Tree-structured Proportional Hazards Regression Modeling[*]

Hongshik Ahn and Wei-Yin Loh
Department of Statistics, University of Wisconsin-Madison,
Madison, Wisconsin 53706-1685, U.S.A.

Abstract

A method for fitting piecewise proportional hazards models to censored survival data is described. Stratification is performed recursively, using a combination of statistical tests and residual analysis. 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 predetermined value. The proposed method can thus also serve as a formal goodness-of-fit test for the proportional hazards model. Real and simulated data are used for illustration.
KEY WORDS: Bootstrap; Cox regression; Recursive partitioning; Survival analysis.

1 Introduction

The problem of modeling censored survival data has attracted much attention in recent years. One of the most popular techniques is the Cox (1972) proportional hazards model, which postulates that the logarithm of the hazard rate is a linear function of the covariates. Like the normal regression method that it emulates, Cox's method is powerful because it permits the fitting of a very rich class of models. On the other hand, it suffers from the following disadvantages:

- The proportional hazards model is often difficult to interpret, especially when there are numerous covariates, some of which being correlated. This is a well-known problem in the normal regression literature.
- Unlike normal regression, where there are standard lack-of-fit tests, there are no such tests for proportional hazards models. (See, however, Wei, 1984 and Lin and Wei, 1991 for some recent progress.)

One way to resolve both criticisms is to stratify the data according to particular covariate values and fit separate Cox models to each stratum. The manner in which the data are stratified can yield useful information about the data structure. Furthermore, since the data within a stratum would be more homogeneous, they may be fitted by models with fewer covariates, and hence alleviate the difficulties associated with interpretation and collinearity.

2 The proportional hazards model

Let $T_1, \ldots, T_n$ and $C_1, \ldots, C_n$ be independent random variables, where C_i is the censoring time associated with the survival time T_i, $i = 1, \ldots, n$. We observe $(Y_1, \delta_1), \ldots, (Y_n, \delta_n)$, where $Y_i =$

[*] Research partially supported by NSF grant DMS88-03271 and ARO grant DAAL03-91-G-0111.

$\min\{T_i, C_i\}$, $\delta_i = I(T_i \leq C_i)$ and $I(\cdot)$ is the indicator function. Assume that for each i, a p-dimensional covariate vector $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})$ independent of T_i is available. Let $h(t;\boldsymbol{x}) = f(t,\boldsymbol{x})/\{1 - F(t,\boldsymbol{x})\}$ be the hazard rate at time t for an individual with covariate vector $\boldsymbol{x}$, where $f(t,\boldsymbol{x})$ is the density and $F(t,\boldsymbol{x})$ the distribution function of T at $\boldsymbol{x}$. The proportional hazards model is given by $h(t;\boldsymbol{x}) = h_0(t)\exp(\boldsymbol{x}\boldsymbol{\beta})$, where $\boldsymbol{\beta}$ is a vector of unknown regression coefficients and $h_0(t)$ is the (unknown) hazard function for an individual with covariate vector $\boldsymbol{x} = 0$.

Denote the r ordered distinct times to death among the n survival times by $t_1 < \ldots < t_r$ and let D_i be the set of individuals failing at time t_i, m_i be the number of deaths occurring at t_i, (i.e., the number of elements in D_i), R_i be the set of individuals alive and under observation just prior to t_i, and $\boldsymbol{x}_l(t_i)$ be the p-dimensional vector of covariates for the lth individual at time t_i. Consider the function (Kalbfleisch and Prentice 1972, Peto 1972, Breslow 1974, Miller 1981)

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{r} \frac{\exp(\sum_{j\in D_i} \boldsymbol{x}_j(t_i)\boldsymbol{\beta})}{(\sum_{l\in R_i} \exp(\boldsymbol{x}_l(t_i)\boldsymbol{\beta}))^{m_i}}. \tag{1}$$

The function (1) does not depend on $h_0(t)$ and can be maximized to yield an estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$. Although $L(\boldsymbol{\beta})$ is not a likelihood function in the usual sense, it is well-known (see, e.g., Cox 1975) that the function can be treated as an ordinary likelihood for purposes of inference about $\boldsymbol{\beta}$.

Let $H(t,\boldsymbol{x}) = \int_0^t h(s,\boldsymbol{x})\,ds$ be the cumulative hazard function, estimated by (see, e.g., Link 1979)

$$\hat{H}(t,\boldsymbol{x}) = \exp(\boldsymbol{x}\hat{\boldsymbol{\beta}}) \left[\sum_{i=1}^{l}(t_i - t_{i-1})\hat{h}_{0i} + (t - t_l)\hat{h}_{0l+1}\right],$$

where $t_0 = 0$, $t_l < t \leq t_{l+1}$ and $\hat{h}_{0i} = m_i/\{(t_i - t_{i-1})\sum_{j\in R_i}\exp(\boldsymbol{x}_j\hat{\boldsymbol{\beta}})\}$ is the maximum likelihood estimate of h_0 at t_i. The estimated survival function for covariate pattern $\mathbf{u}$ is $\hat{S}(t,\mathbf{u}) = [\hat{S}_0(t)]^{\exp(u\hat{\beta})}$, where $\hat{S}_0(t) = \exp[-\hat{H}(t,\mathbf{0})]$ is the estimated baseline survival function. The log-minus-log plot of $\ln[-\ln\hat{S}(t,\bar{\boldsymbol{x}})]$ is obtained by plotting the function $\ln\hat{H}(t,\bar{\boldsymbol{x}})$ against t, where $\bar{\boldsymbol{x}}$ is the mean covariate vector.

