

Log-normal Regression Modeling through Recursive Partitioning

Hongshik Ahn

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

Abstract

This article discusses a method for fitting log-normal regression models to censored survival data through binary decision trees. Recursive partitioning is performed by analysis of the distributions of residuals and cross-validation estimates of the average squared error. Several forms of strata selection and bootstrapping are examined to study their relative effectiveness. If the Newton-Raphson method for determining the maximum likelihood estimates fails because of heavy censoring, a method relying only on the first derivatives of the log likelihood function is used. The proposed method helps to identify the local effect of the covariates. The methods are illustrated with real and simulated data. Especially, a data set having categorical variables and missing values is used for modeling the tree-structured log-normal regression.

KEY WORDS: Bootstrap; Censoring; Cross-validation; Parametric regression; Regression tree; Survival analysis.

1 Introduction

A variety of methods have been developed for handling regression problems. 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. Recently, Breiman *et al.* (1984) developed the Classification and Regression Trees (CART) method of selecting a tree of appropriate size using cross-validation. Tree-structured regression became possible due to rapid advances in computer technology. Tree-structured methods using recursive partitioning provide a powerful analysis tool for exploring the structure of a set of data and for predicting the outcomes of new cases. With tree-structured regression techniques, some of the restrictive classical assumptions about the relationship between the response variable and the independent variables can be avoided. Moreover, a tree-structure provides easier interpretation than a regression equation since the tree identifies effects of covariates in subgroups whereas regression examines effects in the whole sample.

During the past several years, many researchers including Segal (1988), Ciampi and Thiffault (1989), Davis and Anderson (1989), Bloch and Segal (1989), Loh (1991), Ciampi (1991), LeBlanc and Crowley (1992), Schmoor *et al.* (1993), Ahn (1994) and Ahn and Loh (1994) extended the tree-structured regression method to censored survival data. There are two broad categories of regression models for censored survival data. One approach is using parametric families of lifetime distributions and the other is not assuming particular parametric families of survival distributions. The former category of the methods includes exponential, Weibull, log-normal and log-gamma regression models. The latter category was introduced by Cox (1972), Miller (1976), Buckley and James (1979) and Koul, Susarla and Van Ryzin (1981).

One way to examine the relationship of covariates to survival time is through a regression model in which survival time has a distribution that depends on the covariates. Parametric regression models would be appropriate for this situation. Among the parametric models, the log-normal distribution has been widely used as a lifetime distribution model. In the log-normal regression model, the lifetime T is log-normal and $Y = \ln T$ is normally distributed with mean $\mathbf{x}\boldsymbol{\beta}$ and variance σ^2 . Log-normal regression models with survival data are discussed in Boag (1949), Feinleib (1960), Glasser (1965) and Nelson and Hahn (1972, 1973). The methods based on uncensored data are well

known in regression analysis. However, the model is often difficult to interpret, especially when there are many correlated covariates. To avoid this disadvantage, the proposed method stratifies the data according to particular covariate values and fits separate log-normal regression models to each stratum. This approach can yield useful information about the data structure.

In this paper, we investigate the tree-structured log-normal regression modeling for censored survival data. The proposed method selects its splits by analysis of the distributions of the residuals (see Loh, 1991). This method has the same objective as the work done by Ciampi (1991). The latter proposed regression trees with the generalized regression model to build local regressions, but it is not based on the residuals. The proposed method performs regression and uses a tree construction on the residuals of this regression to improve the fit of the regression, or equivalently, to find strata which are as homogeneous as possible with respect to the estimates of the regression coefficients. This method uses a cross-validators multi-step look-ahead stopping rule to determine tree size as introduced by Chaudhuri *et al* (1994).

The program is written in FORTRAN 77 and the basic algorithm is described in Ahn (1992). The program uses a dynamic memory allocation. Therefore, it can handle any number of observations and up to 50 covariates. The maximum depth of the trees (both the main tree and the trees tested in the cross-validation) is unlimited.

The log-normal regression model and some notation are introduced in Section 2. The tree-structured algorithm is explained in Section 3. Some examples with real and simulated data are given in Section 4. In that section, one example includes categorical variables and missing values. Discussion is given in Section 5.

2 Log-normal Regression Model

Let T_1, \dots, T_n and C_1, \dots, C_n be independent random variables, where C_i is the censoring time associated with the survival time T_i , $i = 1, \dots, n$. We observe $(W_1, \delta_1), \dots, (W_n, \delta_n)$, where $W_i = \min\{T_i, C_i\}$, $\delta_i = I(T_i \leq C_i)$ and $I(\cdot)$ is the indicator function. Assume that for each i , a p -dimensional covariate vector $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})$ independent of T_i is available.

