Experiment 5 is designed to study change of the power with different choices of two logistic models.

In Experiments 1 and 2, the mean of the response variable was generated from the dose-response function with mean

$$\mu_i = 1/\{1 + \exp[-(\beta_0 + \beta_1 d_i)]\}, \quad i = 1, \dots, n,$$

where  $d_i \in \{0, 1, 2, 3, 4, 5\}$  and  $\beta = (\beta_0, \beta_1) = (-2.944, .758)$ . The values of  $\beta_0$  and  $\beta_1$  gave the mean  $\mu_i = .05$  at  $d_i = 0$  and  $\mu_i = .7$  at  $d_i = 5$ .

In Experiments 3 and 4, the means of the response variable were generated from the two dose-response functions such that

$$\mu_i = \begin{cases} 1/\{1 + \exp[-(\beta_0 + \beta_1 d_i)]\}, & \text{for } d_i = 0, 1, 2 \\ 1/\{1 + \exp[-(\gamma_0 + \gamma_1 d_i)]\}, & \text{for } d_i = 3, 4, 5, \end{cases}$$

where  $(\beta_0, \beta_1) = (-2.944, 0)$  and  $(\gamma_0, \gamma_1) = (-6.736, 1.517)$ . The values of  $\beta_0$  and  $\beta_1$  gave a constant mean  $\mu_i = .05$  for  $d_i \in (0, 2.5)$ , and the values of  $\gamma_0$  and  $\gamma_1$  gave the mean  $\mu_i = .05$  at  $d = 2.5$  and  $\mu_i = .7$  at  $d_i = 5$ . This model will be referred to as Model A1.

In Experiment 5, the means of the response variable were generated from the two dose-response functions such that

$$\mu_i = \begin{cases} 1/\{1 + \exp[-(\beta_0 + \beta_1 d_i)]\}, & \text{for } d_i = 0, 1, 2, 3 \\ 1/\{1 + \exp[-(\gamma_0 + \gamma_1 d_i)]\}, & \text{for } d_i = 4, 5, \end{cases}$$

where  $(\beta_0, \beta_1) = (-2.944, 0)$  and  $(\gamma_0, \gamma_1) = (-11.792, 2.528)$ . The values of  $\beta_0$  and  $\beta_1$  gave a constant mean  $\mu_i = .05$  for  $d_i \in (0, 3.5)$ , and the values of  $\gamma_0$  and  $\gamma_1$  gave the mean  $\mu_i = .05$  at  $d = 3.5$  and  $\mu_i = .7$  at  $d_i = 5$ . This model will be referred to as Model A2.

The simulation design was based on the context of a developmental toxicity experiment. The number of litters per dose group was taken to be 10, giving a total of 60 litters per trial. This design was used because it also represents many other toxicological bioassays. The litter size  $n_i$  was chosen at random using the relative frequency distribution from actual developmental toxicity experimental data given by Haseman and Hogan (1975). The value of  $n_i$  ranges from 1 to 20, and the mean of the distribution is about 12.

Each trial data set was analyzed by the logistic regression tree model for  $\rho = 0$  and  $\rho \neq 0$  using the two bootstrap estimation methods for choosing  $f$  and  $\eta$ . Two hundred simulation data sets were generated. The performance of the proposed procedure was compared with the logistic regression trees by Chaudhuri et al. (1995). The procedure in Chaudhuri et al. (1995) can build trees using one of five criteria: (1) Direct stopping rule, (2) Pruning by test sample, (3) Taking the large tree, (4) Pruning by cross-validation, and (5) Pruning by Efron optimism (AIC). See Lo (1993) for further details on the criteria. In this paper, we examine criteria 1, 4 and 5 only, since criterion 3 gives large trees without pruning and a test sample is needed for criterion 2. Two hundred samples were generated to test the procedures by Chaudhuri et al. (1995).

For all the examples in this paper, 10-fold cross-validation was used in our procedure and the value of  $\alpha$  in the bootstrap was chosen to be .05.

## 5.1 One Logistic Model Without Over-Dispersion

Because the data are generated from a logistic model for all the cases, it is expected that a single logistic regression fit is sufficient and no split is required. Simulation results are given in the fourth column of Table 5. For the procedure without over-dispersion, the probabilities of a Type I error for the B1 and B2 methods were 4.5% and 5.5%, respectively. The probability of a Type I error appear to be quite satisfactory. For the procedure with over-dispersion, the probabilities of a Type I error were 4.5% and 3%, respectively, for the B1 and B2 methods. The B2 method of the procedure with over-dispersion seems to be a little more conservative than the procedure without over-dispersion.

The third column of Table 6 gives the results from the simulation for the procedure by Chaudhuri et al. (1995). Pruning by cross-validation gave as many as 95 non-trivial trees out of 200 trials. However, the direct pruning (the threshold value is  $\alpha = .1$ ) and pruning by Efron optimism gave 6.5% and 7%, respectively, of the probabilities of a Type I error. The latter two procedures are comparable to the proposed procedure without over-dispersion.

## 5.2 One Logistic Model With Over-Dispersion

The response variable was generated from an over-dispersed binomial distribution with  $\rho = .1$ ,  $\rho = .3$ , and  $\rho = .5$ .

The last three columns of Table 5 show the simulation results. For the procedure without over-dispersion, the B1 method gave 5% and the B2 method gave 7.5% of the probabilities of a Type I error for the data with

$\rho = .1$ . For the data with  $\rho = .3$ , the B1 method gave 10% and the B2 method gave 4.5% of the probabilities of a Type I error. For the data with  $\rho = .5$ , the B1 and B2 methods gave 11.5% and 3% of the probabilities of a Type I error, respectively. The high probability of a Type I error for the high over-dispersion for the B1 method is not due to the limited number of simulation trials. For  $\rho = .3$  with the B1 method, 9.2% of the probability of a Type I error was obtained in one thousand trials. (To be consistent with other simulations, this value is not given in the table.) For the procedure with over-dispersion, the probabilities of a Type I error for the B1 and B2 methods were 4% and 6.5%, respectively for the data with  $\rho = .1$ ; 3% and 5.5%, respectively for the data with  $\rho = .3$ ; and 6% and 6.5%, respectively for the data with  $\rho = .5$ . For higher intra-cluster correlations ( $\rho = .3$  and  $\rho = .5$ ), the logistic regression tree procedure with over-dispersion controlled the probability of a Type I error better than the procedure without over-dispersion for the B1 method.

Regarding the simulation for the procedure by Chaudhuri et al. (1995), the last column of Table 6 shows that all the three methods considered here gave more than 60% of the probability of a Type I error. Although the method in Chaudhuri et al. (1995) is a good procedure for the data without over-dispersion, it cannot control the probability of a Type I error adequately if the data possess over-dispersion. The proposed regression tree performs much better for handling over-dispersed data. Even for the procedure without over-dispersion, the proposed procedure performs better than that of Chaudhuri et al. (1995).