3 Tree-structured models

Let X denote the $n \times p$ matrix of covariates, $\boldsymbol{x}^k = (x_{1k}, \cdots, x_{nk})$ the n-dimensional vector for the kth covariate and $\boldsymbol{x}_i = (x_{i1}, \cdots, x_{ip})$, the p-dimensional vector of covariates for the ith case. Methods for recursive stratification of the data leading to binary (Cox) 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 $x_k \leq c$? 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 Cox residuals along each x_k-axis and select the one for which the residuals appear most non-random. We employ two methods introduced in Loh (1991).

3.1.1 Method R

1. Compute the Cox residuals by $e_i = \exp(\boldsymbol{x_i}\hat{\boldsymbol{\beta}})H(t,\mathbf{0})$, $i = 1, \ldots, n$.

2. Plot the cumulative hazard function

$$\hat{H}(e) = \sum_{j=1}^{l} \frac{m_i}{\#R_j} + (e - e_{(l)}) \frac{m_{l+1}}{\#R_{l+1}}$$

   against e. Here $e_{(0)} = 0$, $e_{(l)} < e \leq e_{(l+1)}$, $\{e_{(i)}\}$ are the sorted residuals, and $\#R_j$ is the number of cases in the risk set R_j.

3. Superimpose a least squares line on the cumulative hazard plot and label an observation as class 1 if its associated residual is above the line, and class 2 otherwise.

4. For each covariate, perform two-sample t-tests on the two groups of observations for differences in variances (Levene's, 1960, test).

5. Assuming that each t-statistic has a t distribution, compute a P-value for each covariate. (This assumption is used only for the purpose of ranking the covariates; the P-values are not used for inference.)

6. Suppose the ith covariate yields the smallest P-value. The data in the node are split into two parts, with one subset containing all cases with the ith 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.

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.

3.1.2 Method M

Instead of using the cumulative hazard plot, we can look for patterns in the residuals by classifying the covariate vectors into two classes according to the size of their associated residuals. A case belongs to class 1 if its residual is larger than the median of the residuals for the sample and as class 2 otherwise. As before, t tests are used to look for differences of variances between the classes for each covariate. The same procedure as in method R is followed to select the splits.

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 negative loglikelihood (see, e.g., Ciampi *et al.*, 1987) since this is reduces to mean square error in the case of normal regression.

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 negative loglikelihood. If at least one of the trees possesses a smaller negative loglikelihood 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. Briefly, the cases in the node are randomly divided into V subsets. A nested sequence of trees is constructed from the data from $(V - 1)$ subsets and the remaining subset is used as test sample to decide if the data in the former should be partitioned. This procedure is applied V times, each time leaving out a different subset as test sample. Because of variability due to cross-validation, a cross-validation tree is considered "superior" to the trivial tree if it has an estimate of negative loglikelihood that is at least $(1-f)$ times smaller than that for the trivial tree. Here f is either user-specified or estimated by the bootstrap (see below). 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 Cox 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 Cox 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 proportional hazards 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 $\hat{\alpha}(\hat{f},\hat{f})$ is closest to α.

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

3. Fixing $\eta=0.5$, choose the value of f for which $\hat{\alpha}(\hat{f},0.5)$ 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 several real and 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 proportional hazards model

In the first simulation experiment, survival times were generated from an exponential distribution with mean e^{-2x_i}, where $x_i \in \{\pm 1, \pm 2, \pm 3, \pm 4\}$. 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 100 so that about 25% 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.

Table 1: Simulation results for one proportional hazards model using the bootstrap to choose f and/or η. Nominal significance level is $\alpha = 0.05$; 25% censoring; 50 simulations.

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

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

	M method		R method	
Bootstrap method	#splits	freq.	#splits	freq.
1st ($f = \eta$)	0	36	0	20
	1	14	1	30
2nd ($f = 0$)	0	42	0	17
	1	8	1	32
			2	1
3rd ($\eta = .5$)	0	33	0	23
	1	17	1	27

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 a single Cox model, the correct trees are those with no splits. The probability of a type I error appears to be quite satisfactory.

4.1.2 Two proportional hazards models

In the second experiment, data were generated from two proportional hazards 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 R method than for the M method in every case.

4.2 Stanford heart transplant data

The Stanford heart transplant data were previously analyzed by Miller and Halpern (1982) using the Cox, Buckley-James (1979) and Miller regression methods. (See also Crowley and Hu, 1977, Kalbfleisch and Prentice, 1980, and Aitkin *et al.*, 1983.) Fitting log(survival time) on age and 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

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.

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

Table 4: Likelihood ratio, score and Wald test statistics and P-values for 157 cases

	Age only		Mismatch only		Age and mismatch	
	χ^2 (1 d.f.)	P-value	χ^2 (1 d.f.)	P-value	χ^2 (2 d.f.)	P-value
Likelihood ratio	7.43	0.0064	0.84	0.3583	8.44	0.0147
Score	6.85	0.0089	0.86	0.3544	7.85	0.0198
Wald	6.78	0.0092	0.86	0.3545	7.78	0.0205

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(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 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. Table 4 gives likelihood ratio, score and Wald test statistics and P-values for the data with 157 cases.

4.2.1 Using the M method

Applying our method with the first bootstrap estimation method gave a tree with one split, on age at 41.7 years, with a P-value of 2×10^{-5}. The second bootstrap method (with $f = 0$) produced the tree in Figure 2 which has one more split. The second split was on age again, at 48.95 years. Figure 3 shows the Kaplan-Meier estimates of the survival distributions for the data in the three terminal nodes. Regression estimates and P-values for the coefficients are given in Table 5. Only for the group of patients whose ages were greater than 48 years was age significant. Neither covariate was significant in the other two groups. The median survival times were 697, 486 and 145 days in the three nodes. The third bootstrap method holding η fixed at .5 produced a trivial tree with no splits.