2.1 Model

The probability density function of Y given \mathbf{x} is $f(y|\mathbf{x}) = (\sqrt{2\pi}\sigma)^{-1} \exp\{-(y - \mathbf{x}\boldsymbol{\beta})^2/(2\sigma^2)\}$ and the survival function of Y given \mathbf{x} is $S(y|\mathbf{x}) = 1 - \Phi\{(y - \mathbf{x}\boldsymbol{\beta})/\sigma\}$, where $\Phi(\cdot)$ is the standard normal cumulative distribution function. Therefore,

$$Z = (Y - \mathbf{x}\boldsymbol{\beta})/\sigma \tag{1}$$

has the standard normal distribution.

2.2 Maximum Likelihood Methods

The Newton-Raphson method is used to find the maximum likelihood estimates of the parameters. However, if censoring is heavy, this method fails to converge unless initial values are very close to the maximum likelihood estimates. In this case, another iterative procedure is provided using a method that requires only the first derivatives of the log likelihood function.

2.2.1 The Newton-Raphson method

Since we work with log times, $y_i = \ln w_i$ represents a log lifetime or log censoring time. From the probability density function and survival function of Y , the log likelihood function for a censored

sample based on n observations is

$$\ln L(\boldsymbol{\beta}, \sigma) = \sum_{i=1}^n [-\delta_i \{\ln(\sqrt{2\pi}\sigma) + (y_i - \mathbf{x}_i\boldsymbol{\beta})^2/(2\sigma^2)\} + (1 - \delta_i) \ln\{1 - \Phi((y_i - \mathbf{x}_i\boldsymbol{\beta})/\sigma)\}],$$

where $\phi(\cdot)$ is the standard normal probability density function. Let $z_i = (y_i - \mathbf{x}_i\boldsymbol{\beta})/\sigma$ and $A(z) = \phi(z)/(1 - \Phi(z))$. The first and second derivatives of $\ln L$ can be obtained from the above formula. The maximum likelihood equations $\partial \ln L/\partial \beta_r = 0$, $r = 1, \dots, p$, and $\partial \ln L/\partial \sigma = 0$ are solved by the Newton-Raphson method to get the maximum likelihood estimates of $\boldsymbol{\beta}$ and σ^2 . Negative values of σ can arise during the iteration. This is avoided by replacing each negative value with one-half the σ value in the previous iteration (see page 315, Lawless (1981)).

2.2.2 An alternative iteration method

In the case that the Newton-Raphson method fails to converge, a generalization of the method of Sampford and Taylor (1959) is tried. The procedure discussed in Lawless (1981, pages 316-317) is implemented in our program. The maximum likelihood equations are

$$\partial \ln L/\partial \beta_r = \sigma^{-1} \sum_{i=1}^n \{(v_i - \mathbf{x}_i\boldsymbol{\beta})/\sigma\} x_{ir} = 0, \quad r = 1, \dots, n \quad (2)$$

and

$$\partial \ln L/\partial \sigma = \sum_{i=1}^n [-\delta_i n + \{(v_i - \mathbf{x}_i\boldsymbol{\beta})/\sigma\}^2 - (1 - \delta_i)A(z_i)(A(z_i) - z_i)] = 0, \quad (3)$$

where $v_i = \delta_i y_i + (1 - \delta_i)\{\mathbf{x}_i\boldsymbol{\beta} + \sigma A(z_i)\}$. The solution to (2) is

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}, \quad (4)$$

where $V = (v_1, \dots, v_n)'$ and (3) gives

$$\tilde{\sigma}^2 = \sum_{i=1}^n (v_i - \mathbf{x}_i \tilde{\boldsymbol{\beta}})^2 / \sum_{i=1}^n [\delta_i n + (1 - \delta_i) A(z_i) \{A(z_i)(A(z_i) - z_i)\}]. \quad (5)$$

Equations (4) and (5) are used to calculate new estimates $\tilde{\boldsymbol{\beta}}$ and $\tilde{\sigma}^2$ of $\boldsymbol{\beta}$ and σ^2 . The procedure is repeated. This procedure converges to the maximum likelihood estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ if suitable initial values are used. This procedure is essentially an EM algorithm. Wolynetz (1974) reports that this scheme converges more surely than Newton-Raphson iteration, though more slowly. (See also Wolynetz, 1979). A maximum of 30 iterations are allowed in the program.

3 Tree-structured Algorithm

Let $X = (\mathbf{x}'_1, \dots, \mathbf{x}'_n)'$ be the $n \times p$ matrix of covariates, where $\mathbf{x}^k = (x_{1k}, \dots, x_{nk})'$ is the n -dimensional vector for the k th covariate and $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})$ is the p -dimensional vector of covariates for the i th case. Any covariate that takes categorical values is transformed into dummy vectors of indicator variables. Although the proposed method is considered a regression tree, it also plays a role of a classification tree when a dummy variable is chosen to be split. See Section 4.2 for further explanations on categorical variables. The missing values are replaced with the mean value (for continuous) or modal value (for categorical).