### 5.3 Two Logistic Models Without Over-Dispersion

In this simulation experiment, it is expected that a single logistic regression fit to the root node is not adequate and each tree should have a split.

The simulation results for Model A1 are shown in the fourth column of Table 7. For the procedure without over-dispersion, the powers (percentage of the trees with at least one split) were 92% for the B1 method and 67.5% for the B2 method. For the procedure with over-dispersion, the powers were 91.5% and 82.5% for the B1 and B2 methods, respectively. The logistic regression tree procedure with over-dispersion performed as well as the tree procedure without over-dispersion for the B1 method. For the B2 method, the power was substantially larger for the procedure with over-dispersion than the procedure without over-dispersion.

The procedure in Chaudhuri et al. (1995) was investigated for the data with  $\rho = 0$ . The powers from the direct stopping rule, pruning cross-validation, and pruning by Efron optimism of their procedure were over 98% (see Table 8). Their procedure seemed to have more power than the proposed procedure for data without over-dispersion.

### 5.4 Two Logistic Models With Over-Dispersion

Given the mean values, the response variable was generated from Model A1 with  $\rho = .1$ ,  $\rho = .3$  and  $\rho = .5$ .

The last three columns of Table 7 show the simulation results. For the procedure without over-dispersion, the B1 method gave 62.5% and the B2 method gave 33% for the power for the data with  $\rho = .1$ . For the data with  $\rho = .3$  and  $\rho = .5$ , the powers were substantially reduced. For the tree procedure with over-dispersion, the reduction of the power was less severe as  $\rho$  increased. These results show that some power may be gained by explicit modeling of the over-dispersion. The power is improved by using the adequate model. Chaudhuri et al.'s (1995) procedure is not reported here because its Type I error rate was felt to be too large under the setting  $\rho \neq 0$ .

### 5.5 Further Study on the Power

Further simulations were conducted to determine at what point the proposed method is able to discriminate between two logistic models. The response variable was generated from Model A2 with  $\rho = .3$  and  $\rho = .5$  with the mean values given in the beginning of Section 5. The smaller over-dispersions were not considered in this section because they had larger power in the previous subsections. Since the B2 method appears underpowered compared to the B1 method, the former is not considered in this experiment. Table 9 shows the simulation results. The procedure without over-dispersion gave 57% and 25% for the power for the data with  $\rho = .3$  and  $\rho = .5$ , respectively. The procedure with over-dispersion gave 81.5% and 41.5% for the power for the data with  $\rho = .3$  and  $\rho = .5$ , respectively. For the procedure with over-dispersion, higher values of  $\gamma_1$

in Model A2 are expected to have better power. The power for  $\rho = .5$  is lower than that for  $\rho = .3$ , but the former is an extreme case. In this experiment, the procedure with over-dispersion gave a satisfactory power for the data with  $\rho = .3$ .

## 6 Discussion

This paper presents a tree-structured regression algorithm for the analysis of effects of covariates on over-dispersed binomial response data using a quasi-likelihood logistic approach. The splitting criterion in cross-validation is based on the quasi-deviance function. Unlike the Chaudhuri et al. (1995) procedure, which uses the cross-validation estimate in CART's pruning method to find the final tree, the proposed procedure uses a cross-validators multistep look-ahead stopping rule and bootstrap resampling to determine the proper depth of a tree. In the CART, after a large tree is constructed, a nested sequence of subtrees is obtained by progressively deleting branches according to the pruning method. In the proposed method, pruning is done under the  $V$ -fold cross-validation at each node. The cross-validation decides whether the node should be split. In other words, the CART method uses pruning only once, but our procedure uses pruning at each node. A referee pointed out that in the present paper, quasi-likelihood methods are not incorporated when the  $t$ -test and Levene's test are applied to ordinal covariates such as dose and potentially over-dispersed continuous covariates such as average fetal weight. Using a non-parametric method based upon ranks for the former and a quasi-likelihood approach for the latter could be worth considering for a future study. As the referee mentioned, this would be keeping in the spirit of the quasi-likelihood method for modeling the binomial outcomes.

Chaudhuri et al. (1995) uses pseudo-residuals from the smoothed binary response variable in the residual analysis. However, because the proposed algorithm uses the proportion of successes (or failures) and the number of trials, it requires data to be clustered. Nevertheless, the residuals from the logistic model are already wellshaped and do not need smoothing. In the binary regression, the residuals have two values only, but in the binomial regression the residuals have more values and they are close to those of the normal residuals. The proposed procedure uses the cluster-based model and selects its splits by analysis of the distributions of the residuals. This procedure can be adopted to modeling binary data when the within-cluster covariates are different. If no intracluster correlation is involved, Chaudhuri et al.'s (1995) procedure can be used to analyze the binary data as an alternative.

For the example given in Section 4, the standard errors at the terminal node increased after each split. Because the split is in favor of a better fit, the power does not always decrease after a split. The node is not split if better fits cannot be obtained in the children nodes. Thus, the power to detect effects is already captured in the splits. Ahn and Loh (1994) found that, in a proportional hazards regression tree, a covariate is insignificant at the root node, but it becomes significant at a terminal node. Ahn (1994) and Chaudhuri et al. (1994) also show similar results with different regression tree models. As we described in Section 3.2, the deviance function is used in cross-validation to check if any further split needs to occur. In cross-validation, a penalty ( $f$ ) is given to the non-trivial trees (trees with split) when they are compared with the tree without split. The penalty function is obtained from the bootstrap. This sequence of a procedure in the stopping rule prevents the tree from having excessive splits in growing the tree.

The example given in Section 4 has more doses than usual. The sample size of the data is also greater than that in typical toxicity experiments. In Figure 4, however, the sample was split into low- (0 and 15 mg/kg/day) and high-dose groups. The low-dose group has 175 cases and only two dose levels. The sample in that node was further split into 0 mg/kg/day and 15 mg/kg/day dose groups. The 0 mg/kg/day dose group (with 89 cases) was split again. The sample in node(1, 1) (0 and 15 mg/kg/day dose group with 175 cases) might be closer to the usual toxicology data, and the left subtree (the subtree start with node(1, 1)) of the figure shows that the example can reflect the performance of the method for use in more modest data. The sample size of the data in our simulation was only 60, but the probability of a Type I error was proper and the power of the method was sufficient for Model A2 with  $\rho = .3$ . For the proportional hazards regression model, Ahn and Loh (1994) grew regression trees using data with 157 cases.

The simulation results show that the proposed algorithms control the probability of a Type I error satisfactorily. The B1 method performs better than the B2 method. It controls the probability of a Type I error well and has better power. Therefore, we recommend the B1 method be used. The simulation also

shows that the procedure based on the standard binomial model performs satisfactorily for the data without over-dispersion, but it fails for the data with over-dispersion. The procedure for over-dispersion performs well for data without over-dispersion as well as for data with over-dispersion. The simulations show that Chaudhuri et al.'s (1995) procedure controls the probability of a Type I error well for some criteria of building trees and gives excellent power if the binary responses are not correlated. However, although the procedure proposed here uses proportions and hence fewer data points, it controls the probability of a Type I error very well for both correlated and uncorrelated data.

