

Tree-structured Exponential Regression Modeling*

Hongshik Ahn

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

Abstract

A method for fitting piecewise exponential regression models to censored survival data is described. Stratification is performed recursively, using a combination of statistical tests and residual analysis. The splitting criterion employed in cross-validation is the average squared error of the residuals. The bootstrap is employed to keep the probability of a type I error (the error of discovering two or more strata when there is only one) of the method close to a pre-determined value. The proposed method can thus also serve as a formal goodness-of-fit test for the exponential regression model. Real and simulated data are used for illustration.

KEY WORDS: Bootstrap; Exponential regression; Recursive partitioning; Survival analysis.

1 Introduction

Many regression techniques are available for use with censored survival data. We can divide these into two broad categories. One employs parametric families of lifetime distributions and extends models such as exponential, Weibull, log-normal and log-gamma models to include covariates. The second approach is distribution-free and assumes less about underlying distributions than do the

*Research partially supported by NSF grant DMS88-03271 and ARO grant DAAL03-91-G-0111. The author would like to express sincere thanks to Professor Wei-Yin Loh for his discussion of this paper and most valuable suggestions.

parametric models. The former category of the methods is to examine the relationship of covariates to survival time through a distribution model in which the latter has a distribution that depends on the covariates.

Among the parametric models, the exponential distribution is important in application. It was the first lifetime model for which statistical methods were extensively developed in the life testing literature. However, as in the normal regression, the model is often difficult to interpret, especially when there are numerous covariates, some of which being correlated. One way to avoid this disadvantage is to stratify the data according to particular covariate values and fit separate exponential regression models to each stratum. In this paper, we explore tree-structured exponential regression for survival data.

2 Exponential regression model

An exponential regression model is appropriate when each individual has a constant hazard function. Glasser (1967), Cox and Snell (1968), Prentice (1973), Lawless (1976), Feigl and Zelen (1965) have studied this 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 T given \mathbf{x} is

$$f(t|\mathbf{x}) = \theta \mathbf{x}^{-1} \exp(-t/\theta \mathbf{x}), t > 0, \quad (1)$$

where $\theta_{\mathbf{x}} = E(T|\mathbf{x})$. The most useful functional form of $\theta_{\mathbf{x}}$ is

$$\theta_{\mathbf{x}} = \exp(\mathbf{x}\boldsymbol{\beta}), \quad (2)$$

where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$ is a vector of the regression coefficients. Hence, we can rewrite (1) as

$$f(t|\mathbf{x}) = e^{-\mathbf{x}\boldsymbol{\beta}} \exp(-te^{-\mathbf{x}\boldsymbol{\beta}}), \quad t > 0.$$

If we let $Y = \ln T$, the probability density function of Y given \mathbf{x} is

$$f(y|\mathbf{x}) = \exp\{(y - \mathbf{x}\boldsymbol{\beta}) - \exp(y - \mathbf{x}\boldsymbol{\beta})\}, \quad -\infty < y < \infty \quad (3)$$

from (1) and (2) and the survival function of Y given \mathbf{x} is

$$S(y|\mathbf{x}) = \exp\{-\exp(y - \mathbf{x}\boldsymbol{\beta})\}. \quad (4)$$

Also, we can write

$$Y = \mathbf{x}\boldsymbol{\beta} + Z,$$

where Z has an extreme value distribution with probability density function

$$f(z|\mathbf{x}) = \exp(z - e^z), \quad -\infty < z < \infty.$$

This is a location-scale regression model with error Z . Maximum likelihood estimation is used to estimate $\boldsymbol{\beta}$.

2.2 Maximum likelihood methods

Since we work with log times, $y_i = \ln t_i$ represents a log lifetime or log censoring time. From the probability density function (3) and survival function (4) of Y , the likelihood function for a censored sample based on n observations is

$$L(\boldsymbol{\beta}) = \prod_{i=1}^n [\exp\{(y_i - \mathbf{x}_i\boldsymbol{\beta}) - \exp(y_i - \mathbf{x}_i\boldsymbol{\beta})\}]^{\delta_i} [\exp\{-\exp(y_i - \mathbf{x}_i\boldsymbol{\beta})\}]^{1-\delta_i},$$

where

$$\delta_i = \begin{cases} 0 & \text{if the } i\text{th individual is censored} \\ 1 & \text{otherwise} \end{cases}$$

and thus

$$\ln L(\boldsymbol{\beta}) = \sum_{i=1}^n \{\delta_i(y_i - \mathbf{x}_i\boldsymbol{\beta}) - \exp(y_i - \mathbf{x}_i\boldsymbol{\beta})\}.$$

The first and second derivatives of $\ln L$ are

$$\partial \ln L / \partial \beta_r = - \sum_{i=1}^n x_{ir} \{\exp(y_i - \mathbf{x}_i\boldsymbol{\beta}) - \delta_i\}, \quad r = 1, \dots, p,$$

$$\partial^2 \ln L / (\partial \beta_r \partial \beta_s) = - \sum_{i=1}^n x_{ir} x_{is} \exp(y_i - \mathbf{x}_i\boldsymbol{\beta}), \quad r, s = 1, \dots, p.$$

The maximum likelihood equations

$$\partial \ln L / \partial \beta_r = 0, \quad r = 1, \dots, p$$

are solved by the Newton-Raphson method to get the m.l.e. $\hat{\boldsymbol{\beta}}$. The $p \times p$ observed information matrix is $I = (-\partial^2 \ln L / \partial \beta_r \partial \beta_s)_{\hat{\boldsymbol{\beta}}}$.

In the ‘‘LIFEREG’’ procedure of SAS (SAS Institute Inc., 1985), the exponential distribution is

treated as a Weibull distribution with the scale parameter restricted to the value 1. For restrictions placed on the scale parameter, one degree of freedom Lagrange Multiplier test statistics may be computed. These statistics are computed as

$$\chi^2 = b^2/M,$$

where b is the derivative of the log-likelihood with respect to the scale parameter at the restricted maximum and

$$M = (I_{ii} - I_{ij}I_{jj}^{-1}I_{ji})^{-1},$$

where I is the observed information matrix and the i subscript refers to the scale parameter and the j subscript to the unrestricted parameters. The information matrix is evaluated at the restricted maximum. Under the null hypothesis, these statistics are asymptotically distributed as χ^2 with one degree of freedom.

3 Tree-structured models

Let X denote the $n \times p$ matrix of covariates, $\mathbf{x}^k = (x_{1k}, \dots, x_{nk})$ the n -dimensional vector for the k th covariate and $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})$, the p -dimensional vector of covariates for the i th case. Methods for recursive stratification of the data leading to binary exponential regression trees are described in this section.