3.1 Splitting Rules

A binary tree is constructed by splitting the data in each node into two subnodes. Each split is formed by a question of the form: Is $x_{ik} \leq c_k$? To choose k , we study the distributions of the residuals along each x^k -axis and select the one for which the residuals appear most non-random. At each node, we assign the cases that satisfy the inequality to the left subnode and to the right subnode otherwise. We employ the following two approaches (the M and R methods) proposed in

Loh (1991) for classifying the data into two cases. At each node,

1. Fit the log-normal regression model using the sample at the node and get the residuals z_i , where $z_i = (y_i - \mathbf{x}_i \hat{\boldsymbol{\beta}}) / \hat{\sigma}$, where y_i represents a log lifetime or log censoring time.
2. In the R method, a case belongs to class 1 if its corresponding residual is positive and belongs to class 2 otherwise. Since (1) has the standard normal distribution, the residuals behave as in least squares regression for uncensored data.

In the M method, a case belongs to class 1 if its corresponding residual is larger than the median of the residuals for the sample and belongs to class 2 otherwise. This method is used because the mean of the residuals is not near zero due to the censoring.

3. For each covariate, perform two-sample t -tests for means and Levene's tests for variances (Levene, 1960) on the two groups of observations.
4. The P -value from the larger of the t statistic and Levene's statistic is computed for each covariate.
5. Suppose the k th covariate yields the smallest P -value. The data in the node are split into two groups, with one subset containing all cases with the k th covariate value less than or equal to c and the other subset containing the remaining cases, where c is the average of the two sample means.

3.2 Stopping Rules

In order to determine whether a node should be split or not, average squared error is used as a measure of goodness-of-fit. In tree-structured log-normal regression, the values of the splitting threshold η and fractional reduction f in cross-validation determine the shape of the tree. In V -fold cross-validation, a nested sequence of trees is constructed from the data from $(V - 1)$ subsets

and the remaining subset is used as a test sample. The procedure is repeated V times, each time leaving out a different subset as test sample. At each time, if a cross-validation tree has an estimate of average squared error smaller than that for the trivial trees at least $(1 - f)$ times, then it is considered better than the trivial tree. If the number of the better cross-validation trees is greater than ηV , then the node is split. See Ahn and Loh (1994) for further details.

3.3 Bootstrap Parameter Selection

In this section, we give a detailed discussion of the bootstrap choice of the parameters introduced in Ahn (1992) and Ahn and Loh (1994). To choose f and η , test if the root node needs to be split for the data. Let the null hypothesis be ‘The root node should not be split’. Then the Type I error is

$$\alpha = P(\text{Split the root node} | H_0 : \text{The root node should not be split}). \quad (6)$$

Probability (6) is evaluated using different values of the (f, η) pair. The (f, η) are chosen to be the values for which (6) is closest to the pre-selected α . Three different bootstrap estimation methods are tried for finding f and η . The first bootstrap method is to fix the values of f and η to be equal and to search for the best common value. The second method is to fix f at 0 and search for the best η value. Finally, the third method is to fix η at .5 and search for the best f value. In each method, the significance level α of the test is chosen prior to searching. Now we discuss the first method in detail. The other methods estimate the parameters the same way. In the first bootstrap method, fix $f = \eta$. Let the estimate of (6) be

$$g_1(f) = \hat{\alpha}(f, \eta) = P_{\hat{F}}(\text{Split the root node} | H_0 : \text{The root node should not be split}). \quad (7)$$

Since the tree size gets larger as f and η become smaller, g_1 is a nonincreasing function of f . Using one hundred samples, the following procedure is performed at each value of f : Starting from a

small value of f and η ($f = \eta = .1$), increase f by .1 and evaluate (7). Stop at $f = f_0$ such that $g_1(f_0) \leq \alpha$ and $g_1(f_0 - .1) > \alpha$. Take $f = \eta = f_0$ for the R method and $f = \eta = f_0 - .05$ for the M method as the best choice of (f, η) , since the estimated probabilities of Type I errors are closer to α in that way.

The survival times and censoring distributions are generated as follows.

1. **Generation of bootstrap survival times T^* :** For the log-normal regression model, survival times are generated from the residuals. Let z_1, \dots, z_n be the residuals. To generate a bootstrap observation, first randomly choose n numbers from $\{1, \dots, n\}$ with replacement. Let the set of chosen numbers be $\{r_1, \dots, r_n\}$; then the bootstrap observation $t_i^* = z_{r_i}$ is obtained. The associated covariate vector is \mathbf{x}_{r_i} .
2. **Generation of bootstrap censoring time C^* :** Assume that the real censoring times C are independent of the real survival times and the covariates. By changing the censoring indicators $(d_{r_1}, \dots, d_{r_n})$ into $(1 - d_{r_1}, \dots, 1 - d_{r_n})$, switch the censoring status. Using (t_1^*, \dots, t_n^*) and the new censoring indicators, find the Kaplan-Meier estimate of the censoring distribution. Let $\hat{S}(t)$ be the Kaplan-Meier estimate. To generate the bootstrap censoring times C^* , generate random variables $U_i \sim \text{Uniform}(0, 1), i = 1, \dots, n$ and let $C_i^* = \max\{t : \hat{S}(t) \geq U_i\}$.
3. Let (t_1^*, \dots, t_n^*) and (c_1^*, \dots, c_n^*) be the bootstrap survival and censoring times. Compute the bootstrap observations $(y_1^*, d_1^*), \dots, (y_n^*, d_n^*)$, where $y_i^* = \min\{t_i^*, c_i^*\}$ and $d_i^* = I(t_i^* \leq c_i^*)$. These data and the observed covariate values are used to get bootstrap estimates of the probability of the Type I error of splitting the root node when it should not be split.