One of the uses of bioassay data is to predict the incidence of toxic effects at low doses that humans may encounter. A well-known problem in quantitative risk estimation is that the shape of the dose-response in the low-dose range cannot be observed with adequate precision. As we found in Section 4, the proposed method identifies the low-dose effect better than the conventional methods. Note that the higher dose groups (45 mg/kg/day and 60 mg/kg/day) have fewer litters than the lower dose groups. This is because the compound was so toxic at high doses that it killed dams and/or entire litters. However, this is not a serious problem in interpreting the data if the method is able to distinguish the different shapes of the dose-response function for the low- and high-dose groups because only the low dose region is of primary interest.

The primary application of the procedure illustrated has been addressed specifically for dose-response modeling of data from a teratological study. The regression tree algorithm can be used to model other toxic effects. The linear logistic dose-response function used in this paper is for mathematical convenience. Other dose-response functions such as one-hit, Weibull, or probit models can also be used. Ideally, the dose-response model should be derived from sound biological theories. Given the inherent uncertainty regarding the exact dose-response function in the low dose region, the tree-structured regression analysis can provide a simple approach for predicting risk and determining a safe dose level.

The entire algorithm is coded in a FORTRAN program. Computation for the simulations and data analysis were performed on an Alpha workstation. However, the program can also be run on an IBM compatible PC. It took approximately 2 hours and 35 minutes to obtain the tree in Figure 4 (Model 2) in Section 4.2, using the Alpha workstation. Computing time for each data set of the simulations in Section 5 was about 20 to 25 minutes on the Alpha workstation. A DOS executable of the program is available on request.

## Acknowledgements

This work was done while the first author was in Division of Biometry and Risk Assessment, National Center for Toxicological Research. The authors wish to thank Drs Wei-Yin Loh and Wenda Lo for kindly providing their logistic regression tree program. The authors appreciate Professor Wei-Yin Loh for his helpful comments on this paper. The authors are also grateful to the associate editor and two anonymous referees for many helpful comments that substantially improved this paper.

## References

- Ahn, H. (1994). Tree-structured extreme value model regression. *Communications in Statistics - Theory and Methods*, **23**, 153-174.
- Ahn, H., and Chen, J. J. (1995). Generation of over-dispersed and under-dispersed binomial variates. *Journal of Computational and Graphical Statistics* **4**, 55-64.
- Ahn, H., and Loh, W.-Y. (1994). Tree-structured proportional hazards regression modeling. *Biometrics* **50**, 471-485.
- Becker, R. A., Chambers, J. M., and Wilks, A. R. (1988). *The New S Language*. Pacific Grove, California: Wadsworth.
- Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*. Belmont, California: Wadsworth.

- Catalano, P. J., and Ryan, L. M. (1992). Bivariate latent variable models for clustered discrete and continuous outcomes. *Journal of the American Statistical Association* **87**, 651-658.
- 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. (1995). Generalized regression trees. *Statistica Sinica* **5**, 641-666.
- Chen, J. J. (1993). A malformation incidence dose-response model incorporating fetal weight and/or litter size as covariates. *Risk Analysis* **13**, 559-564.
- Ciampi, A., and Thiffault, J. (1989). Pruning regression trees for censored survival data: The RECPAM approach. *Communications in Statistics - Theory and Methods* **18**, 3378-3388.
- Collett, D. (1991). *Modelling Binary Data*. New York: Chapman and Hall.
- Davis, R. B., and Anderson, J. R. (1989). Exponential survival trees. *Statistics in Medicine* **8**, 947-961.
- Emrich, L. J., and Piedmonte, M. R. (1991). A method for generating high-dimensional multivariate binary variables. *The American Statistician* **45**, 302-304.
- Gold, L. S., Manley, N. B., and Ames, B. N. (1992). Extrapolation of carcinogenicity between species: Qualitative and quantitative factors. *Risk Analysis* **12**, 579-588.
- Haseman, J. K., and Hogan, M. D. (1975). Selection of the experimental unit in teratology studies. *Teratology* **12**, 165-172.
- Holson, J. F., Gaines, T. B., Nelson, C. J., LaBorde, J. B., Gaylor, D. W., Sheehan, D. M., and Young, J. F. (1992). Developmental toxicity of 2,4,5-trichlorophenoxyacetic acid I: Multireplicated dose-response studies in four inbred strains and one outbred stock of mice. *Fundamental and Applied Toxicology* **19**, 286-297.
- Krewski, D., Murdoch, D., and Dewanji, A. (1986). *Statistical modeling and extrapolation of carcinogenesis data*. In: *Modern Statistical Methods in Chronic Disease Epidemiology*, S. H. Moolgavkar and R. L. Prentice (eds), 259-282. New York: Wiley.
- Kupper, L. L., Portier, C., Hogan, M. D., and Yamamoto, E. (1986). The impact of litter effects on dose-response modeling in teratology. *Biometrics* **42**, 85-98.
- LeBlanc, M., and Crowley, J. (1992). Relative risk trees for censored survival data. *Biometrics* **48**, 411-426.
- Levene, H. (1960). Robust tests for equality of variances. In *Contributions to Probability and Statistics*, I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow and H. G. Mann (eds), 278-292. Stanford, California: Stanford University Press.
- Liang, K.-Y., and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13-22.
- Lo, W.-D. (1993). Logistic regression trees. Unpublished Ph.D. dissertation, University of Wisconsin-Madison, Dept. of Statistics.
- Loh, W.-Y. (1991). Survival modeling through recursive stratification. *Computational Statistics and Data Analysis* **12**, 295-313.
- 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* (2nd ed.). London: Chapman and Hall.
- Morgan, B. J. T. (1992). *Analysis of Quantal Response Data*. London: Chapman and Hall.
- Morgan, J. N., and Sonquist, J. A. (1963). Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association* **58**, 415-434.
- Moulton, L. H., and Zeger, S. L. (1989). Analyzing repeated measures on generalized linear models via the bootstrap. *Biometrics* **45**, 381-394.
- Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics* **44**, 1033-1048.
- Rosner, B., and Milton, R. C. (1988). Significance testing for correlated binary outcome data. *Biometrics* **44**, 505-512.
- Ryan, L. M., Catalano, P. J., Kimmel, C. A., and Kimmel, G. L. (1991). Relationship between fetal weight and malformation in developmental toxicity studies. *Teratology* **44**, 215-223.
- Segal, M. R. (1988). Regression trees for censored data. *Biometrics* **44**, 35-47.
- Segal, M. R., and Bloch, D. A. (1989). A comparison of estimated proportional hazards models and regression trees. *Statistics in Medicine* **8**, 539-550.
- Stiratelli, R., Laird, N., and Ware, J. H. (1984). Random-effects models for serial observations with binary response. *Biometrics* **40**, 961-971.
- Urquhart, N. S. (1982). Adjustment in covariance when one factor affects the covariate. *Biometrics* **38**, 651-660.
- Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models and the Gauss-Newton method. *Biometrika* **61**, 439-447.
- Zeger, S. L., Liang, K.-Y., and Albert, P. S. (1988). Models for longitudinal data: A generalized estimating equation approach. *Biometrics* **44**, 1049-1060.