3.1 Splitting rules

A binary tree is constructed by splitting the data in each node into two subnodes. Each split is based on a question of the form: Is $x_{ik} \leq c$? Cases satisfying the inequality are sent to the left subnode and otherwise to the right subnode. 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. Starting from the root node, we check if the model fit is appropriate at the node by examining the residuals from the fitted model. In each node, we fit the given parametric regression model using the sample at the node and get the residuals Z , where

$$Z = (\ln W - \mathbf{x}\hat{\boldsymbol{\beta}})/\hat{\sigma}.$$

The data values may be divided into two groups according to the size of the residuals. We slightly change the two splitting methods (the R and M methods) used in Loh (1991) and implement them in exponential regression trees.

1. In the M method, a covariate vector is considered as a class 1 vector if its corresponding residual is larger than the median of the residuals for the sample and as class 2 otherwise. In the R method, the data values are divided into two groups corresponding to non-negative and negative residuals.
2. In both methods, we compute the t -statistic for means and Levene's statistic for variances (Levene's, 1960, test), of the data values in the two groups.
3. The P -value from the larger of the t and Levene's statistics is computed for each predictor. (This assumption is used only for the purpose of ranking the covariates; the P -values are not used for inference.)
4. Suppose the i th covariate yields the smallest P -value. The data in the node are split into two parts, with one subset containing the remaining cases, where c is the average of the two sample means.

This process is repeated at each subsequent node until either the smallest P -value is less than the significance level determined by cross-validation (see below) or there are too few cases left at the node.

3.2 Stopping rules

In order to determine whether or not a node should be split or declared terminal, a measure of goodness-of-fit is needed. We use average squared error as a loss function for cross-validation of the exponential regression trees.

If an independent test sample is available, it can be used to decide whether a node ought to be split as follows. Construct a nested sequence of trees by splitting the node one or more times. Run the test sample down each tree in the sequence to obtain an estimate of average squared error. If at least one of the trees possesses a smaller average squared error than the node in question, the latter is split. Otherwise it is declared terminal.

In the absence of an independent test sample, the process can be mimicked through V -fold cross-validation. Starting with the root node, the decision to split a node is made through cross-validation. Let G be the tree constructed from $\mathcal{L} = \{(y_i, d_i, \mathbf{x}_i) | 1 \leq i \leq n\}$, where d_i is the censoring indicator and the cases in \mathcal{L} be randomly divided into V subsets $\mathcal{L}_1, \dots, \mathcal{L}_V$ each containing nearly the same number of cases. For each subset \mathcal{L}_v , $v = 1, \dots, V$, use the data in $\mathcal{L}^{(v)} = \mathcal{L} - \mathcal{L}_v$ to construct a regression tree TR^v and then use the data in \mathcal{L}_v to obtain the cross-validation estimate $R^{CV}(v)$, $v = 1, \dots, V$. The V -fold cross-validation estimate is defined as

$$R^{CV} = \frac{1}{V} \sum_{v=1}^V R^{CV}(v),$$

where $N_v \simeq N/V$ is the number of cases in \mathcal{L}_v . The average squared error

$$R^{CV}(v) = \frac{1}{N_v} \sum_{y_j \in \mathcal{L}_v} (y_j - \hat{y}_j)^2, \quad v = 1, \dots, V$$

as the cross-validation estimate.

If a cross-validation tree has an estimate of average squared error that is at least $(1 - f)$ times smaller than that for the trivial tree, then it is considered “superior” to the trivial tree because of variability due to cross-validation. Here f is either user-specified or estimated by the bootstrap. If the proportion of times (out of V) that a superior cross-validation tree is found exceeds another pre-specified number η , the node is split. See the Appendix for further details.

3.3 Categorical and missing values

Any covariate that takes categorical values is transformed into a dummy vector of 0-1 indicator variables for the purpose of fitting exponential regression models. If a continuous covariate has missing values, the latter are substituted with class means estimated from the nonmissing values (Dixon *et al.*, 1985) prior to fitting exponential models. If a categorical covariate has missing values, they are replaced with the mode of the nonmissing values. This step is repeated at every node.

3.4 Bootstrap selection of parameter values

Because the values of f and η help to determine the size of the tree, a procedure is needed to control the probability of spurious splits. We use the classical hypothesis testing approach of bounding the probability α , say, that a non-trivial tree results when in fact a single exponential regression model suffices for all the data.

The bootstrap method offers a convenient way to achieve this goal. Let $\hat{\alpha}(f, \eta)$ be the bootstrap

estimate of α , under the hypothesis that no splits are needed. Let \hat{f} and $\hat{\eta}$ be the values such that $\hat{\alpha}(\hat{f}, \hat{\eta})$ is closest to α . Three methods of reducing the number of candidate values for f and η may be used.

1. Fixing $f = \eta$, choose the value of f for which $\hat{\alpha}(f, f)$ is closest to α .
2. Fixing $f = 0$, select the value of η for which $\hat{\alpha}(0, \eta)$ is closest to α .
3. Fixing $\eta = 0.5$, choose the value of f for which $\hat{\alpha}(f, 0.5)$ is closest to α .

We use a finite grid with increments of 0.1 in each case. Full details of the bootstrap procedure are given in the Appendix.

4 Examples

The proposed methods were tested on the Stanford heart transplant data and several artificial data sets. We report the results in this section. The value of α was chosen to be .05 in all the examples.

4.1 Simulated data

4.1.1 One exponential regression model

In the first simulation experiment, survival times were generated from an exponential distribution with mean $\mu_i = e^{\mathbf{x}_i \boldsymbol{\beta}}$, where $\boldsymbol{\beta} = (2, 2)'$, $\mathbf{x}_i = (1, x_{i1})$ and $x_{i1} \in \{\pm 1, \pm 2, \pm 3, \pm 4\}$. 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 2500 so that about 20% of the observations were censored. Fifty simulation trials were performed for each of the R and M methods and each of the three bootstrap methods for choosing f and η .

The results are given in Table 1. The number of splits and the number of times they were observed are shown in the second and third columns of the table. Since the data were generated from

Table 1: Simulation results for one exponential regression model using the bootstrap to choose f and/or η . Nominal significance level is $\alpha = 0.05$; 20% censoring; 50 simulations.

Bootstrap method	M method		R method	
	#splits	freq.	#splits	freq.
1st ($f = \eta$)	0	48	0	46
	1	2	1	2
			2	2
2nd ($f = 0$)	0	41	0	45
	1	8	1	5
	2	1		
3rd ($\eta = .5$)	0	50	0	45
			1	3
			2	2

Table 2: Simulation results for two exponential regression models, using the bootstrap to find f and/or η . Significance level is $\alpha = 0.05$, 20% censoring, 50 trials.