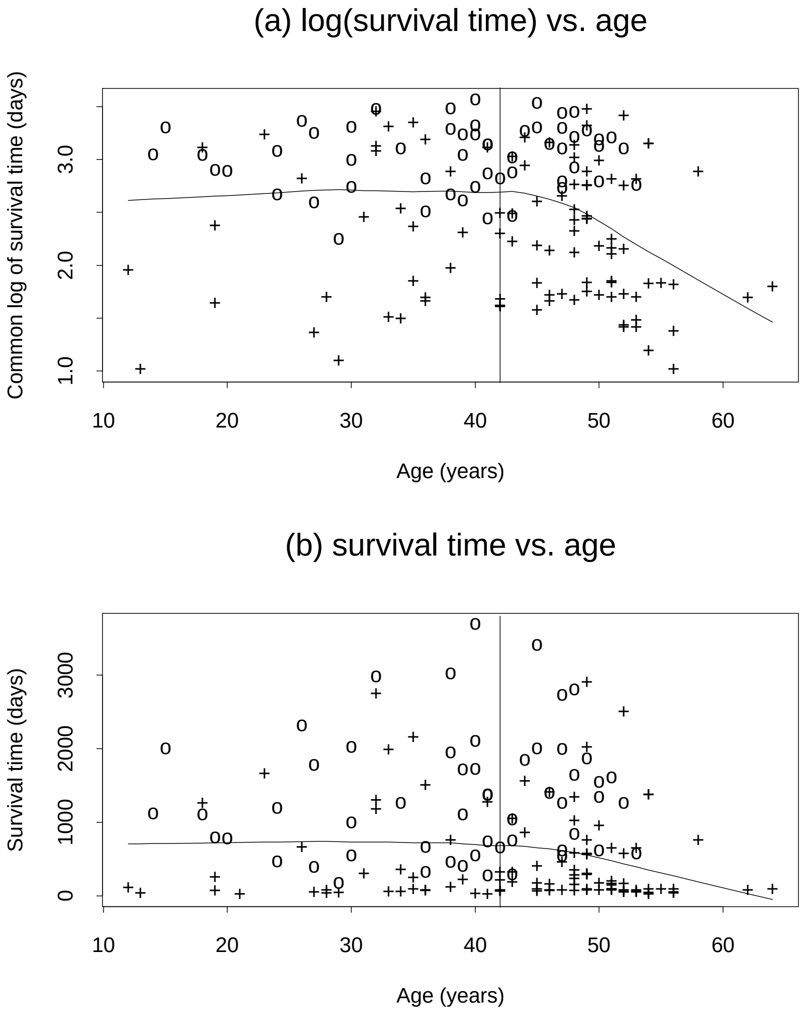

Figure 1: (a) Scatterplot of $\log_{10}$ survival time (days) versus age at transplant (years) with a smooth curve for 152 Stanford heart transplant patients who survived at least 10 days. (b) Scatterplot of survival time versus age at transplant with a smooth curve for 157 Stanford heart transplant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

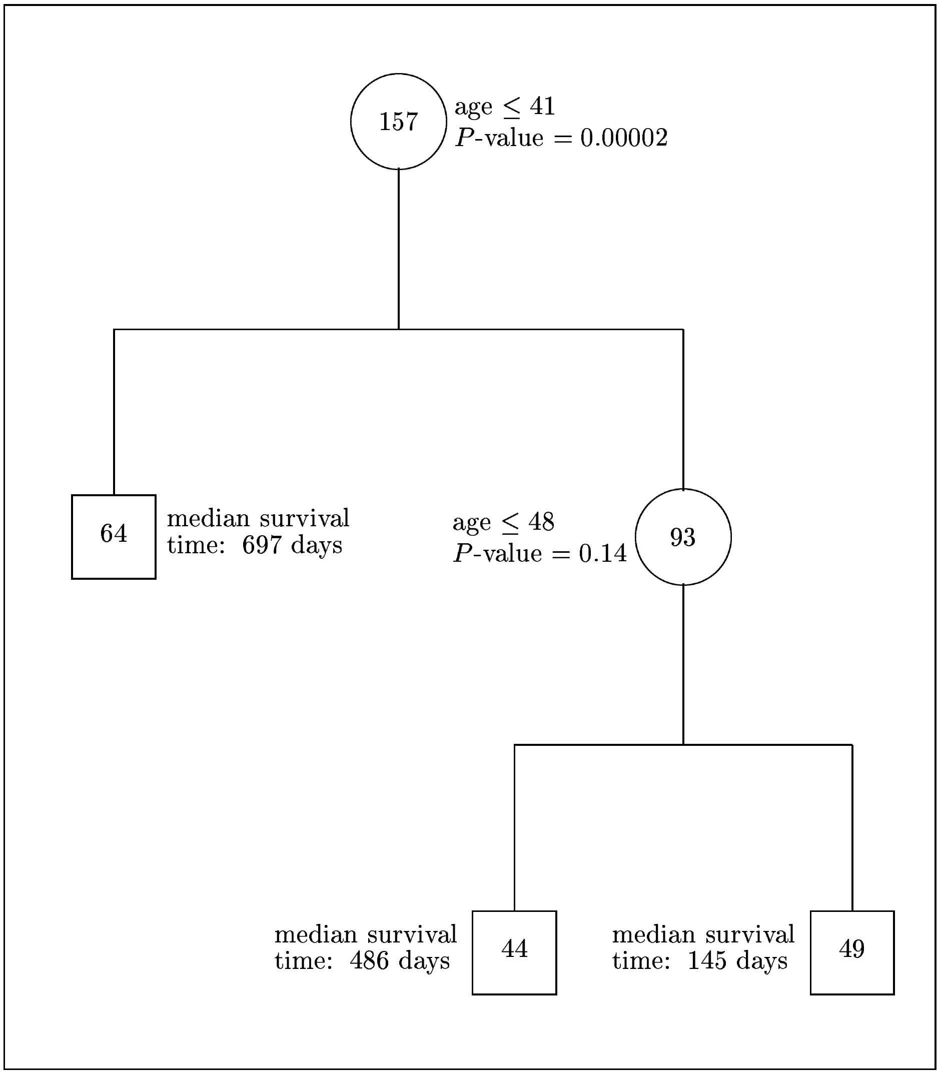

Figure 2: Cox regression tree for heart transplant data with the M-method and $f = 0$, $\eta = .95$. The numbers within circles or squares are sample sizes. The P-value beside each node refers to Levene's test.