## Appendix

### Algorithm for the Trees

#### A.1 Splitting

The following procedure is performed at each node.

1. A logistic regression model is fitted to the data in the node.
2. The residuals  $\epsilon_i = y_i - \mu_i$  are calculated in the node.
3. An observation belongs to class 1 if its residual is larger than the median of the residuals for the sample and to class 2 otherwise.
4. For each covariate, perform two-sample  $t$ -tests on the two groups of observations for differences in means and variances (the latter test is Levene's, 1960).
5. The  $P$ -value from the larger of the  $t$ -statistic and Levene's statistic is computed for each covariate.
6. The covariate selected to split the node is the one that yields the smallest  $P$ -value. Suppose the  $k$ th covariate yields the smallest  $P$ -value. The data in the node are split into two parts, with one subset containing all cases with the  $k$ th covariate value less than  $c$  and the other subset containing the remaining cases, where  $c$  is the average of the two sample means.

The above process is repeated at each subsequent node until either cross-validation determines not to split the node or there are too few cases left at the node.

## A.2 Cross-Validation

Let  $\text{node}(i, j)$  and  $\mathcal{L}(i, j)$  be the current node and the sample in it, respectively. The cases in  $\mathcal{L}(i, j)$  are randomly divided into  $V$  subsets  $\mathcal{L}_1, \dots, \mathcal{L}_V$  each containing nearly the same number of cases. The following process is repeated for  $v = 1, \dots, V$ .

1. Grow a large tree  $T_{v0}$  using the cases in  $\mathcal{L}^{(v)} = \mathcal{L} - \mathcal{L}_v$ . A node is terminal in this tree if there are too few cases in the node or the information matrix is almost singular at some stage of the iterations in the Newton-Raphson method. Let  $p_{ij}$  be the smallest  $P$ -value from two-sample  $t$ -tests for mean and variance for  $\text{node}(i, j)$ . Suppose there are  $r$  distinct values of  $p_{ij}$ 's. Sort the  $P$ -values and add  $p_0 = 0$  and  $p_{r+1} = 1$  so that  $0 = p_0 < p_1 \leq \dots \leq p_r < p_{r+1} = 1$ .
2. Compute  $\delta_l = (p_l + p_{l+1})/2$  for  $l = 0, 1, \dots, r$ .
3. Prune  $T_{v0}$  at level  $\delta_l$  to obtain a tree  $T_l$  and compute the cross-validation estimate  $R^{CV}(v, l)$  of  $T_l$  using  $\mathcal{L}_v$  as test sample as follows. Let  $T_{r+1} = T_{v0}$ .
  - Loop over  $k = r, r-1, \dots, 1, 0$ .
    - (a) Starting from the lowest level to the root node of  $T_{k+1}$ , do the following.
      - i. At level  $i$ , loop over  $j = 1, \dots, 2^i$ , starting with the left node. At  $\text{node}(i, j)$ ,
        - A. If the node is intermediate and the two children nodes are terminal, let  $p$  be the  $P$ -value of the split at the node. If  $p < \delta_k$ , go to the next node. If  $p \geq \delta_k$ , delete the two children nodes and make  $\text{node}(i, j)$  terminal.
        - B. Otherwise, go to the next node.
      - ii. End the loop for  $j$ .
    - (b) End the loop for  $i$ . This gives the pruned tree  $T_k$  at level  $\delta_k$ .
  - End the loop for  $k$ .
4. Let  $f \in (0, 1)$  be the fractional reduction (obtained from the bootstrap) in the deviance.
  - (a) Set  $\theta(v) = 0$ .
  - (b) Loop over  $k = 1, \dots, r$ .
  - (c) If  $R^{CV}(v, k) < (1 - f)R^{CV}(v, 0)$ , set  $\theta(v) = 1$  and exit.
  - (d) Otherwise increase  $k$  to  $k + 1$  and go to (c).

Let  $\eta \in (0, 1)$  be the splitting threshold obtained from the bootstrap and  $\theta = \sum_{v=1}^V \theta(v)$ . If  $\theta > \eta V$ , the node is split; otherwise it is declared terminal.

## A.3 Bootstrap Parameter Selection

The estimating procedure of the B1 method is discussed here. The B2 method estimates the parameters the same way. Using 100 bootstrap samples, the following steps are performed at each value of  $f$ .

Fix  $f = \eta$  and let the estimate of (2) be

$$g_1(f) = \hat{\alpha}(f, \eta) = \hat{P}(\text{Split the root node} | H_0 : \text{The root node should not be split}). \quad (3)$$

1. Starting from  $f = \eta = 0$ , increase the value of  $f$  by .1 and evaluate (3). Because the tree size gets larger as  $f$  and  $\eta$  values become smaller,  $g_1$  is a nonincreasing function of  $f$ .
2. Stop at  $f = f'$  such that  $g_1(f') \leq \alpha$  and  $g_1(f' - .1) > \alpha$ .
3. Starting from  $f' - .1$ , increase the value of  $f$  by .01 and evaluate (3).
4. Stop at  $f = f_0$  such that  $g_1(f_0) \leq \alpha$  and  $g_1(f_0 - .01) > \alpha$ .
5. Take  $f = \eta = f_0$  as the estimate of  $(f, \eta)$ .