Bootstrap method	M method		R method	
	#splits	freq.	#splits	freq.
1st ($f = \eta$)	0	1	0	0
	1	47	1	45
	2	2	2	5
2nd ($f = 0$)	0	27	0	36
	1	23	1	14
3rd ($\eta = .5$)	0	2	0	0
	1	47	1	47
	2	1	2	3

a single exponential regression model, the correct trees are those with no splits. The probability of a type I error appears to be quite satisfactory, especially in the first and third bootstrap methods.

4.1.2 Two exponential regression models

In the second experiment, data were generated from two exponential regression models, the purpose being to compare the power of the individual methods in detecting the need to partition the data. Table 2 gives the simulation results. The powers are larger for the first and third bootstrap estimation methods in both the M and R methods.

Table 3: Coefficient estimates and standard errors for Cox regression of log survival time on age and mismatch for the Stanford heart transplant data, with and without five patients with survival times less than 10 days.

157 cases	Age		Mismatch	
	Estimate	S.E.	Estimate	S.E.
	0.030	0.011	0.167	0.183
152 cases	Age		Age-squared	
	Estimate	S.E.	Estimate	S.E.
	-0.146	0.055	0.0023	0.0007

4.2 Stanford heart transplant data

We apply our method to the Stanford heart transplant data. The data are reported in Miller and Halpern (1982) and consist of 184 patients who received heart transplants. The covariates are age at transplant and $T5$ mismatch score (a measure of dissimilarity between donor and recipient tissue), survival time in days after transplant and a censoring indicator (0 for censored, 1 for uncensored). Following Miller and Halpern (1982) and Segal (1988), the 27 patients who did not have mismatch scores were excluded from our analysis. Other papers that refer to these data are Crowley and Hu (1977), Kalbfleisch and Prentice (1980), Miller and Halpern (1982) and Aitkin *et al.* (1983).

Miller and Halpern (1982) analyzed the data using Cox (1972), Buckley-James (1979) and Miller (1976) regression methods. Fitting $\log(\text{survival time})$ on age and $T5$ mismatch score, they concluded that mismatch score was not significant and that a quadratic model in age was satisfactory. Miller and Halpern's analysis excluded 5 patients with survival times less than 10 days. The results of their analysis using Cox regression are shown in Table 3. 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.

Figure 1(a) shows a scatterplot of $\log(\text{survival time})$ (to base 10) versus age at transplant, with a smooth curve superimposed, for the 152 patients who survived at least 10 days. While it is

Table 4: Regression estimates of the coefficients, Wald test statistics and P -values on the covariates for the Stanford heart transplant data with the exponential regression model. One case with zero survival time was deleted.

Variable	$\hat{\beta}$	S.E.	$\hat{\beta}/\text{S.E.}$	χ^2 (1 d.f.)	P -value (Wald test)
intercept	9.091	0.577	15.75	247.91	< 0.0001
age	-0.040	0.012	-3.29	10.85	0.0010
mismatch	-0.307	0.191	-1.61	2.58	0.1079

Lagrange Multiplier χ^2 for scale is 33.89, P -value < 0.0001.

clear that a quadratic in age is preferable to a linear one, it is seen that the smooth curve is quite different from a parabola. If the data are divided at some point of age between 40 and 45 years, then a linear fit might be adequate in each group. Figure 1(b) shows the same plot without taking logarithms of survival times.

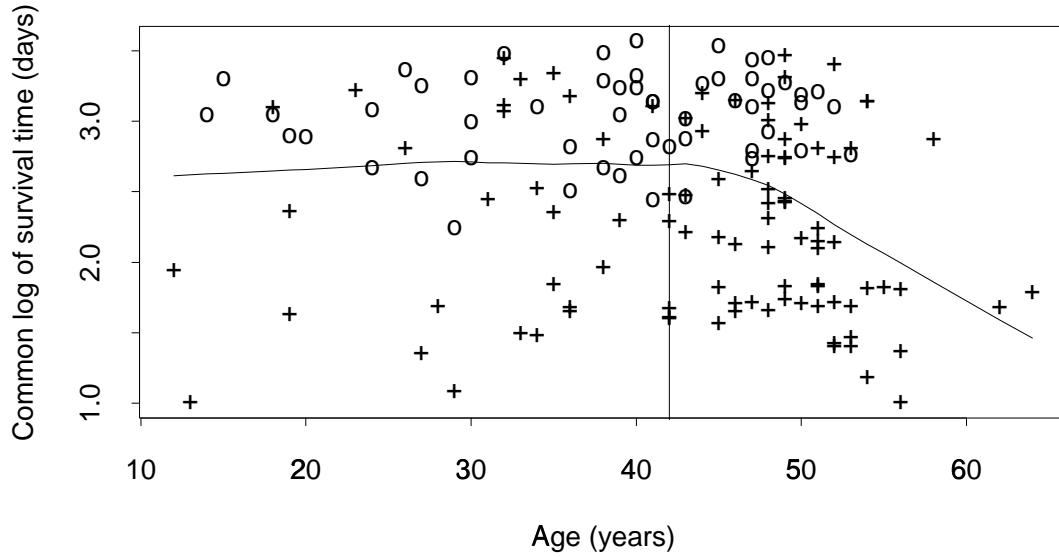
On the other hand, an exponential regression model is fitted to the entire sample. One case with zero survival time is deleted before fitting the model since we use the log-survival times in parametric regression models. Thus 156 cases are used in this paper. Table 4 gives the regression estimates, Wald test statistics and the P -values for the coefficients. As in Cox regression, survival time is shorter for older patients (P -value is 0.001), but mismatch score is not significant at level 0.05. The Lagrange multiplier χ^2 test statistic for the scale obtained from SAS is 33.89 and the P -value is less than 0.0001. Therefore, the Weibull regression model seems better than the exponential regression model for the whole data.

Before we report our results of the exponential regression trees, we define $node(i, j)$ as the j th node from the left at the i th level. The root node is $node(0, 1)$.