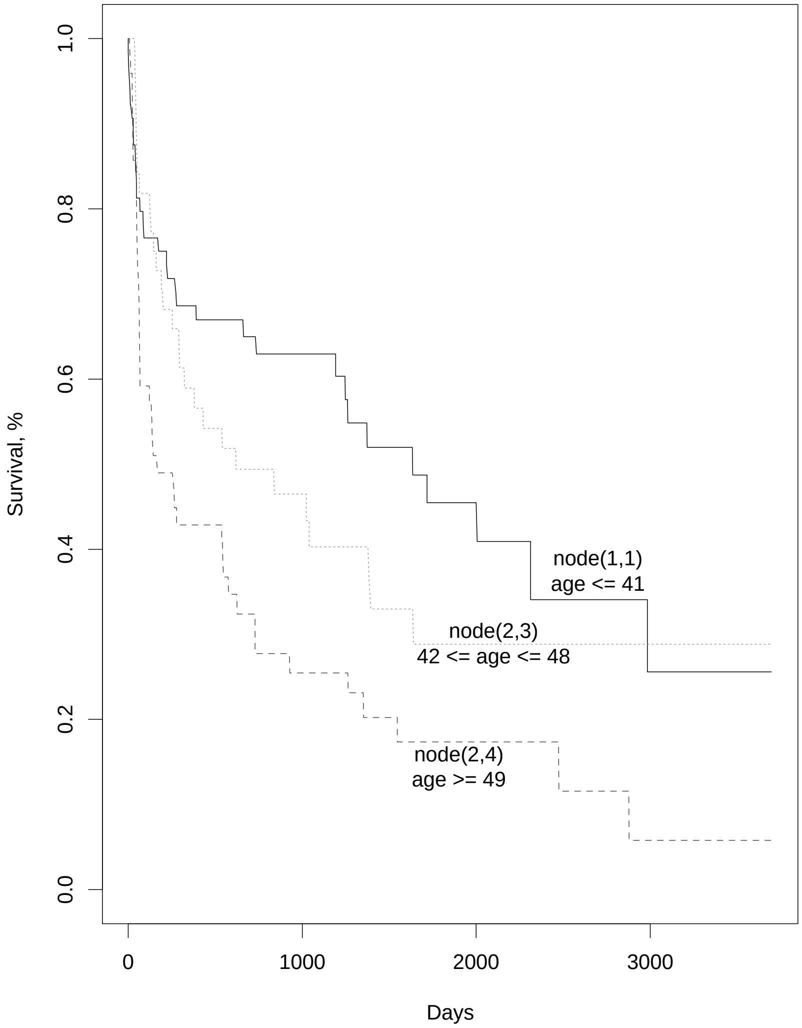

Figure 3: Kaplan-Meier survival curves for the terminal nodes of the tree in Figure 2.

Table 5: Regression estimates and P-values of likelihood ratio, score and Wald tests at the terminal nodes of the tree in Figure 2.

	Variable	$\hat{\beta}$	S.E.	$\hat{\beta}$/ S.E.	P-value		
Node					Likelihood ratio	Score	Wald
age ≤ 41	age	-0.026	0.022	-1.20	0.238	0.226	0.229
	mismatch	-0.333	0.339	-0.98	0.319	0.325	0.326
	both				0.303	0.297	0.300
$41 < \text{age} \leq 48$	age	-0.090	0.094	-0.95	0.343	0.339	0.341
	mismatch	0.215	0.431	0.50	0.617	0.617	0.617
	both				0.526	0.527	0.530
age > 48	age	0.134	0.048	2.76	0.011	0.005	0.006
	mismatch	0.488	0.261	1.87	0.069	0.059	0.061
	both				0.007	0.005	0.006

Table 6: Regression estimates and P-values of likelihood ratio, score and Wald tests for the two lowest terminal nodes of the tree in Figure 4.

	Variable	$\hat{\beta}$	S.E.	$\hat{\beta}$/ S.E.	P-value		
Node					Likelihood ratio	Score	Wald
$41 < \text{age} \leq 48$	age	-0.154	0.144	-1.07	0.295	0.280	0.287
&	mismatch	-2.894	1.213	-2.39	0.013	0.010	0.017
mismatch ≤ 1.06	both				0.040	0.037	0.057
$41 < \text{age} \leq 48$	age	-0.102	0.126	-0.81	0.414	0.414	0.416
&	mismatch	0.586	1.073	0.55	0.587	0.585	0.585
mismatch > 1.06	both				0.579	0.585	0.588

4.2.2 Using the R method

The R method gave trivial trees for the first and third bootstrap methods. The second bootstrap method with $f = 0$ gave the tree in Figure 4. It has one more split (on mismatch score) than that for the M-method. Figure 5 shows the Kaplan-Meier estimates of the survival distributions for the terminal nodes and Table 6 gives the regression estimates. Mismatch score was significant in the left node but not the right. Age was not significant in either node. Median survival time is seen to be shorter for the patients with larger mismatch scores.