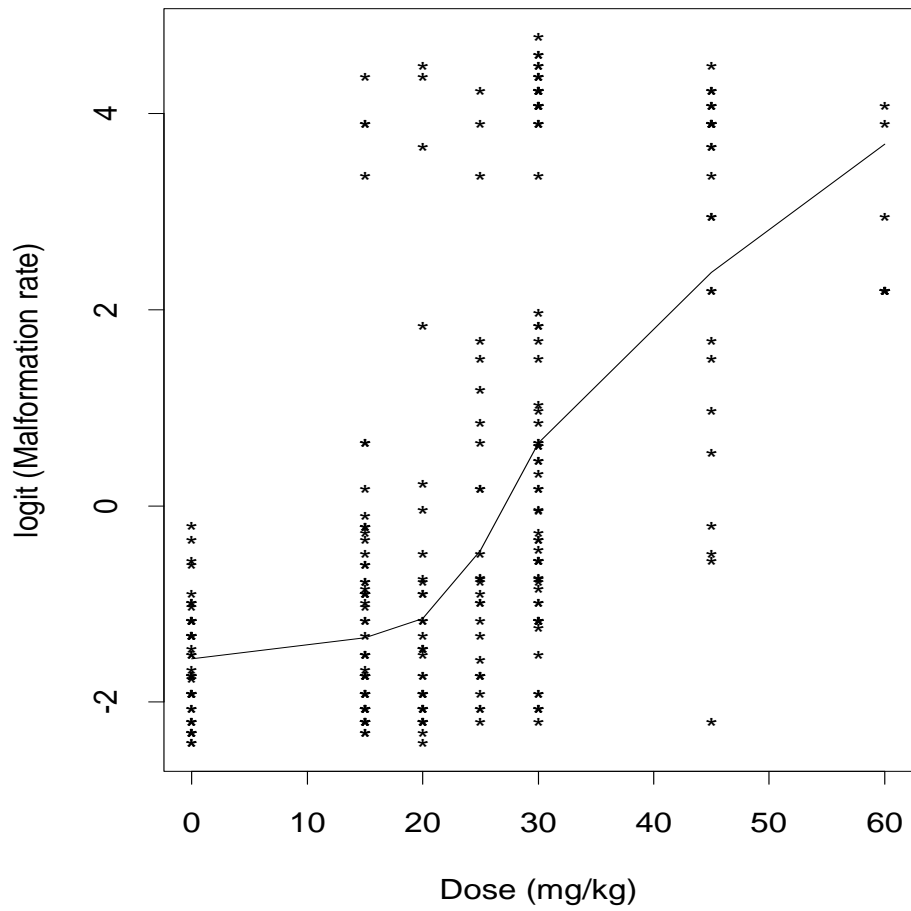


Figure 1: Scatterplot of logit(malformation) versus dose with a smooth curve superimposed for data from a developmental toxicology study of 2, 4, 5-trichlorophenoxyacetic acid for A/JAX strain.

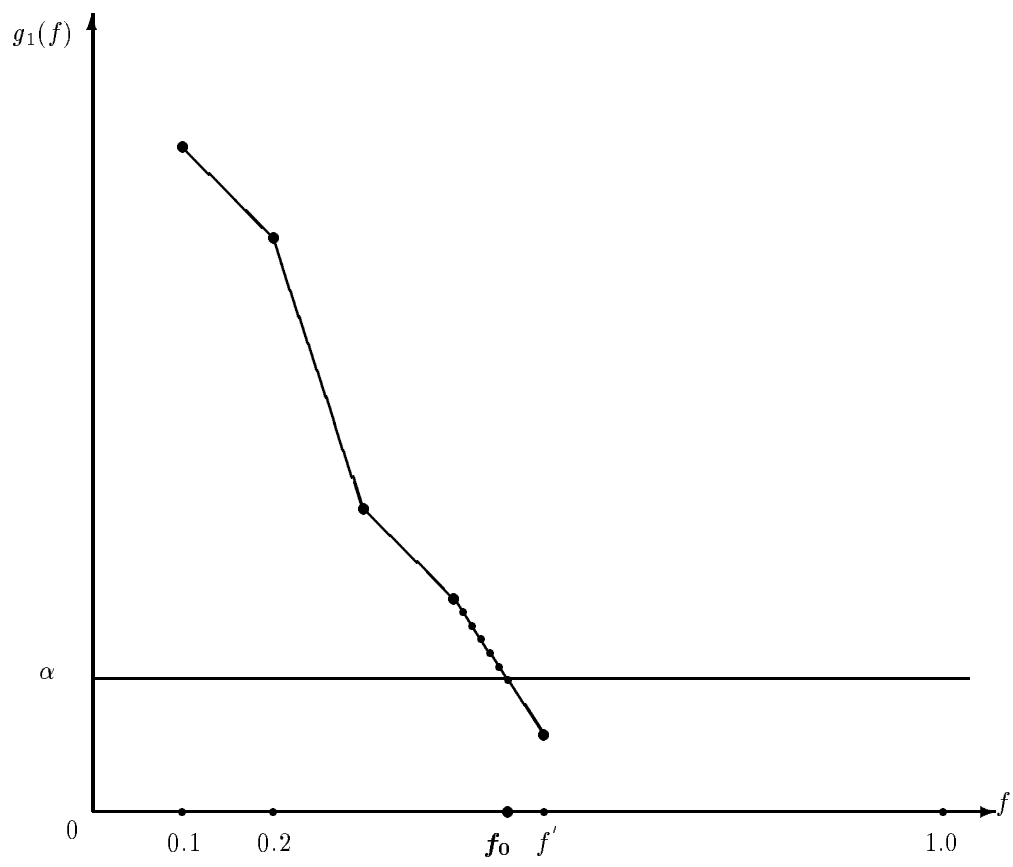


Figure 2: Searching for  $f$  in the bootstrap estimation method (B1).