4.2.1 Using the M method

The M method gave trivial trees for the second bootstrap estimation method. The third bootstrap estimation method with $\eta = 0.5$ gave the tree in Figure 2. Figure 3 shows the Kaplan-Meier estimates of the survival distributions of the terminal nodes and Table 5 gives the regression esti-

(a) $\log(\text{survival time})$ vs. age



(b) survival time vs. age

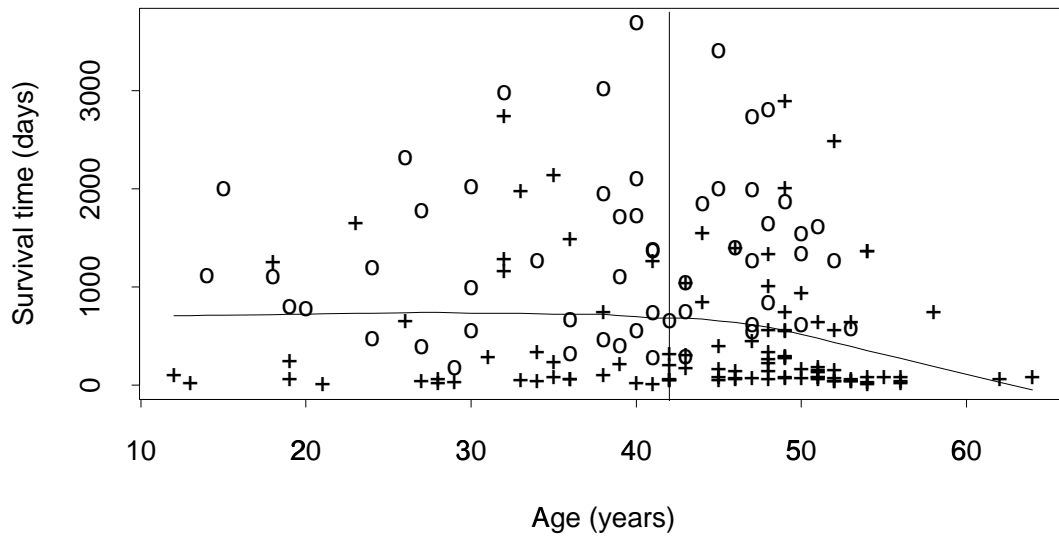


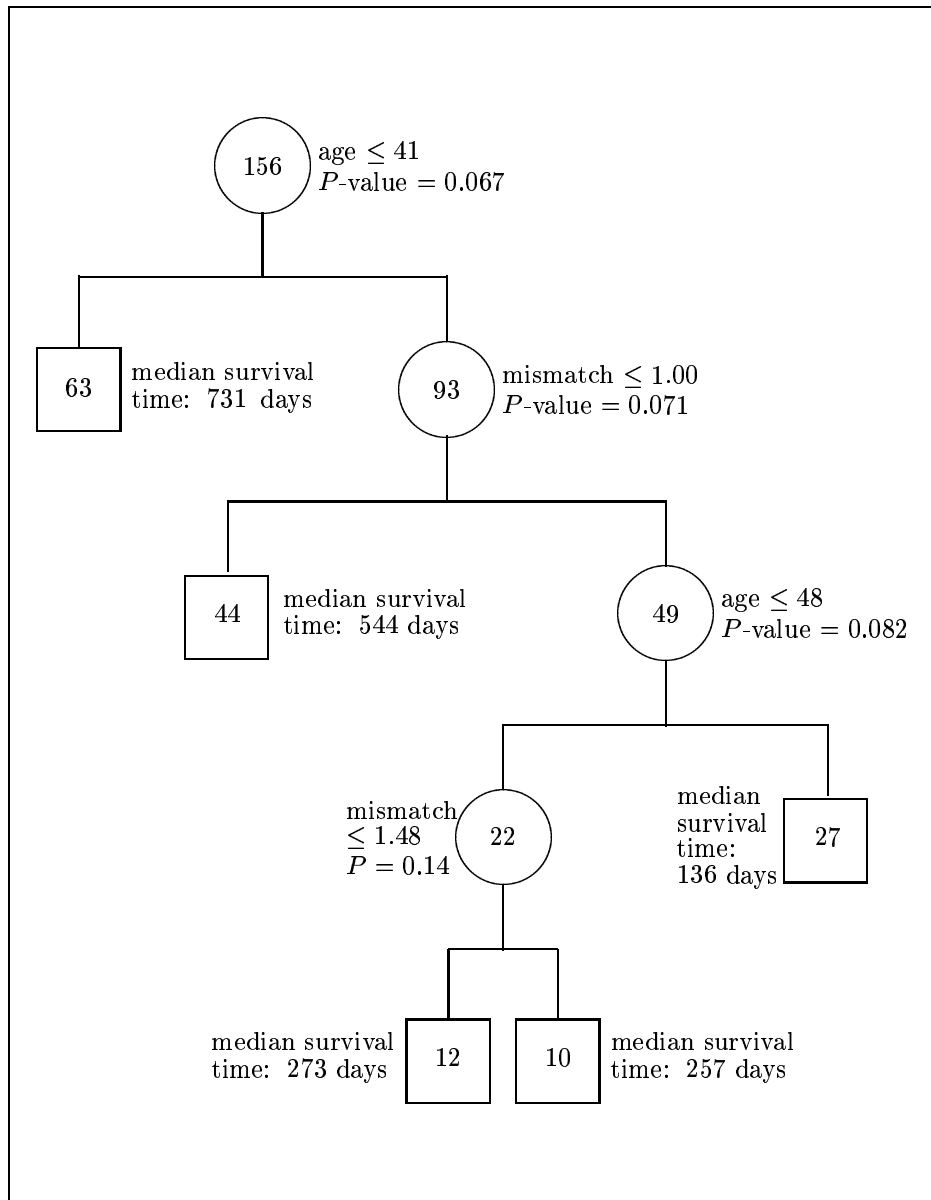
Figure 1: (a) Scatterplot of \log_{10} survival time (days) versus age at transplant (years) with a smooth curve for 152 Stanford heart transplant patients who survived at least 10 days. (b) Scatterplot of survival time versus age at transplant with a smooth curve for 157 Stanford heart transplant patients.

Table 5: Regression estimates of the coefficients and P -values of Wald test at the terminal nodes of the tree in Figure 2.

Node	Variable	$\hat{\beta}$	S.E.	$\hat{\beta}/\text{S.E.}$	P -value (Wald test)
age \leq 41	intercept	6.254	0.732	8.55	< 0.0001
	age	0.033	0.021	-1.55	0.1204
	mismatch	0.270	0.335	0.81	0.4193
Lagrange Multiplier χ^2 for scale is 13.73, P -value is 0.0002.					
age > 41 & mismatch \leq 1.00	intercept	13.540	2.314	5.85	< 0.0001
	age	-0.151	0.046	-3.24	0.0012
	mismatch	1.337	0.719	1.86	0.0628
Lagrange Multiplier χ^2 for scale is 7.85, P -value < 0.0051.					
age > 48 & mismatch > 1.00	intercept	26.440	7.810	4.83	0.0007
	age	-0.386	0.160	-1.71	0.0158
	mismatch	-0.248	0.577	-1.96	0.6679
Lagrange Multiplier χ^2 for scale is 3.77, P -value is 0.0521.					
41 < age \leq 48 & 1.00 < age \leq 1.48	intercept	-3.223	7.454	-0.43	0.6655
	age	0.145	0.157	0.92	0.3557
	mismatch	2.750	2.474	1.11	0.2663
Lagrange Multiplier χ^2 for scale is 1.05, P -value is 0.3060.					
41 < age \leq 48 & mismatch > 1.48	intercept	2.916	9.604	0.30	0.7614
	age	0.299	0.231	1.29	0.1969
	mismatch	-5.736	1.764	-3.25	0.0011
Lagrange Multiplier χ^2 for scale is 0.34, P -value is 0.5578.					