4 Examples

Some real and simulated data are examined to demonstrate the performance of the log-normal regression trees. In all the examples, 10-fold cross-validation is used and the significance level for

the bootstrap is chosen to be .05.

4.1 Stanford Heart Transplant Data

The Stanford heart transplant data published in Andrews and Herzberg (1985) are used in this analysis. There are 157 patients who have non-missing T5 mismatch scores. The data were previously analyzed by Miller and Halpern (1982) using the Cox (1972), Buckley-James (1979) and Miller (1976) regression methods. Fitting $\log_{10}(\text{survival time})$ on age and mismatch score, they concluded that mismatch score was not significant and that a quadratic model in age was satisfactory. For their analysis, Miller and Halpern deleted the 5 patients with survival times less than 10 days so that the data are more symmetrical. Wei, Ying and Lin (1990) re-analyzed the data using linear regression based on rank tests and also concluded that a quadratic model in age was better than a linear one. Ahn and Loh (1994) suggested that a piecewise linear fit for each of young and old age groups might be adequate.

A log-normal regression model is fitted to the entire sample. One case with zero survival time is changed to .5 since the log-survival times are used in log-normal regression models. Table 1 gives the regression estimates, Wald test statistics and the P -values for the coefficients. Survival times are shorter for older patients (P -value is .052), but mismatch score is not significant at level .1.

Next, the data are fitted using log-normal regression trees. Before we report our results of the log-normal regression trees, we define $node(i, j)$ as the j th node from the left (including the empty nodes) at the i th level. The root node is $node(0, 1)$.

4.1.1 Using the M method

Applying our method with the third (with $\eta = .5$) bootstrap estimation method gives a tree with one split, on age at 41 years. The first (with $f = \eta$) and second (with $f = 0$) bootstrap methods produced the tree in Figure 1 which has three more splits. No more split occurs at $node(1, 1)$, but

the sample in $node(1, 2)$ is split at $mismatch = 1.00$ so that 44 cases with $age > 41$ and $mismatch \leq 1.00$ are put into $node(2, 3)$ and 49 cases with $age > 41$ and $mismatch > 1.00$ are put into $node(2, 4)$. There are further splits in $node(2, 3)$ at age 48 years and in $node(2, 4)$ at age 48 years. Figure 2 shows the Kaplan-Meier estimates of the survival distributions for the data in the five terminal nodes. Regression estimates and P -values for the coefficients in the terminal nodes are given in Table 2. Age is significant only in $node(3, 8)$ and mismatch score is significant only in $node(3, 5)$ at level .05. The median survival times are seen to be shorter for the patients with larger mismatch scores. (Compare $node(3, 5)$ with $node(3, 7)$, and $node(3, 6)$ with $node(3, 8)$.)

4.1.2 Using the R method

In the R method, the first bootstrap method gives a tree with one split, on age at 41 years. The second bootstrap method produces the tree in Figure 3 which has two more splits. The second split is on mismatch score at 1.15 and the third split is on age at 48 years. The sample in $node(1, 1)$ is the same as in $node(1, 1)$ of the tree obtained from the M method. Figure 4 shows the Kaplan-Meier estimates of the survival distributions for the data for the four terminal nodes and Table 3 gives the regression estimates in $node(2, 3)$, $node(3, 7)$ and $node(3, 8)$. The estimates in $node(1, 1)$ are the same as in Table 2. Age is significant only in $node(2, 3)$ at level .05. Mismatch score is not significant in any terminal node of the tree. The median survival times are 697, 544, 292 and 129 days in the four terminal nodes. The third bootstrap method produces a trivial tree with no splits.

4.1.3 Conclusions from the analysis

From the above results, we see that the first split occurs at the same place on age in both the M and R methods. The second split (if it occurs) occurs on mismatch score in both methods and the split points are very close. At level 2, the two splits in the M method and the one split in the R method occur at the same place (48 years of age). From this, we conclude that the M and

R methods work quite similarly for the data. The six trees from the above analysis range from the trivial one with no split to one with four splits. As Ahn and Loh (1994) suggest, they can be treated as an informal confidence set. We recommend the tree with one split (the third bootstrap with M and the first bootstrap with R methods), since it is about halfway between the smallest and the largest trees. Also, it is obtained in both the M and R methods.