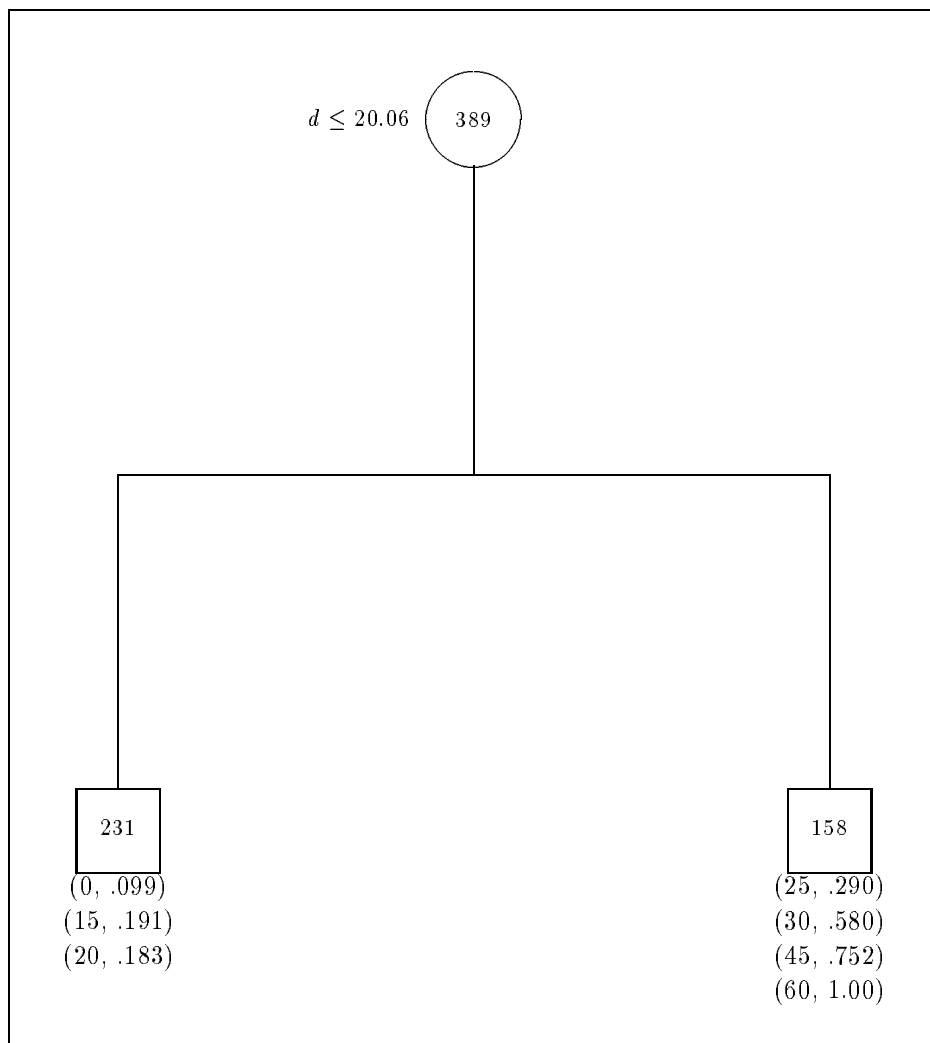


Figure 3: Logistic regression tree with dose response model (Model 1). The numbers within circle or squares are sample sizes. The numbers at the bottom of the squares are dose levels (first entry of each pair) and malformation rates (second entry) for the node. The bootstrap method used here is B1 with  $f = \eta = .10$ .

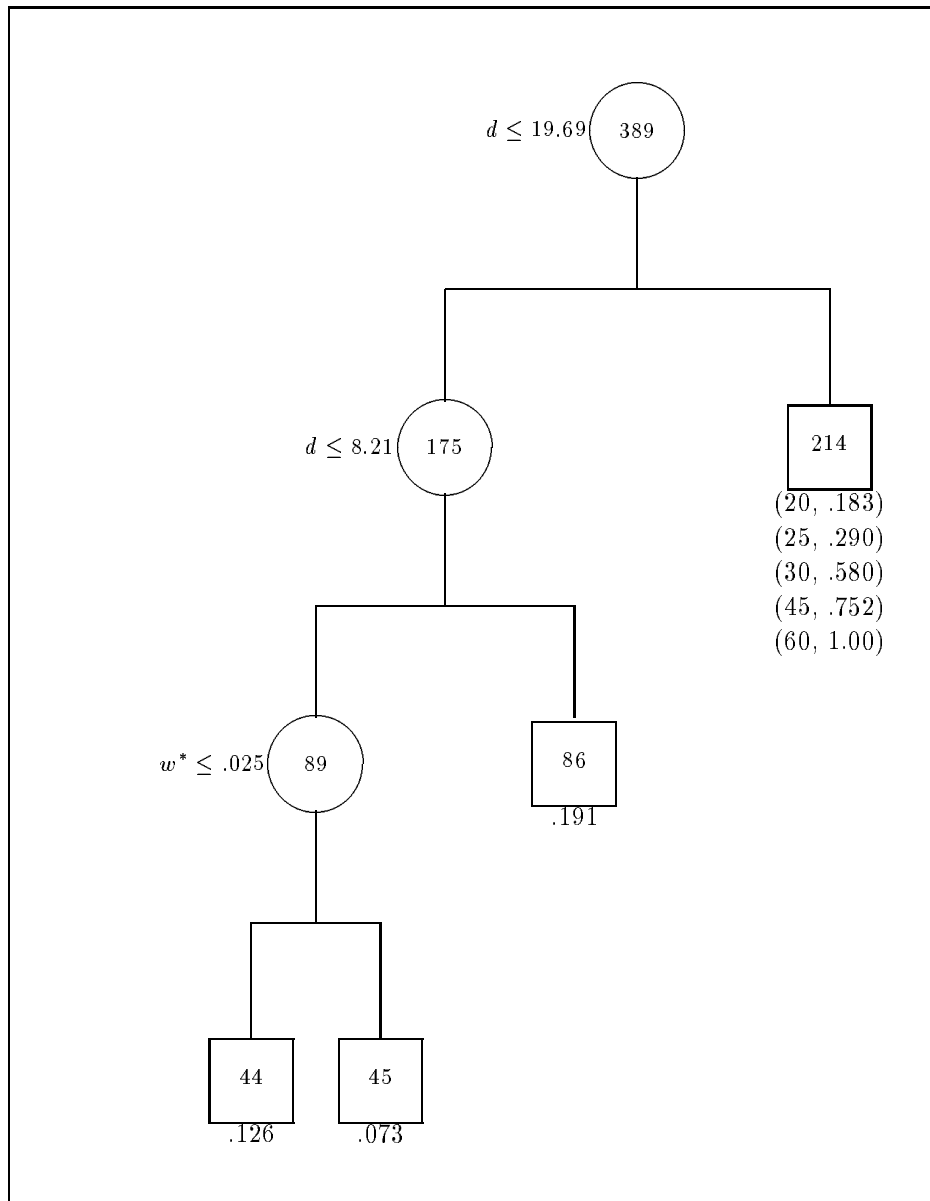


Figure 4: Logistic regression tree with the Gaussian regression chain model (with over-dispersion; Model 2). The numbers within circles or squares are sample sizes. The numbers at the bottom of the squares are malformation rates. In the case that one node includes more than one dose level, the first entry of the pair is dose level and the second entry is malformation rate. The bootstrap method used here is B1 with  $f = \eta = .12$ .

Table 1: Data from a developmental toxicology study of 2, 4, 5-trichlorophenoxyacetic acid in A/JAX mice.\*