4.3 Mayo liver transplant data

The next example is based on data on liver transplantation performed at the Mayo Clinic. The data are described in Dickson *et al.* (1989) and Markus *et al.* (1989) and published in Fleming and Harrington (1991, Appendix D.1). There were 418 patients for each of whom the survival time and censoring indicator (1 = death, 0 = censored) was known. Dickson *et al.* (1989) selected five variables: (i) age in years, (ii) albumin (in mg/dl), (iii) serum bilirubin (in mg/dl), (iv) presence of edema (0 = no edema and no diuretic therapy for edema; 0.5 = edema present for which no diuretic therapy was given, or edema resolved with diuretic therapy; 1 = edema despite diuretic therapy), and (v) prothrombin time (in seconds, abbreviated to “protime” here).

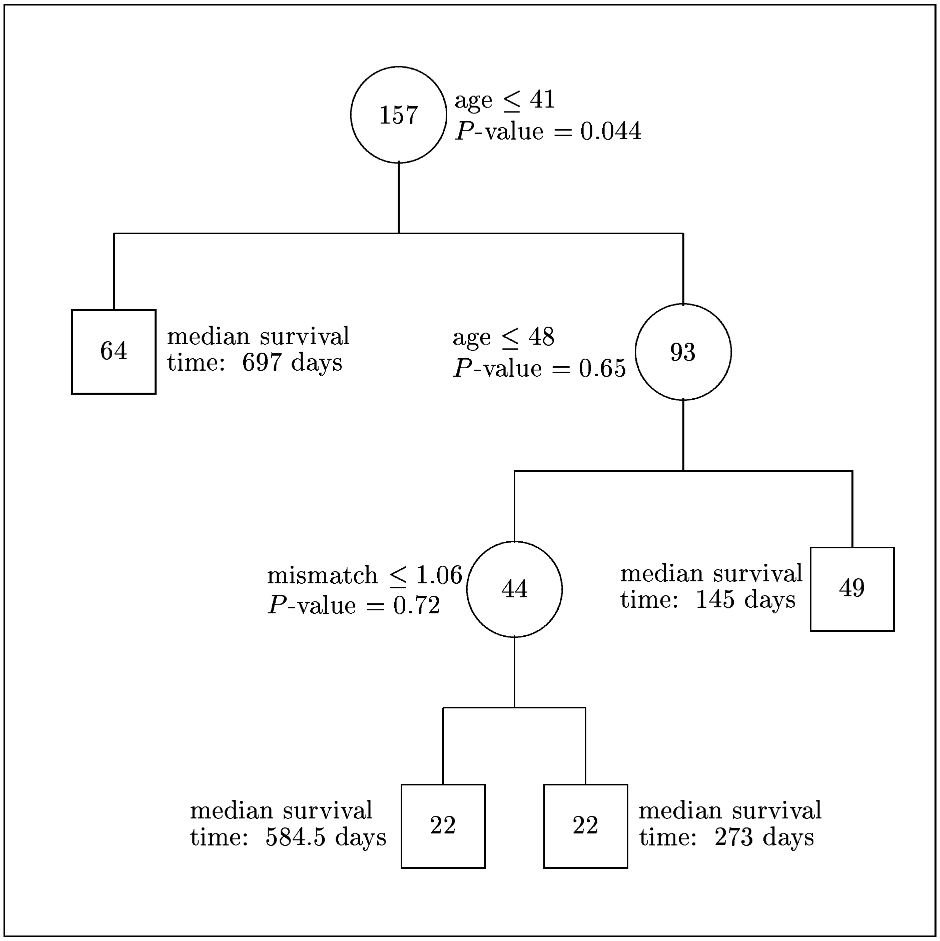

Figure 4: Cox regression tree with the second bootstrap and R methods for the heart transplant data. The numbers within circles or squares are sample sizes. The P-value beside each node refers to Levene's test.

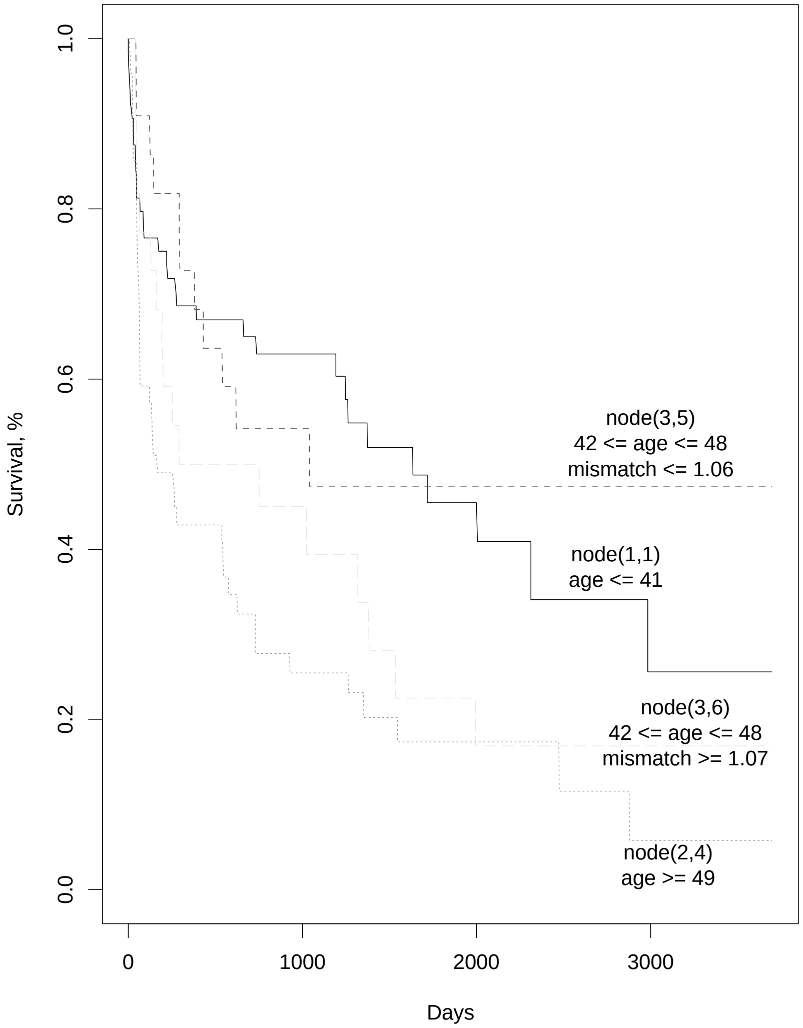

Figure 5: Kaplan-Meier survival curves for the terminal nodes of the tree in Figure 4.