For this example, the log-normal regression fit for the whole sample shows that age is a significant factor for all the age groups. However, the proposed method shows that age affects the survival time significantly for only the old patients. This result coincides with the conclusion from the analysis of the data in Ahn and Loh (1994).

4.2 Squamous Carcinoma Data

Kalbfleisch and Prentice (1980, Appendix 1) gives the data for a part of a clinical trial carried out by the Radiation Therapy Oncology Group in the United States. The data of the patients with squamous carcinoma on three sites in the oropharynx were reported by six institutes. We included six variables for analysis: (i) sex (1 = male; 2 = female), (ii) grade (1 = well differentiated; 2 = moderately differentiated; 3 = poorly differentiated), (iii) age in years at time of diagnosis, (iv) condition (1 = no disability; 2 = restricted work; 3 = requires assistance with self care; 4 = bed confined), (v) T staging (1 = primary tumor measuring 2 cm or less in largest diameter; 2 = primary tumor measuring 2 cm to 4 cm in largest diameter, minimal infiltration in depth; 3 = primary tumor measuring more than 4 cm; 4 = massive invasive tumor), and (vi) N staging (0 = no clinical evidence of node metastases; 1 = single positive node, 3 cm or less in diameter, not fixed; 2 = single positive node, more than 3 cm in diameter, not fixed; 3 = multiple positive nodes or fixed positive nodes). For the fourth variable (condition), levels 2, 3 and 4 were combined and considered as disability status, since there is not much difference among 2, 3 and 4 according to the log-normal regression fit. The data also include the survival time and censoring indicator (1 =

death; 0 = censored) for each patient. There are two missing values in condition and one missing value in grade, and those are included in our analysis. We transformed the categorical values into dummy vectors of 0-1 indicator variables for the purpose of fitting log-normal regression models. They are, Sex (1 = male, 0 = female), G1 (1 if grade = 1, 0 otherwise), G2 (1 if grade = 1 or 2, 0 otherwise), Cond (1 if condition = 1, 0 otherwise), T1 (1 if T staging = 1, 0 otherwise), T2 (1 if T staging = 1 or 2, 0 otherwise), T3 (1 if T staging = 1, 2 or 3, 0 otherwise), N0 (1 if N staging = 0, 0 otherwise), N1 (1 if N staging = 0 or 1, 0 otherwise), and N2 (1 if N staging = 0, 1 or 2, 0 otherwise). In the case that a dummy variable is chosen to be split, the model is fitted at the low level of the variable at the left subnode, and the model is fitted at the high level at the right subnode. Hence, for categorical variables, the proposed method has the role of a classification tree. Fitting a log-normal regression model to the whole sample, Cond is the most significant factor, and T3 is found to be significant at level .05 (see Table 4). The survival time tends to be longer for the patients without disability.

The first bootstrap with M method gives a tree with one split. The sample is split at Cond = .7. That is, the cases with Cond = 0 are put into the left subnode and the cases with Cond = 1 are put into the right subnode. Since the variable Cond has constant value at both the terminal nodes, it is not included in the model. In *node(1, 1)*, T1 is also excluded in the model since it has a constant value. Figure 5 shows the Kaplan-Meier estimates of the survival times for the terminal nodes and Table 5 gives the regression estimates and *P*-values. For the patients with disability (*node(1, 1)*), no covariate is significant at level .05. For the patients without disability, however, T3 and N2 are significant. The median survival times are 228.5 days for the patients with disability and 544 days for the patients without disability.

The log-normal regression fit for the whole sample shows that primary tumor measuring affects the survival time significantly. However, the proposed method shows that the data for the patients with and without disability may need to be analyzed using separate log-normal regression models.

For the patients with disability, none of the variables significantly affect the survival time. In contrast, for the patients without disability, both primary tumor measuring and node metastases are significant factors.

The above observation is impossible to deduce from Table 4. The sample is split in the most significant variable and two different log-normal regression models are fitted in the two subnodes. The fitted models became simpler in the tree obtained in this section. Since no covariate is significant in one of the two strata, the Kaplan-Meier curves for this stratum is a sufficient summary of the information in the data. The other bootstrap methods give trivial trees.

4.3 Simulations

4.3.1 One log-normal regression model

In the first simulation experiment, survival times were generated from a log-normal distribution with mean $\mu_i = \mathbf{x}_i\boldsymbol{\beta}$ and variance $\sigma = 1$, where $\boldsymbol{\beta} = (1, 1)'$, $\mathbf{x}_i = (1, x_{i1})$ and $x_{i1} \in \{-4, -3, -2, -1, 1, 2, 3, 4\}$. Each design point was replicated eight times, giving a total of 64 cases per trial. Censoring times were independently generated from an exponential distribution with mean 100 so that the proportion of censoring is about 20%. Two hundred simulation trials were performed for each of the R and M methods and each of the three bootstrap methods for choosing f and η .