mates and P -values for the coefficients. The method resulted in 4 splits, yielding 5 terminal nodes ($node(1, 1)$ for age \leq 41 with median survival time 731 days, $node(2, 3)$ for age > 41 and mismatch \leq 1.00 with median survival time 544 days, $node(3, 8)$ for age > 48 and mismatch > 1.00 with median survival time 136 days, $node(4, 13)$ for 41 < age \leq 48 and 1.00 < mismatch \leq 1.48 with median survival time 273 days and $node(4, 14)$ for 41 < age \leq 48 and mismatch > 1.48 with median survival time 257 days).

In $node(1, 1)$ (age \leq 41) and $node(4, 13)$ (41 < age \leq 48 and 1.00 < age \leq 1.48), the two factors are not significant at level 0.05. The survival time is shorter for the older patients in $node(2, 3)$ (age > 41 and mismatch \leq 1.00) and $node(3, 8)$ (age > 48 and mismatch > 1.00). Mismatch score is not significant in all the terminal nodes except $node(4, 14)$ (41 < age \leq 48 and mismatch



æ

Figure 2: Exponential regression tree for the heart transplant data with the M method and $f = 0.05, \eta = 0.5$. The numbers within circles or squares are sample sizes. The P -value beside each node refers to the maximum of the t -test for the means and Levene's test. One case with zero survival time was deleted.

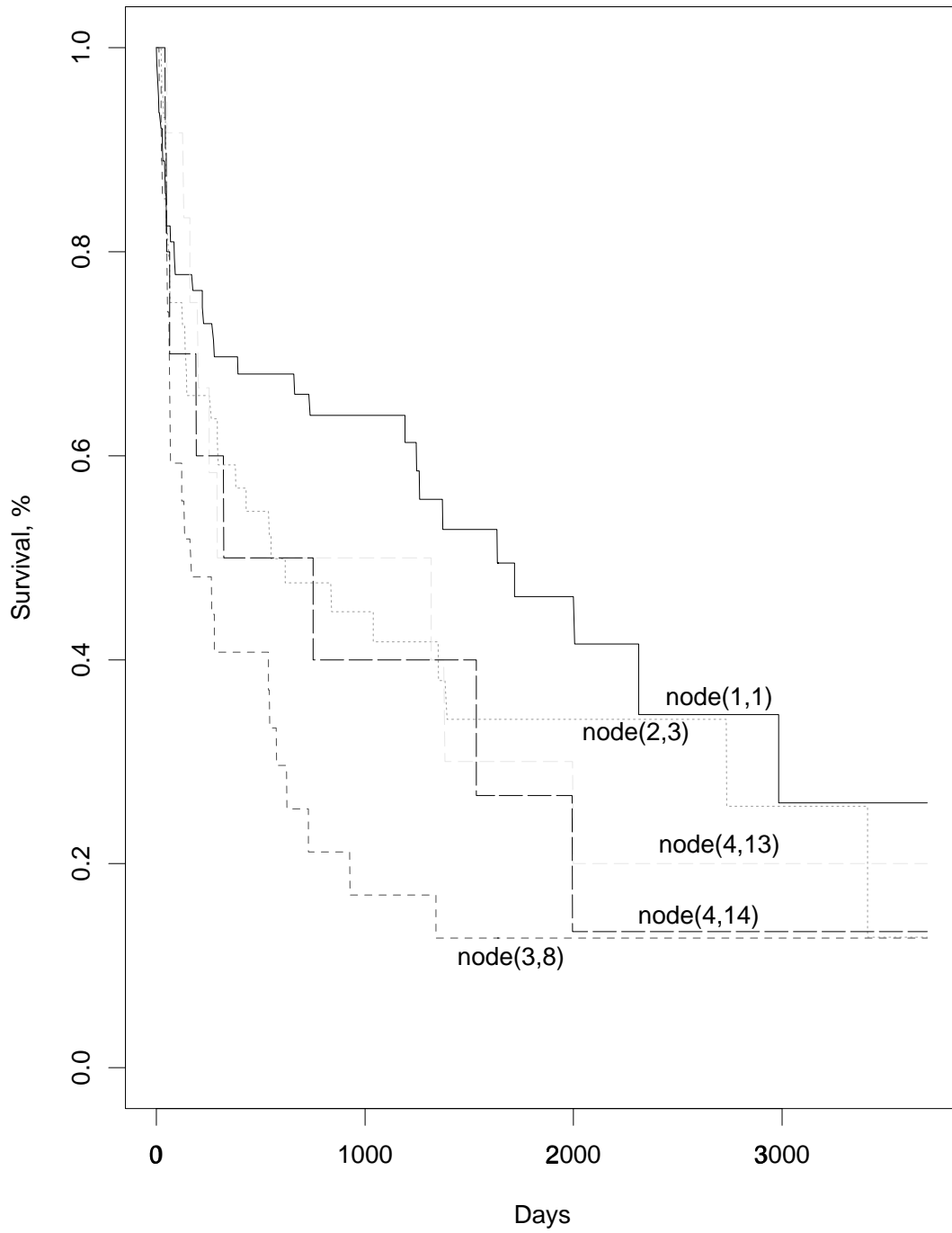


Figure 3: Kaplan-Meier survival curves for the terminal nodes of the tree in Figure 2.