Table 7: Regression estimates of the coefficients and the P-values of likelihood ratio, score and Wald tests for the liver transplant data.

Variable	$\hat{\beta}$	S.E.	$\hat{\beta}$/S.E.	P-value		
				Likelihood ratio	Score	Wald
age	0.039	0.008	5.15	< 0.0001	< 0.0001	< 0.0001
log(albumin)	-2.533	0.648	-3.91	0.0002	0.0001	0.0001
log(bilirubin)	0.871	0.083	10.54	< 0.0001	< 0.0001	< 0.0001
edema	0.859	0.271	3.17	0.0022	0.0013	0.0015
log(protime)	2.380	0.767	3.10	0.0037	0.0020	0.0019

Dickson *et al.* (1989) fitted a Cox proportional hazards model with covariates age, log(albumin), log(bilirubin), edema and log(protime). The regression estimates of the coefficients and other summary statistics are given in Table 7. Using martingale residuals, Lin, Wei and Ying (1992) found that even though a log transformation of bilirubin was suggested, the resulting model was still not entirely satisfactory.

We report here only the result of applying our procedure with the R method. The bootstrap with $f = 0$ gave the tree in Figure 6. The sample was first split at edema = 0.1 with a P-value less than 10^{-5}. Figure 7 shows the Kaplan-Meier estimates of the survival times for the four terminal nodes. Regression estimates and P-values are given in Table 8.

The covariate log(bilirubin) was significant in every terminal node of the tree at level 0.05. Survival times tended to be shorter for the patients with larger log(bilirubin) values. Age affected the survival time significantly at level 0.05 except in one node. The covariate log(albumin) was significant at level .01 in only one node. Survival times were shorter for the patients with smaller log(albumin) values in that node. The other bootstrap methods gave slightly larger trees while the M method gave trees with equal or fewer numbers of terminal nodes.

5 Conclusion

This research was motivated by the twin goals of (i) providing a formal goodness-of-fit test for the proportional hazards 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 proportional hazards 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 combination of the R method and bootstrapping with $f = 0$, however, seems to yield the best power.