Simulation results are given in Table 6. The number of splits and the number of times they were observed are shown in the table. Since the data were generated from a single log-normal regression model, we expect no split. In this simulation experiment, we expect to have the probability of a Type I error around .05. The first and third bootstrap methods seem to be quite conservative in both the M and R methods. The second bootstrap method gave eight and twelve non-trivial trees in the M and R methods, respectively. The estimated probabilities of a Type I error in the M and R methods are thus $.040 \pm .014$ (the “ \pm ” quantities are simulation standard errors.) and $.060 \pm .017$,

respectively for the second bootstrap method.

4.3.2 Two log-normal regression models

In the second experiment, data were generated from two log-normal regression models. We expect exactly one split so that the tree has two terminal nodes. The purpose of this experiment is to compare the power of the individual methods in detecting the need to partition the data. Table 7 gives the simulation results. The powers are larger for the first and third bootstrap estimation methods in both the M and R methods. In both methods, most of the non-trivial trees obtained in this simulation experiment have exactly one split.

4.3.3 Conclusions from the simulations

From the above simulations, we conclude that the R method gives more power than the M method. The second bootstrap method in both the M and R methods control the probability of a Type I error quite satisfactorily. In particular, the second bootstrap with R method gives more power.

5 Discussion

This paper presents a tree-structured log-normal regression algorithm for the analysis of censored survival data. Unlike the usual regression tree for censored survival data, the proposed method selects its splits by analysis of the distributions of the residuals. Instead of using the cross-validation estimate in CART's pruning method to find the final tree, the proposed method uses a multi-step look-ahead stopping rule and bootstrap resampling to determine the size of a tree.

From the results in this paper, we conclude that the manner in which the data are stratified can yield useful information about the data structure. Also, we found that tree structure helps detection of conditional information of the data. Since the data within a stratum would be more homogeneous, they are fitted with fewer covariates. This makes interpretation easier. Also, interpretation is

simpler because only models with linear terms are employed. Furthermore, categorical covariates and missing values are allowed in the log-normal regression trees. As we discussed in Section 4.2, the proposed method functions as a regression tree for continuous covariates and a classification tree for categorical variables. Bootstrap estimation of the probability that a single log-normal regression model is erroneously rejected as inadequate provides the formal test of fit.

Several simulation examples were considered to study the effectiveness of the tree-structured method. The probability of a Type I error is quite satisfactory in the second ($f = 0$) bootstrap estimation method. The first ($f = \eta$) and third ($\eta = .5$) bootstrap methods seem to yield the best power in both the M and R methods. As a general method, we recommend the second bootstrap with R method since it controls the probability of a Type I error quite well and yields large power.

There are many situations that the proposed method can be applied. For example, the method can be applied in medical studies dealing with fatal disease. Also, the procedure can be used in the investigation of the effect of carcinogenic substances on tumor onset or survival time in laboratory animals (Gart *et al.*, 1986).

Acknowledgements

This research was partially supported by NSF grant DMS88-03271 and ARO grant DAAL03-91-G-0111 while the author was at the University of Wisconsin-Madison. The author thanks professor Wei-Yin Loh for his discussion of the paper and most valuable suggestions. The author also wishes to thank the referees for comments that substantially improved the paper.

References

Ahn, H. (1992). Survival modeling through regression trees. Unpublished Ph.D. Thesis.

Department of Statistics, University of Wisconsin-Madison.

- Ahn, H. (1994). Tree-structured extreme value model regression. *Communications in Statistics - Theory and Methods*, **23**, 153-174.
- Ahn, H. and Loh, W.-Y. (1994). Tree-structured proportional hazards regression modeling. *Biometrics*, **50**, 471-485.
- Andrews, D.F. and Herzberg, A.M. (1985). *Data*. Springer-Verlag, New York, 45-50.
- Bloch, D.A. and Segal, M.R. (1989). Empirical comparison of approaches to forming strata. *JASA*, **84**, 897-905.
- Boag, J.W. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *J. R. Statist. Soc. B*, **11**, 15-53.
- Breiman, L., Friedman, J.H., Olshen R.A. and Stone, C.J. (1984). *Classification and regression trees*. Wadsworth, Belmont, California.
- Buckley, J. and James, I. (1979). Linear regression with censored data. *Biometrika*, **66**, 429-436.
- Chaudhuri, P., Huang, M.-C., Loh, W.-Y. and Yao, R. (1994). Piecewise-polynomial regression trees. *Statistica Sinica*, **4**, 143-167.
- Ciampi, A. (1991). Generalized regression trees. *Computational Statistics and Data Analysis*, **12**, 57-78.
- 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.
- Cox, D.R. (1972). Regression models and life-tables. *J. R. Statist. Soc. B*, **34**, 187-202.