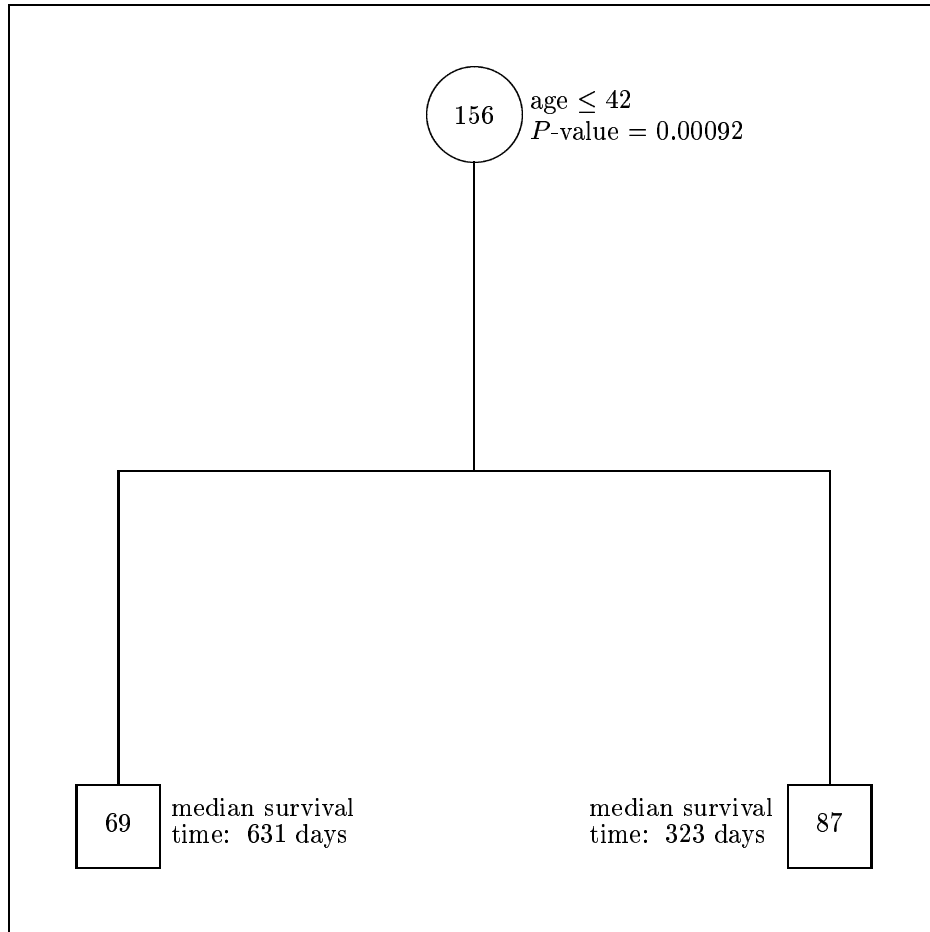
Table 6: Regression estimates and P -values of Wald test at the terminal nodes of the exponential regression tree in Figure 4.

Node	Variable	$\hat{\beta}$	S.E.	$\hat{\beta}/\text{S.E.}$	P -value (Wald test)
age \leq 42	intercept	6.877	0.741	9.28	< 0.0001
	age	0.012	0.021	0.57	0.5685
	mismatch	0.169	0.307	0.55	0.5816
Lagrange Multiplier χ^2 for scale is 16.67, P -value < 0.0001.					
age > 42	intercept	15.298	1.693	9.04	< 0.0001
	age	-0.159	0.033	-4.77	< 0.0001
	mismatch	-0.565	0.230	-2.46	0.0141
Lagrange Multiplier χ^2 for scale is 14.06, P -value is 0.0002.					

> 1.48). The P -values of the Lagrange multiplier χ^2 test for the restriction of the scale at 1 are significant in $node(1, 1)$ and $node(2, 3)$, but not significant in $node(3, 8)$, $node(4, 13)$ and $node(4, 14)$ at level 0.05. Although the Weibull regression model is more adequate than the exponential regression model for the whole data, the exponential regression model is adequate in the three nodes $node(3, 8)$, $node(4, 13)$ and $node(4, 14)$. The first bootstrap method with $f = \eta$ gave three more splits than the tree obtained from the third method. The splits occur at $node(2, 3)$, $node(3, 6)$ and $node(3, 8)$.

4.2.2 Using the R method

The R method gave a trivial tree for the first bootstrap method. The second bootstrap method with $f = 0$ and third bootstrap method with $\eta = 0.5$ gave trees with one split at age 42.7 years with a P -value of 0.0009. Figure 4 shows the tree obtained using these methods and Figure 5 displays the Kaplan-Meier estimates of the survival distributions for the terminal nodes. Table 6 gives the regression estimates. Neither covariate was significant at level 0.05 for the group of patients whose ages were less than 42 years. For the other group, survival times tend to be shorter for the older patients and for the patients with larger mismatch scores.



æ

Figure 4: Exponential regression tree with the second and the third bootstrap and R methods for the heart transplant data. The numbers within circles or squares are sample sizes. The *P*-value beside each node refers to the maximum of the *t*-test for the means and Levene's test. One case with zero survival time was deleted.