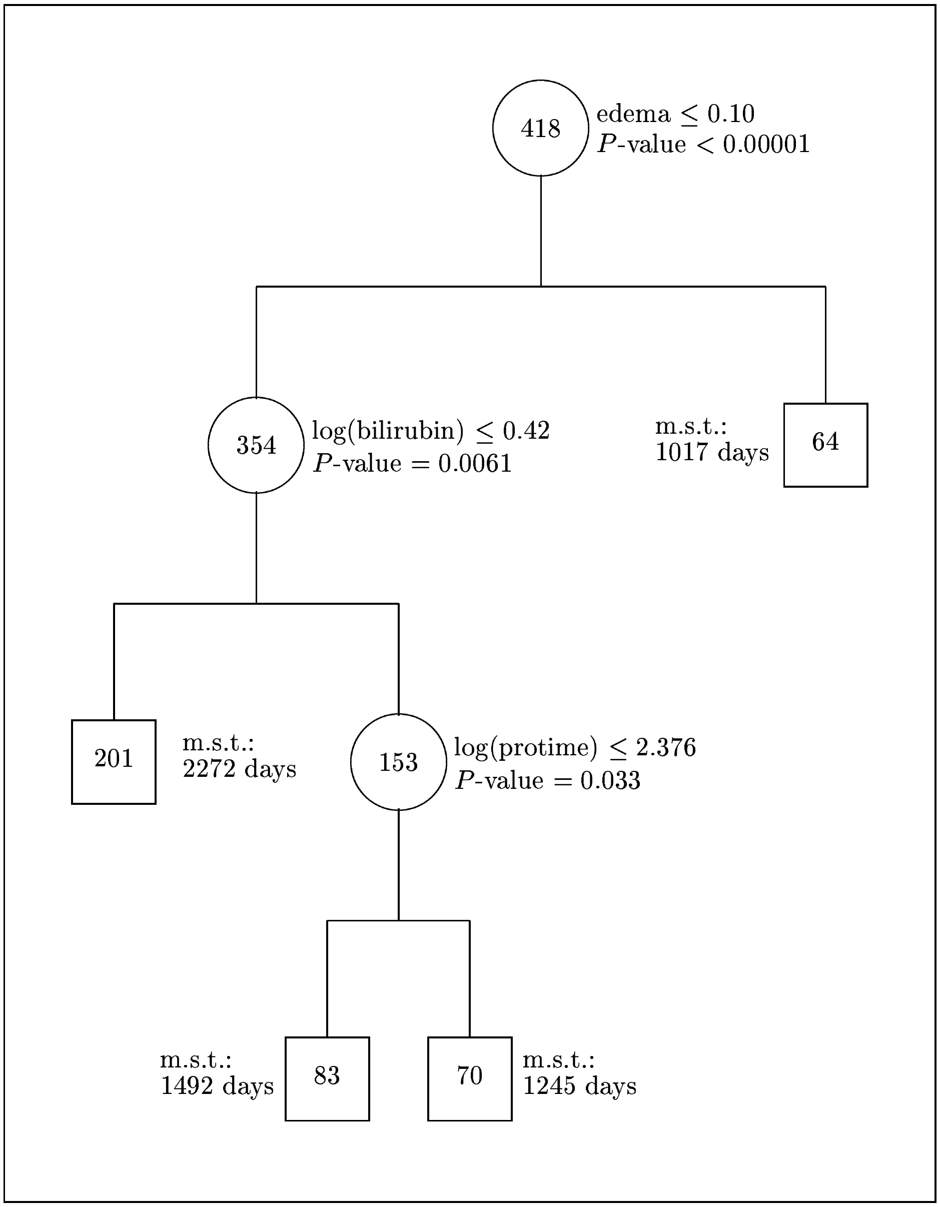

Figure 6: Cox regression tree from the R method for the liver transplant data. The value of f was fixed at 0 and η estimated by the bootstrap to be .95. M.s.t. stands for median survival time. The P-value beside each node refers to Levene's test.

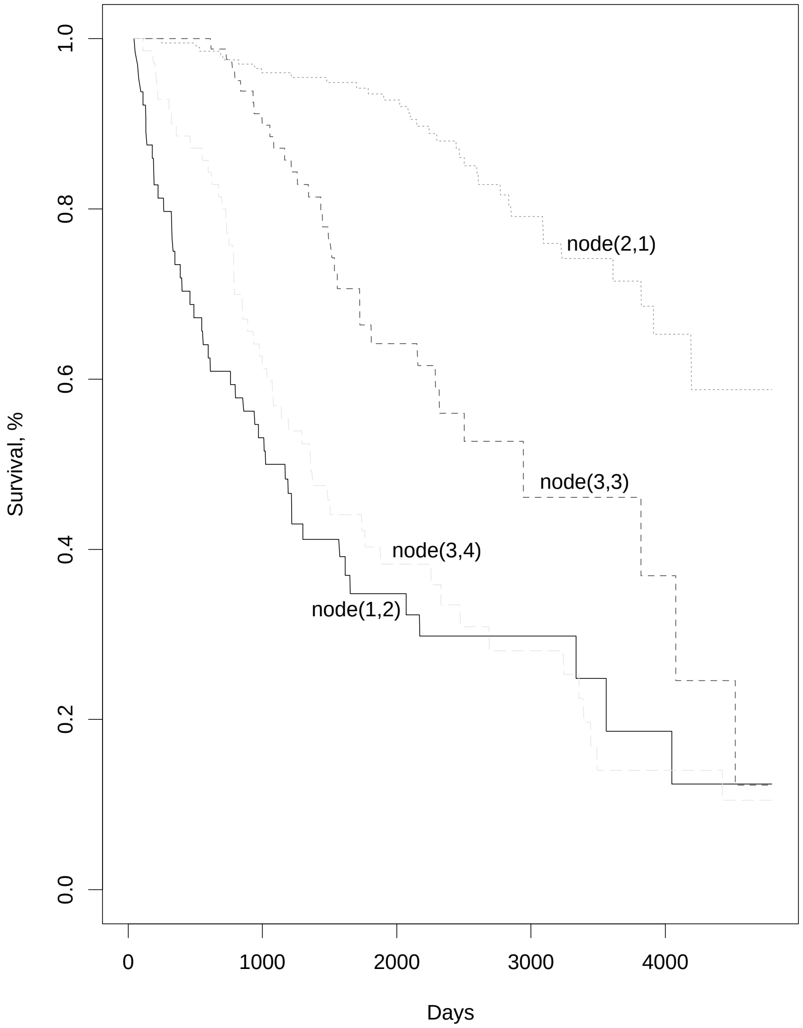

Figure 7: Estimated (Kaplan-Meier) survival curves for the four terminal nodes of the tree in Figure 6.

Table 8: Regression estimates of coefficients and P-values for the liver data tree in Figure 6.

Node	Variable	$\hat{\beta}$	S.E.	$\hat{\beta}$/S.E.	P-value		
					like. ratio	score	Wald
Edema $> .1$	age	0.052	0.018	2.83	0.0031	0.0035	0.0046
	log(albumin)	-1.233	1.351	-0.91	0.3666	0.3581	0.3589
	log(bilirubin)	0.692	0.185	3.75	0.0001	0.0001	0.0002
	edema	1.326	0.794	1.67	0.0949	0.0921	0.0952
	log(protime)	3.161	1.986	1.59	0.1101	0.1090	0.1116
Edema $\leq .1$ & log(bilirubin) $\leq .42$	age	0.056	0.022	2.59	0.0076	0.0087	0.0096
	log(albumin)	-2.525	1.493	-1.69	0.1047	0.0904	0.0911
	log(bilirubin)	1.057	0.533	1.98	0.0411	0.0466	0.0473
	log(protime)	0.331	2.001	0.17	0.8679	0.8666	0.8666
Edema $\leq .1$, log(bilirubin) $> .42$ & log(protime) ≤ 2.376	age	0.059	0.022	2.68	0.0076	0.0065	0.0075
	log(albumin)	-0.911	1.598	-0.57	0.5781	0.5692	0.5687
	log(bilirubin)	1.104	0.285	3.87	0.0002	< 0.0001	0.0001
	log(protime)	1.635	5.083	0.32	0.7471	0.7477	0.7477
Edema $\leq .1$, log(bilirubin) $> .42$ & log(protime) > 2.376	age	0.016	0.012	1.37	0.1742	0.1699	0.1713
	log(albumin)	-4.978	1.328	-3.75	0.0002	0.0001	0.0002
	log(bilirubin)	0.602	0.214	2.81	0.0058	0.0044	0.0050
	log(protime)	0.030	1.542	0.019	0.9847	0.9847	0.9847

References

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.

Breslow, N. (1974). "Covariance analysis of censored survival data". *Biometrics,* **30**, 89-99.