- Davis, R.B. and Anderson, J.R. (1989). Exponential survival trees. *Statistics in Medicine*, **8**, 947-961.
- Feinleib, M. (1960). A method of analyzing lognormally distributed survival data with incomplete follow-up. *JASA*, **55**, 534-545.
- Gart, J.J., Krewski, D., Lee, P.N., Tarone, R.E. and Wahrendorf, J. (1986). *Statistical Methods in Cancer Research, (Vol. 3) The Design and Analysis of Long-Term Animal Experiments*. Oxford University Press.
- Glasser, M. (1965). Regression analysis with dependent variable censored. *Biometrics*, **21**, 300-307.
- Kalbfleisch, J.D. and Prentice, R.L. (1980). *The statistical analysis of failure time data*. Wiley, New York, 225-229.
- Koul, H., Susarla, V. and Van Ryzin, J. (1981). Regression analysis with randomly right-censored data. *The Annals of Statistics*, **9**, 1276-1288.
- Lawless, J.F. (1981). *Statistical models and methods for lifetime data*. Wiley, New York.
- 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*. Olkin, I., *et al* (eds.) Stanford University Press, 278-292.
- Loh, W.-Y. (1991). Survival modeling through recursive stratification. *Computational Statistics and Data Analysis*, **12**, 295-313.
- Miller, R.G. (1976). Least squares regression with censored data. *Biometrika*, **63**, 449-464.

- Miller, R.G. and Halpern, J. (1982). Regression with censored data. *Biometrika*, **69**, 521-531.
- Morgan, J.N. and Sonquist, J.A. (1963). Problems in the analysis of survey data, and a proposal. *JASA*, **58**, 415-434.
- Nelson, W.B. and Hahn, G.J. (1972). Linear estimation of a regression relationship from censored data. Part I - Simple methods and their applications. *Technometrics*, **14**, 247-269.
- Nelson, W.B. and Hahn, G.J. (1973). Linear estimation of a regression relationship from censored data. Part II - Best linear unbiased estimation and theory. *Technometrics*, **15**, 133-150.
- Sampford, M.R. and Taylor, J. (1959). Censored observations in randomized block experiments. *J. R. Statist. Soc. B*, **21**, 214-237.
- Segal, M.R. (1988). Regression trees for censored data. *Biometrics*, **44**, 35-47.
- Schmoor, C., Ulm, K. and Schumacher, M. (1993). Comparison of the Cox model and the regression tree procedure in analysing a randomized clinical trial. *Statistics in Medicine*, **12**, 2351-2366.
- Wei, L.J., Ying Z. and Lin, D.Y. (1990). Linear regression analysis of censored survival data based on rank tests. *Biometrika*, **77**, 845-851.
- Wolynetz, M.S. (1974). Analysis of Type I censored normally distributed data, Unpublished Ph.D. Thesis. University of Waterloo, Waterloo, Ontario, Canada.
- Wolynetz, M.S., (1979). Statistical algorithms AS 138 and AS 139. *Applied Statistics*, **28**, 185-206.

Table 1: Regression estimates of the coefficients and P -values on the covariates for the Stanford heart transplant data with the log-normal regression model. One case with zero survival time was deleted.

Variable	Estimate	S.E.	$\hat{\beta}/\text{S.E.}$	P -value (Wald test)
intercept	8.000	0.921	8.68	< 0.0001
age	-0.039	0.020	-1.94	0.0520
mismatch	-0.079	0.361	-0.22	0.8279
scale	2.453	0.183		

Table 2: Regression estimates and P -values of Wald tests at the terminal nodes of the tree in Figure 1.

Node	Parameter	Estimate	S.E.	$\hat{\beta}/\text{S.E.}$	P -value (Wald test)
age \leq 41 <i>node</i> (1, 1)	intercept	4.215	1.998	2.11	0.0349
	age	0.053	0.055	0.96	0.3360
	mismatch	1.283	0.857	1.50	0.1347
	scale	3.203	0.434		
41 < age \leq 48 & mismatch \leq 1.00 <i>node</i> (3, 5)	intercept	-5.007	8.864	-0.56	0.5722
	age	0.203	0.190	1.06	0.2869
	mismatch	4.317	1.591	2.71	0.0067
	scale	1.715	0.401		
age > 48 & mismatch \leq 1.00 <i>node</i> (3, 6)	intercept	14.405	6.024	2.39	0.0168
	age	-0.173	0.115	-1.51	0.1305
	mismatch	0.356	2.412	0.15	0.8826
	scale	2.152	0.373		
41 < age \leq 48 & mismatch > 1.00 <i>node</i> (3, 7)	intercept	2.149	8.094	0.27	0.7906
	age	0.123	0.171	0.72	0.4700
	mismatch	-1.189	1.308	-0.91	0.3635
	scale	1.665	0.300		
age > 48 & mismatch > 1.00 <i>node</i> (3, 8)	intercept	26.611	6.878	3.87	0.0001
	age	-0.412	0.138	-2.97	0.0029
	mismatch	-0.069	0.527	-0.13	0.8953
	scale	1.380	0.205		

Table 3: Regression estimates and P -values of Wald test for the two terminal nodes of the tree in Figure 3. The estimates of age ≤ 41 is in Table 3.