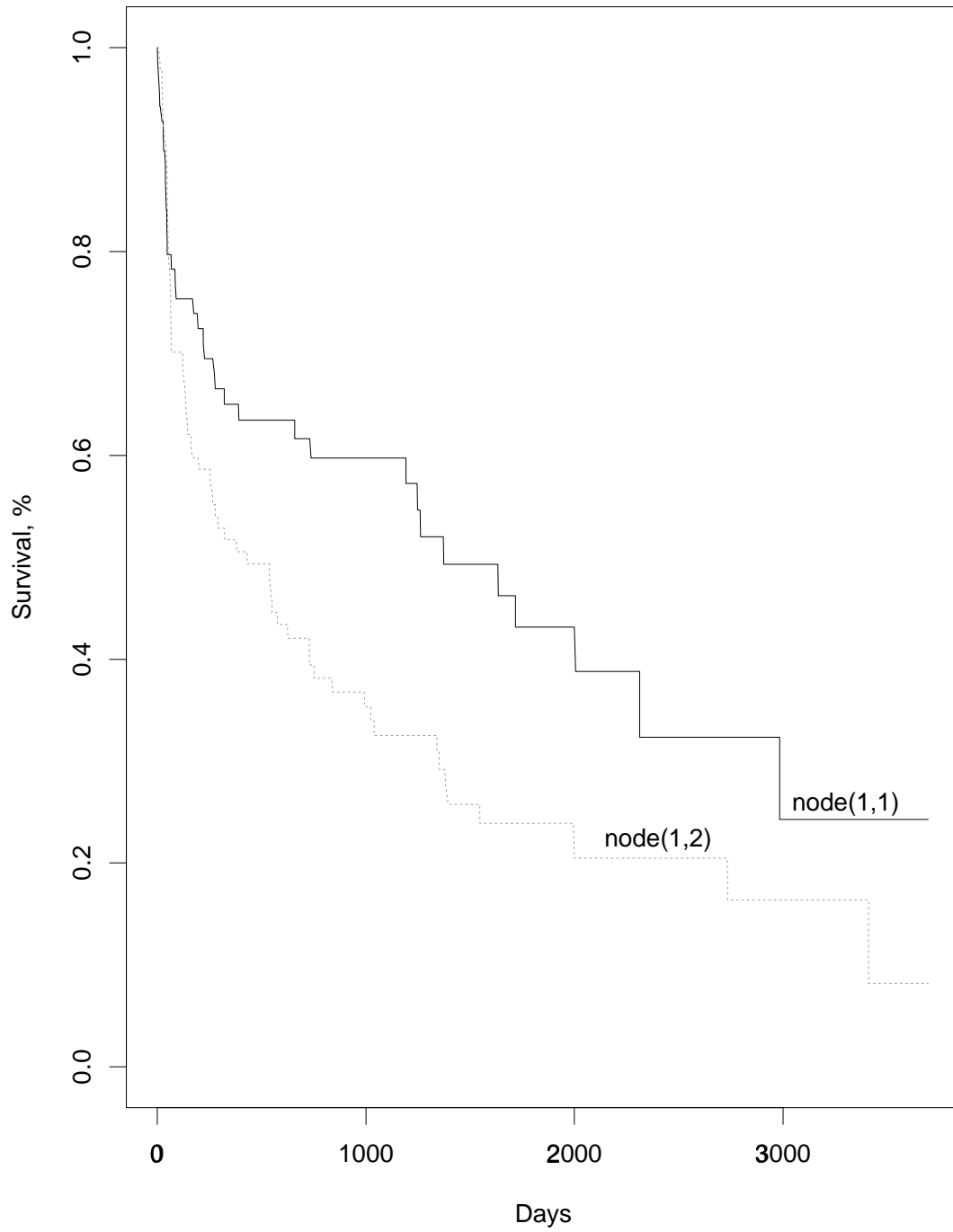


Figure 5: Kaplan-Meier survival curves for the terminal nodes of the tree in Figure 4.

5 Conclusion

This research was motivated by the twin goals of (i) providing a formal goodness-of-fit test for the exponential regression model and (ii) keeping the fitted models as simple as possible, for ease of interpretation. It turns out that these goals can be met by using recursive stratification and the bootstrap. Stratification is a powerful technique that can break down complex models into simple pieces linked together by a decision tree. Bootstrap estimation of the probability that a single exponential regression model is erroneously rejected as inadequate provides the formal test of fit.

Several forms of strata selection and bootstrapping were considered to study their relative effectiveness as well as to demonstrate the variety of techniques available. The examples suggest that any form of bootstrapping would produce satisfactory control of the probability of a type I error. The first ($f = \eta$) and third ($\eta = 0.5$) bootstrap methods seem to yield the best power in both the M and R methods.

Appendix: Outline of the algorithm

We present a sketch of our algorithm here. First we need some notation.

- \mathbf{t} : response vector
- \mathbf{y} : $\ln \mathbf{t}$
- X : matrix of the covariates
- \mathbf{d} : vector of censoring indicators
- $intcpt$: inclusion indicator for the intercept term in the regression. Equals 0 if we do not add a vector of 1's to the design matrix, equals 1 if we add a vector of 1's for the intercept term.

The default value is 0.

- *mindat*: user-specified terminal node sample size
- *tol*: maximum error in Newton-Raphson iterations
- *flag*: diagnostic flag for exponential regression. Equals 1 if the Newton-Raphson method of exponential regression fails to converge in 20 iterations, equals 2 if the information matrix is almost singular at some stage of the iterations in the Newton-Raphson method, and equals 0 otherwise.
- $\text{node}(i, j)$: j th node from the left at the i th level. The root node is $\text{node}(0, 1)$.
- $\text{pos}(i, j, k)$: position of the subsample in $\text{node}(i, j)$ among the whole sample. $\text{pos}(i, j, 1)$ is the initial position, $\text{pos}(i, j, 2)$ is the end position of the subsample.
- *count*: the number of exponential regression fits at the current level. If *count* = 0 after checking all the nodes, then there is no more split after the node; therefore, exit the loop.

Main program

The exponential regression tree algorithm is as follows:

1. Read data \mathbf{t} , X and \mathbf{d} .
2. Choose splitting method (R or M).
3. Add the intercept term if $\text{intcpt} = 1$.
4. Initialize the values of tol , f , η , mindat and set $\text{flag} = 0$.
5. Set $\mathbf{y} = \ln \mathbf{t}$.
6. If a covariate contains categorical variables, then indicator variables are generated for the levels of the variable.

7. Loop over levels $i = 0, 1, \dots$

- $count = 0$.

- Loop over nodes $j = 1, \dots, 2^i$ (i th level has at most 2^i nodes). If $pos(i, j, 1) \neq 0$, then do:

- (a) Get data from X , \mathbf{y} , \mathbf{d} for the node. The initial position is $pos(i, j, 1)$ and the end position is $pos(i, j, 2)$.

- (b) If necessary, estimate the missing data. \mathbf{y} is sorted in ascending order and X , \mathbf{d} are reordered according to \mathbf{y} .

- (c) Fit the exponential regression model. If $flag = 1$ or $flag = 2$, set the node directly above as a terminal and exit.

- (d) Get the residuals,

- (e) Apply the R or M method to divide the sample into two classes provided that the numbers of observations in both of the classes are greater than $mindat$.

- (f) Apply cross-validation to split the node.

- (g) Increase count by 1.

- end the loop for j (for node in the same level).

- if $count = 0$, then exit.

8. end the loop for i (level).

Cross-validation subprogram

Cross-validation is used to decide whether or not to split a node. Let $node(i, j)$ and $\mathcal{L}(i, j)$ be the current node and the sample in it respectively. $\mathcal{L}(i, j)$ is randomly divided into V nearly equal parts $\mathcal{L}_1, \dots, \mathcal{L}_V$ and the following process repeated for $v = 1, \dots, V$.

1. Grow a large tree T_{v_0} using the cases in $\mathcal{L}^{(v)} = \mathcal{L} - \mathcal{L}_v$. A node is terminal in this tree only if one or more of the following conditions hold:

- (a) There are too few cases in the node.
- (b) The Newton-Raphson method does not converge in 20 iterations when fitting an exponential regression model to one of the subnodes.
- (c) 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 t and Levene's tests for node (i, j) . Suppose there are s distinct values of p_{ij} 's. Sort the p_{ij} 's in ascending order and adjoin $p_0 = 0$ and $p_{s+1} = 1$ to this set so that $0 = p_0 < p_1 < \dots < p_s < p_{s+1} = 1$.