Buckley, J. and James, I. (1979). "Linear regression with censored data". *Biometrika,* **66**, 429-436.

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". *Journal of the Royal Statistical Society B,* **34**, 187-202.

Cox, D. R. (1975). "Partial likelihood". *Biometrika,* **62**, 269-276.

Crowley, J. and Hu, M. (1977). "Covariance analysis of heart transplant survival data". *Journal of the American Statistical Association,* **72**, 27-36.

Dickson, E. R., Grambsch, P. M., Fleming, T. R., Fisher, L. D. and Langworthy, A. (1989). "Prognosis in Primary Biliary Cirrhosis: Model for Decision Making". *The New England Journal of Medicine,* **10**, 1-7.

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.

Fleming, T. R., and Harrington, D. P. (1991). *Counting Processes and Survival Analysis.* Wiley, New York.

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.

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.

Lin, D. Y. and Wei, L. J. (1991). "Goodness-of-fit tests for the general Cox regression model". *Statistica Sinica,* **1**, 1-17.

Lin, D. Y., Wei, L. J. and Ying, Z. (1992). "Checking the Cox model with cumulative sums of martingale-based residuals". *Technical Report No. 111.* Department of Biostatistics, University of Washington.

Link, C. L. (1979). "Confidence intervals for the survival function using Cox's proportional hazard model with covariates". *Technical Report No. 45, Division of Biostatistics, Stanford University.*

Loh, W.-Y. (1991). "Survival modeling through recursive stratification". *Computational Statistics and Data Analysis.* **12**, 295-313.

Markus, B. H., Dickson, E. R., Grambsch, P. M., Fleming, T. R., Mazzaferro, V., Klintmalm, G. B. G., Weisner, R. H., Van Thiel, D. H. and Starzl, T. E. (1989). "Efficacy of liver transplantation in patients with primary biliary cirrhosis". *The New England Journal of Medicine,* **320**, 1709-1713.

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.

Peto, R. (1972). "Discussions on Cox's paper" *Journal of the Royal Statistical Society B*, **34**, 205-207.

Segal, M. R. (1988). "Regression trees for censored data". *Biometrics,* **44**, 35-47.

Wei, L. J. (1984). "Testing goodness of fit for proportional hazards model with censored observations". *Journal of the American Statistical Association,* **79**, 649-652.

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.

Appendix: Outline of the algorithm

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

- $\mathbf{y}$: response vector
- X: matrix of the covariates
- $\mathbf{d}$: vector of censoring indicators
- *mindat*: user-specified terminal node sample size
- *tol*: maximum error in Newton-Raphson iterations
- $flag$: diagnostic flag for Cox regression. Equals 1 if the Newton-Raphson method of Cox 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.
- node(i, j): jth node from the left at the ith level. The root node is node$(0, 1)$.
- $pos(i, j, k)$: position of the subsample in node(i, j) among the whole sample. $pos(i, j, 1)$ is the initial position, $pos(i, j, 2)$ is the end position of the subsample.
- *count*: the number of Cox 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 Cox regression tree algorithm is as follows:

1. Read data $\mathbf{y}, X$ and $\mathbf{d}$.
2. Choose splitting method (R or M).
3. Initialize the values of $tol, f, \eta, mindat$ and set $flag = 0$.
4. If a covariate contains categorical variables, then indicator variables are generated for the levels of the variable.
5. Loop over levels $i = 0, 1, \ldots$.
    - $count = 0$.
        - Loop over nodes $j = 1, \ldots, 2^i$ (ith 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 Cox regression model. If $flag = 1$ or $flag = 2$, set the node directly above as a terminal and exit.
            (d) Get the Cox residuals, the cumulative hazard function of the residuals and fit a least-squares line to the points. Also, evaluate the residuals from the linear fit.

        (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.

6. 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 $\mathcal{L}_1,\ldots,\mathcal{L}_V$ and the following process repeated for $v=1,\ldots,V$.

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 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 a Cox 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 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 $0=p_0<p_1<\ldots<p_s<p_{s+1}=1$.

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

3. Prune T_{v0} 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 $T_{s+1}=T_{v0}$.

    - Loop over $k=s,s-1,\ldots,1,0$.
        - Starting from the lowest level to the root node of T_{k+1} (loop over levels $i=a,\ldots,0$, where a is the lowest level of T_{k+1}), do the following.
            (a) At level i, loop over $j=1,\ldots,2^i$, 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 mean square error.

    Set $\theta(v) = 0$.

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

    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 Cox 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 $\hat{\beta}^0 = (0, \ldots, 0)'$ as the initial value of the regression parameters in the iteration. If the Newton-Raphson method does not converge within 20 iterations, the program stops. Further, 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^*. Let $\hat{S}(t|\mathbf{x}) = \hat{S}_0(t)^{\exp(x\beta)}$ be the estimated model from the observed data. To generate a bootstrap observation T^* with survival function $\hat{S}(t|\mathbf{x})$, perform the following steps.

    - Generate random variables $U_i \sim \text{Uniform}(0,1)$, $i = 1, \ldots, n$.
    - Let T_i^* be defined by $\hat{S}(T_i^*|\mathbf{x}) = U_i$. This implies $\hat{S}_0(T_i^*)^{\exp(x_i\beta)} = U_i$ and hence

    $$T_i^* = \max\{t : \hat{S}_0(t) \geq U_i^{1/\exp(x_i\hat{\beta})},\ t \in \{y_1, \cdots, y_n\}\}.$$

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 ith 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 $(t_1^*, \cdots, t_n^*)$ and $(c_1^*, \cdots, c_n^*)$ be the bootstrap survival and censoring times. Compute the bootstrap observations $(y_1^*, d_1^*), \cdots, (y_n^*, d_n^*)$, 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.