Dose	$x$	$n$	wgt	$x$	$n$	wgt	$x$	$n$	wgt	$x$	$n$	wgt	$x$	$n$	wgt	$x$	$n$	wgt	
0	1	6	58.67	0	7	61.71	0	8	61.38	1	10	65.20	1	7	60.00	0	6	68.83	
	2	12	48.33	0	3	66.33	1	10	62.30	0	7	59.00	2	10	55.50	2	7	46.14	
	2	9	64.00	1	8	54.50	0	10	61.50	0	7	62.71	1	7	63.86	1	6	53.50	
	2	9	54.89	1	5	50.20	0	10	58.60	0	8	58.13	1	6	59.33	2	9	59.44	
	0	7	57.71	1	7	61.71	0	8	52.63	0	9	63.67	1	10	60.60	1	9	58.22	
	2	7	69.43	0	9	55.22	1	2	69.50	1	8	53.63	1	11	55.36	3	11	61.64	
	2	7	58.29	0	9	67.67	0	11	56.36	2	8	62.13	0	10	62.40	1	6	42.00	
	0	8	60.38	0	7	59.71	0	8	66.25	3	8	56.00	1	5	57.80	0	10	62.50	
	1	9	68.33	0	8	57.25	1	9	57.56	0	7	43.00	0	7	62.29	0	9	63.56	
	2	8	49.13	1	7	67.14	1	10	61.00	0	8	57.00	2	8	57.50	2	8	63.75	
	1	11	56.18	2	13	60.00	3	7	55.86	0	6	56.67	0	6	59.50	0	10	56.30	
	0	8	48.63	0	9	57.89	0	7	50.00	2	9	49.44	0	7	56.00	0	8	57.63	
	0	6	50.00	0	8	62.75	0	9	54.44	0	9	58.78	2	8	45.25	4	11	52.64	
	3	10	58.00	2	9	56.22	0	11	58.82	0	9	66.00	0	9	67.11	1	7	55.14	
	0	8	59.63	0	8	62.75	0	11	62.82	0	8	57.38	1	6	55.67				
	15	0	10	56.20	0	9	59.89	0	8	56.50	2	12	58.75	1	8	63.38	0	8	60.63
		2	9	51.33	2	8	48.88	3	10	57.70	0	7	57.14	0	8	58.00	0	9	66.33
0		7	55.57	4	9	52.22	0	6	60.00	0	6	62.83	4	11	63.27	3	10	55.30	
5		5	40.20	4	11	54.36	3	10	52.50	1	9	56.78	3	3	36.33	5	5	35.80	
0		6	67.17	5	11	52.09	0	9	83.56	1	9	55.89	2	6	58.17	6	9	40.22	
1		10	53.20	1	5	56.00	0	3	58.00	0	8	55.75	0	3	59.00	0	7	63.86	
2		5	60.00	1	6	60.17	1	10	54.30	1	7	47.86	0	10	56.70	0	6	59.67	
0		9	57.56	0	8	53.88	1	7	54.71	3	7	54.43	1	9	57.33	2	8	65.00	
0		9	52.56	0	5	58.80	1	6	57.67	6	9	47.56	0	6	47.67	0	7	56.71	
0		5	54.00	3	11	53.27	5	9	50.22	1	9	56.78	1	9	58.11	1	5	51.40	
8		8	40.88	0	6	57.33	1	6	61.83	3	10	57.50	5	11	48.73	0	7	99.00	
0		7	67.71	1	5	60.40	1	8	54.13	1	8	58.38	1	6	58.00	2	4	51.50	
1		7	62.00	1	7	50.00	2	6	60.17	0	9	58.44	0	8	61.38	1	3	59.33	
5		11	49.18	0	9	60.33	1	8	54.38	0	6	58.83	2	7	55.86	0	11	55.73	
1		6	53.50	1	8	61.13													
20	0	8	61.75	1	5	55.80	1	7	55.00	2	9	56.78	5	10	54.80	2	5	56.80	
	0	5	58.20	1	10	50.00	1	7	56.00	1	8	59.38	2	10	51.10	3	9	55.44	
	0	7	55.71	4	4	44.00	4	7	32.14	9	9	43.33	0	8	59.63	0	7	59.43	
	1	6	53.67	2	8	50.75	1	9	60.11	1	8	60.38	2	6	43.17	0	6	54.50	
	1	8	63.00	7	8	46.38	8	8	35.38	0	9	50.11	0	8	64.25	0	7	58.43	
	3	10	53.20	0	9	54.44	0	9	55.33	2	8	55.25	0	6	81.67	0	9	55.11	
	1	6	48.00	0	8	49.00	0	4	44.00	1	8	59.63	1	11	53.18	1	7	55.86	
	0	9	61.78	0	8	60.50	1	9	62.78	2	10	59.50	1	9	54.67	0	8	55.88	
	0	5	60.00	2	10	50.40	0	6	63.00	1	7	63.43	0	4	67.50	3	10	53.30	
	1	9	61.11	0	6	50.50													

(To be continued)

Table 1: (continued)

Dose	$x$	$n$	wgt	$x$	$n$	wgt	$x$	$n$	wgt	$x$	$n$	wgt	$x$	$n$	wgt	$x$	$n$	wgt
25	2	6	51.67	0	10	82.20	0	8	49.00	7	7	50.00	1	7	52.57	2	5	54.60
	1	9	61.00	1	8	51.88	0	5	48.60	0	3	61.00	4	12	52.00	0	6	65.33
	0	6	55.83	1	6	56.17	6	9	50.44	1	6	61.33	7	9	53.56	1	8	50.88
	0	7	67.43	0	6	64.50	3	3	68.33	2	7	64.57	3	9	47.44	2	7	52.57
	6	7	44.43	5	5	39.20	3	10	57.80	5	6	44.50	2	9	54.89	5	9	40.89
	0	6	57.00	5	7	48.71	0	8	77.00	0	7	58.00	5	9	47.56	1	6	60.67
	0	7	51.29	2	8	54.75	2	11	53.36	0	9	75.00						
	30	6	6	46.33	6	6	38.67	4	10	59.10	2	6	50.00	1	8	55.63	4	8
6		9	43.44	2	8	60.38	2	8	50.75	1	5	56.40	2	8	50.63	5	8	51.63
6		9	41.00	3	8	63.25	4	9	55.78	1	9	52.56	5	5	32.60	4	12	59.67
1		4	43.25	7	7	45.57	0	5	56.80	6	8	48.50	0	7	55.57	2	7	52.71
10		10	49.50	8	8	40.88	5	6	57.67	5	8	62.63	5	5	53.40	1	8	61.50
1		7	59.86	5	10	54.00	6	6	44.50	1	3	51.00	12	12	41.25	2	8	55.38
9		9	47.11	7	8	53.50	3	7	55.00	3	3	32.33	8	9	46.89	3	5	47.40
5		5	41.60	3	8	48.63	1	8	46.00	1	7	54.57	2	7	54.29	7	7	53.14
7		7	43.71	6	6	34.50	2	8	83.00	3	7	52.00	3	4	59.00	0	7	66.43
3		9	54.33	10	10	43.90	5	9	47.00	8	8	56.63	4	6	47.33	3	8	59.25
0		4	51.25	5	7	51.57	0	6	56.83	7	8	38.25	6	6	35.00	7	7	49.29
7		7	48.71	8	8	44.38	5	9	47.89	7	7	46.00	0	4	56.00	9	9	43.78
3		8	51.50	6	7	45.43	4	6	48.67	0	7	49.86						
45		7	7	45.57	1	1	43.00	1	9	51.67	3	4	49.00	9	9	40.78	6	6
	5	5	41.40	6	6	39.83	5	5	46.60	6	6	40.50	6	7	53.29	5	6	50.33
	2	3	47.00	5	5	44.20	7	7	42.00	4	4	37.00	5	5	51.60	3	3	39.33
	1	1	42.00	5	5	43.60	0	4	63.25	1	2	49.50	0	1	60.00	0	1	33.00
	4	4	39.00	5	5	48.20	2	2	33.50	7	7	43.71	0	1	52.00	2	5	65.60
	2	2	47.00	0	11	82.09	3	8	51.00									
60	1	1	36.00	1	1	38.00	6	6	46.50	2	2	44.00	1	1	30.00	1	1	39.00
	5	5	45.20	1	1	32.00	1	1	43.00									

\* Dose: in mg/kg/day;  $x$ : number of malformations;  $n$ : number of fetuses per litter; wgt: average fetal body weight in (g/100).