2. Compute $\gamma_l = (p_l + p_{l+1})/2$ for $l = 0, 1, \dots, s$.

3. Prune T_{v_0} at level γ_l to obtain 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 $T_{s+1} = T_{v_0}$.

- Loop over $k = s, s - 1, \dots, 1, 0$.
 - Starting from the lowest level to the root node of T_{k+1} (loop over levels $i = a, \dots, 0$, where a is the lowest level of T_{k+1}), do the following.
 - (a) At level i , loop over $j = 1, \dots, 2^i$, starting with the left node. At node (i, j) :
 - i. If the node is intermediate and the two children nodes are terminal, let p be the P -value of the split at the node.
 - A. If $p < \gamma_k$, go to the next node.
 - B. If $p \geq \gamma_k$, delete the two children nodes and make node (i, j) terminal.
 - ii. Otherwise, go to the next node.

- (b) End the loop for j .
 - End the loop for i . This gives the pruned tree T_k at level γ_k .
 - End the loop for k .
4. Let $f \in (0, 1)$ be the user-specified fractional reduction in the average squared error.

Set $\theta(v) = 0$.

Loop over $k = 1, \dots, s$.

If $R^{CV}(v, k) < (1 - f)R^{CV}(v, 0)$, set $\theta(v) = 1$ and exit.

Otherwise increment k to $k + 1$ and go to the preceding line.

Let $\eta \in (0, 1)$ be the pre-selected splitting threshold and $\theta = \sum_{v=1}^V \theta(v)$. If $\theta > \eta V$, the node t is split; otherwise it is declared terminal. (This method was used in Huang, 1989, for tree-structured regression.)

Computational details

In applying the Newton-Raphson method in exponential regression, we use the LU decomposition to find the solution at each iteration instead of calculating the inverse matrix. The LU factorization algorithm for a symmetric matrix (Hager, 1988, page 87) is used. We use the least squares estimate $\mathbf{b} = (b_1, \dots, b_p)$ of $\boldsymbol{\beta}$ as the initial value of the regression parameters in the iteration. If the Newton-Raphson method does not converge within 20 iterations, do the following:

1. Choose $b_i + 0.5\text{se}(b_i)$ or $b_i - 0.5\text{se}(b_i)$ as initial guess for up to four i 's and go to step 2.
2. If all the initial values in (a) are unsuccessful, then choose $b_i + 0.7\text{se}(b_i)$ or $b_i - 0.7\text{se}(b_i)$ as initial guess for up to four i 's and go to step 2.

3. If all the initial values in (a) and (b) are unsuccessful, then choose $b_i + 0.3se(b_i)$ or $b_i - 0.3se(b_i)$ as initial guess for up to four i 's and go to step 2.
4. If all the initial values in (a), (b) and (c) are unsuccessful, then declare that the Newton-Raphson method diverges and exit.

More detailed description of this procedure is given in Ahn (1992).

The information matrix in the Newton-Raphson method is singular if there are no complete (uncensored) observations among the data to be analyzed. Thus the program will also stop if an information matrix is found to be almost singular.

Bootstrap parameter selection

The bootstrap algorithm for choosing the best values for f and η consists of the following three components.

1. Generation of bootstrap survival times T^* . For the exponential regression model, we generate survival times from the residuals. To generate a bootstrap observation T^* , we randomly choose n numbers from $\{1, 2, \dots, n\}$ with replacement. Let the set of chosen numbers be $\{i_1, \dots, i_n\}$. It is possible that some of the i'_j s are the same. Let the estimate of the j th survival time be $t_j^* = z_{i_j}$, where z_{i_j} is the i_j th residual of the given parametric regression model. The associated covariate vector is \mathbf{x}^j .
2. Estimation of the censoring distribution and generation of bootstrap censoring times C^* . We assume that the real censoring times C are independent of the real survival times T and the covariates X . From the real data \mathbf{y} and censoring indicator \mathbf{d} , switch the censoring status so that if the i th individual is "died" ($d_i = 1$), it is changed to "censored" ($d_i = 0$) and vice

versa. Our estimate of the censoring distribution is then given by the Kaplan-Meier estimate of the modified data. Bootstrap censoring times C^* may now be generated as above.

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 $y_i^* = \min\{t_i^*, c_i^*\}$, $d_i^* = I(t_i^* \leq c_i^*)$. These artificial data and the observed covariate values are then used to get bootstrap estimates of the probability of the type I error of splitting the root node when it should not be split.

References

- Ahn, H. (1992) "Survival Modeling through Regression Trees". Unpublished Ph.D. Thesis, Department of Statistics, University of Wisconsin-Madison.
- Aitkin, M., Laird, N. and Francis, B. (1983). "A reanalysis of the Stanford heart transplant data". *Journal of the American Statistical Association*, **78**, 264-274.
- Buckley, J. and James, I. (1979). "Linear regression with censored data". *Biometrika*, **66**, 429-436.
- Cox, D. R. (1972). "Regression models and life-tables". *Journal of the Royal Statistical Society B*, **34**, 187-202.
- Cox, D. R. and Snell, E. J. (1968). "A general definition of residuals". *Journal of the Royal Statistical Society B*, **30**, 248-275.
- Crowley, J. and Hu, M. (1977). "Covariance analysis of heart transplant survival data". *Journal of the American Statistical Association*, **72**, 27-36.
- Dixon, W. J., Brown, M. B., Engelman, L., Frane, J. W., Hill, M. A., Jennrich, R. I. and Toporek, J. D. (1985). *BMDP Statistical Software*. University of California Press, Berkeley.
- Feigl, P., and Zelen, M. (1965). "Estimation of exponential survival probabilities with concomitant information". *Biometrics*, **21**, 826-838.
- Glasser, M. (1967). "Exponential survival with covariance". *Journal of the American Statistical Association*, **62**, 561-568.
- Hager, W. W. (1988). *Applied numerical linear algebra*. Prentice Hall, Englewood Cliffs, New Jersey.
- Huang, M. C. (1989). "Piecewise linear tree-structured regression". Unpublished Ph.D. Thesis, Department of Statistics, University of Wisconsin-Madison.

- Kalbfleisch, J.D. and Prentice, R.L. (1980). *The statistical analysis of failure time data*. Wiley, New York.
- Lawless, J. F. (1976). "Confidence interval estimation in the inverse power law model". *Applied Statistics*, **25**, 128-138.
- Levene, H. (1960). Robust tests for equality of variances. In *Contributions to probability and Statistics*, (Olkin, I., *et al* eds.), 278-292. Stanford University Press.
- 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.
- Prentice, R. L. (1973). "Exponential survival with censoring and explanatory variables". *Biometrika*, **60**, 279-288.
- SAS Institute Inc. (1985). *SAS User's Guide: Statistics, 1985 edition*. Cary, NC: SAS Institute Inc.
- Segal, M. R. (1988). "Regression trees for censored data". *Biometrics*, **44**, 35-47.
- 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.

Address

Hongshik Ahn
Division of Biometry and Risk Assessment
NCTR/FDA/HFT-20
Jefferson, AR 72079
U.S.A.