Node	Parameter	Estimate	S.E.	$\hat{\beta}/\text{S.E.}$	P -value (Wald test)
age > 41 & mismatch ≤ 1.15 <i>node(2, 3)</i>	intercept	14.236	3.198	4.45	0.0001
	age	-0.180	0.065	-2.79	0.0093
	mismatch	0.960	1.030	0.93	0.0643
	scale	2.012	0.244		
41 < age ≤ 48 & mismatch > 1.15 <i>node(3, 7)</i>	intercept	0.039	9.086	0.004	0.9966
	age	0.205	0.192	1.06	0.2869
	mismatch	-2.117	1.668	-1.27	0.2044
	scale	1.738	0.346		
age > 48 & mismatch > 1.15 <i>node(3, 8)</i>	intercept	21.256	7.897	2.69	0.0071
	age	-0.290	0.158	-1.84	0.0655
	mismatch	-0.538	0.647	-0.83	0.4051
	scale	1.378	0.236		

Table 4: Regression estimates of the coefficients and P -values on the covariates for the Squamous carcinoma data with the log-normal regression model.

Variable	Estimate	S.E.	$\hat{\beta}/\text{S.E.}$	P -value (Wald test)
intercept	5.244	0.467	11.24	< 0.0001
Sex	-0.169	0.168	-1.01	0.3142
G1	-0.010	0.170	-0.06	0.9520
G2	-0.288	0.196	-1.47	0.1411
age	-0.000	0.006	-0.01	0.9914
Cond	1.065	0.165	6.45	< 0.0001
T1	-0.421	0.382	-1.10	0.2701
T2	0.186	0.239	0.78	0.4364
T3	0.481	0.158	3.04	0.0024
N0	-0.090	0.240	-0.38	0.7071
N1	0.098	0.271	0.36	0.7165
N2	0.329	0.207	1.59	0.1118
scale	0.917	0.057		

Table 5: Regression estimates and P -values of Wald tests at the terminal nodes of the tree for the Squamous Carcinoma data. The first bootstrap with M method is used.

Node	Parameter	Estimate	S.E.	$\hat{\beta}/\text{S.E.}$	P -value (Wald test)
Patients with disability <i>node(1, 1)</i>	intercept	5.409	1.164	4.65	< 0.0001
	Sex	-0.077	0.356	-0.22	0.8292
	G1	-0.090	0.415	-0.22	0.8281
	G2	-0.263	0.499	-0.53	0.5981
	Age	-0.002	0.017	-0.14	0.8905
	T2	-0.236	0.650	-0.36	0.7164
	T3	0.585	0.355	1.65	0.0993
	N0	-0.893	0.673	-1.33	0.1843
	N1	0.996	0.707	1.41	0.1588
	N2	-0.145	0.511	-0.28	0.7767
	scale	1.044	0.114		
Patients without disability <i>node(1, 2)</i>	intercept	6.165	0.456	13.52	< 0.0001
	Sex	-0.138	0.191	-0.72	0.4710
	G1	-0.030	0.176	-0.17	0.8657
	G2	-0.340	0.203	-1.68	0.0937
	Age	0.002	0.007	0.26	0.7941
	T1	-0.490	0.354	-1.38	0.1663
	T2	0.304	0.242	1.26	0.2090
	T3	0.409	0.171	2.40	0.0166
	N0	0.093	0.241	0.38	0.7004
	N1	-0.059	0.273	-0.21	0.8299
N2	0.474	0.213	2.23	0.0258	
	scale	0.822	0.063		

Table 6: Simulation results for one log-normal regression model using the bootstrap to choose f and/or η . Nominal significance level is $\alpha = 0.05$; 20% censoring; 200 simulations.

Bootstrap method	M method		R method	
	#splits	freq.	#splits	freq.
1st ($f = \eta$)	0	198	0	199
	1	2	1	0
			2	1
2nd ($f = 0$)	0	192	0	188
	1	7	1	11
	2	1	2	1
3rd ($\eta = .5$)	0	199	0	195
	1	1	1	5

Table 7: Simulation results for two log-normal regression models, using the bootstrap to find f and/or η . Significance level is $\alpha = 0.05$, 20% censoring, 200 trials.

Bootstrap method	M method		R method	
	#splits	freq.	#splits	freq.
1st ($f = \eta$)	0	15	0	9
	1	179	1	181
	2	6	2	10
2nd ($f = 0$)	0	77	0	47
	1	121	1	149
	2	2	2	4
3rd ($\eta = .5$)	0	33	0	13
	1	165	1	179
	2	2	2	8