Table 2: Regression estimates for the whole sample and the samples at the terminal nodes of the tree for Model 1 in Figure 3.

Procedure	Node	Variable	Parameter estimate	S.E. (Model)	$\hat{\beta}/\text{S.E.}$ (Model)	S.E. (Robust)	$\hat{\beta}/\text{S.E.}$ (Robust)
$\rho = 0$	Whole sample (0,1)	intercept	-2.6324	.1059	-24.85	.1700	-15.49
		dose	.0851	.0043	19.76	.0078	10.96
	$d \leq 20$ (1,1)	intercept	-2.1644	.1190	-18.19	.1335	-16.21
		dose	.0402	.0082	4.92	.0109	3.70
	$d \geq 25$ (1,2)	intercept	-2.9541	.3507	-8.42	.8025	-3.68
		dose	.0998	.0115	8.72	.0273	3.65
$\rho$ unspecified	Whole sample (0,1)	intercept	-2.6630	.1831	-14.55	.1675	-15.90
		dose	.0869	.0072	12.02	.0073	11.97
		$\rho$	.2990				
	$d \leq 20$ (1,1)	intercept	-2.1487	.1906	-11.27	.1326	-16.20
		dose	.0391	.0130	3.00	.0109	3.59
		$\rho$	.2184				
	$d \geq 25$ (1,2)	intercept	-2.9736	.5955	-4.99	.6388	-4.65
		dose	.1004	.0190	5.28	.0210	4.78
		$\rho$	.4078				

Table 3: Regression estimates for the whole sample and the samples at the terminal nodes of the tree for Model 2 in Figure 4.

Node	Variable	Parameter estimate	S.E. (Model)	$\hat{\beta}/\text{S.E.}(\text{Model})$	S.E (Robust)	$\hat{\beta}/\text{S.E.}(\text{Robust})$
Whole sample (0,1)	intercept	-2.9956	.1990	-15.05	.1599	-18.73
	dose	.0980	.0079	12.41	.0067	14.52
	$w^*$	-.1390	.0137	-10.12	.0136	-10.20
	$\rho$	.2937				
(1,2)	intercept	-4.6104	.5023	-9.18	.4648	-9.92
	dose	.1513	.0174	8.70	.0167	9.05
	$w^*$	-.1473	.0176	-8.37	.0190	-7.74
	$\rho$	.3094				
(2,2)	intercept	-1.8022	.1499	-12.03	.1550	-11.63
	$w^*$	-.1780	.0242	-7.36	.0270	-6.58
	$\rho$	.0616				
(3,1)	intercept	-1.9434	.1680	-11.57	.1738	-11.18
	$w^*$	-.0345	.0385	-.90	.0342	-1.01
	$\rho$	.0137				
(3,2)	intercept	-2.5372	.2348	-10.80	.2119	-11.97
	$w^*$	.0583	.0759	.77	.0700	.83
	$\rho$	.0495				

Table 4: Predicted malformation rates at each dose level for the whole sample and the samples at the terminal nodes of the trees.

Dose	Observed	Model 1		Model 2	
		Whole	Tree	Whole	Tree
0	.099	.065	.104	.075	.099
15	.191	.204	.173	.218	.203
20	.183	.284	.203	.296	.221
25	.290	.380	.386	.388	.341
30	.580	.486	.510	.488	.486
45	.752	.777	.824	.766	.852
60	1.000	.928	.955	.918	.972

Table 5: Simulation results from the proposed method for data from the null model. Entries are frequencies (%) of splits. Nominal significance level is  $\alpha = .05$ ; 10-fold cross-validation; 200 trials.

Procedure	Bootstrap method	#splits	Model			
			$\rho = 0$	$\rho = .1$	$\rho = .3$	$\rho = .5$
$\rho = 0$	B1 ( $f = \eta$ )	0	95.5	95	90	88.5
		$\geq 1$	4.5	5	10	11.5
	B2 ( $f = 0$ )	0	94.5	92.5	95.5	97
		$\geq 1$	5.5	7.5	4.5	3
$\rho$ unspecified	B1	0	95.5	96	97	94
		$\geq 1$	4.5	4	3	6
	B2	0	97	93.5	94.5	93.5
		$\geq 1$	3	6.5	5.5	6.5

Table 6: Simulation results for the Chaudhuri et al. (1995) method for data from the null model. Entries are frequencies (%) of splits. Two hundred trials.

Method	#splits	Model	
		$\rho = 0$	$\rho = .3$
Direct stopping rule ( $\alpha = .1$ )	0	93.5	35.5
	$\geq 1$	6.5	64.5
Pruning by cross-validation	0	52.5	19
	$\geq 1$	47.5	81
Pruning by Efron optimism	0	93	34.5
	$\geq 1$	7	65.5

Table 7: Simulation results from the proposed method for data from the first alternative model (Model A1). Entries are frequencies (%) of splits. Nominal significance level is  $\alpha = .05$ ; 10-fold cross-validation; 200 trials.

Procedure	Bootstrap method	#splits	Model			
			$\rho = 0$	$\rho = .1$	$\rho = .3$	$\rho = .5$
$\rho = 0$	B1 ( $f = \eta$ )	0	8	37.5	67	85.5
		$\geq 1$	92	62.5	33	14.5
	B2 ( $f = 0$ )	0	32.5	67	95	97.5
		$\geq 1$	67.5	33	5	2.5
$\rho$ unspecified	B1	0	8.5	39.5	54.5	72
		$\geq 1$	91.5	60.5	45.5	28
	B2	0	17.5	50.5	66.5	79
		$\geq 1$	82.5	49.5	33.5	21

Table 8: Simulation results for the Chaudhuri et al. (1995) method for data from the alternative model with  $\rho = 0$ ; 200 trials.

Method	#splits	Frequency (%)
Direct stopping rule ( $\alpha = .1$ )	0	1.5
	$\geq 1$	98.5
Pruning by cross-validation	0	1.5
	$\geq 1$	98.5
Pruning by Efron optimism	0	1
	$\geq 1$	99

Table 9: Simulation results from the proposed method for data from the second alternative model (Model A2). Entries are frequencies (%) of splits. Nominal significance level is  $\alpha = .05$ ; 10-fold cross-validation; using the B1 method; 200 trials.

Procedure	#splits	Model	
		$\rho = .3$	$\rho = .5$
$\rho = 0$	0	43	75
	$\geq 1$	57	25
$\rho$ unspecified	0	18.5	58.5
	$\geq 1$	81.5	41.5