Consistency of tree-based estimators in censored regression with applications in insurance

Xavier Milhaud [0,1,2] & Olivier Lopez [0,1,2,3] & Pierre Thérond [0,4]

[0] *xavier.milhaud@ensae.fr ; olivier.lopez@ensae.fr ; pierre.therond@univ-lyon1.fr*
[1] *Centre de Recherche en Economie et Statistique (LFA lab)*
[2] *Ecole Nationale de la Statistique et de l'Administration Economique*
[3] *Sorbonne Universités, UPMC Université Paris VI, EA 3124, LSTA*
[4] *Institut de Science Financière et d'Assurances, Université Lyon 1*

Résumé. Les arbres de régression et de classification sont devenus très populaires dans les trente dernières années. L'application historique de cette technique concerne l'estimation non-paramétrique d'une espérance conditionnelle, en fonction de certains facteurs de risque représentés par des covariables. Nous adaptons ici cette méthode au cas de données de survie, pour lesquelles la problématique de censure des données doit être traitée. Les propriétés de ces estimateurs par morceaux sont étudiées, et des résultats théoriques permettent de conclure sur la vitesse de convergence de tels estimateurs. Ces résultats sont ensuite validés par une étude simulatoire, puis deux applications sur données réelles en assurance sont proposées afin d'illustrer l'intérêt de la méthode.

Mots-clés. poids Kaplan-Meier, censure, arbre de régression.

Abstract. The use of regression trees as a tool for high-dimensional classification and regression problems has boomed in the last thirty years. Initially designed to estimate non-parametrically the conditional mean of a response given some covariates, this popular technique is here adapted to deal with survival data. We derive key non-asymptotic results and almost sure convergence rates for tree-based estimators provided by the growing step, as well as convergence properties of the associated selection process. Our theoretical results are confirmed by a simulation study and two applications on real-life datasets to illustrate the utility of such a method in practice.

Keywords. Kaplan-Meier weights, censored observations, regression tree.

1 A weighted CART algorithm

In numerous applications of survival analysis, analyzing the heterogeneity of a population is a key issue. For example, in insurance, a strategic question is to determine clusters of individuals which represent different levels of risk. Once such groups have been identified, it becomes possible to improve pricing, reserving or marketing targeting. We show here how to adapt CART methodology (Classification And Regression Trees, Breiman et al (1984)) to a survival analysis context, with such applications in perspective. The main question we face is the presence of censoring that may affect some duration variable, and the necessity to correct the bias it introduces in the statistical methods. Such duration variables naturally appear in the situations we consider, either because we are studying lifetimes or because they are related to the time before a claim is fully settled.

In the sequel, we consider a duration variable $T \in \mathbb{R}^+$, a censoring variable $C \in \mathbb{R}^+$, and a random vector $M \in \mathbb{R}^k$, and define the observations

$$\begin{aligned} Y &= \inf(T, C), \\ \delta &= \mathbf{1}_{T \leq C}, \\ N &= \delta M. \end{aligned}$$

Moreover, let $\mathbf{X} \in \mathcal{X} \subset \mathbb{R}^d$ denote a set of random covariates that may have impact on T and/or M. The observations that we consider in the following are the i.i.d. replications $(N_i, Y_i, \delta_i, \mathbf{X}_i)_{1 \leq i \leq n}$, where the variables M_i correspond to quantities that are observed only when the individual i is fully observed. The classical censored regression framework can be seen as a special case, taking $k = 1$ and $M = T$.

Our aim is to understand the impact of $\mathbf{X}$, and possibly T, on M. More precisely, we wish to estimate a function

$$\pi_0 = \arg\min_{\pi \in \mathcal{P}} E\left[\phi(M, \pi(\mathbf{X}, T))\right],$$

where $\mathcal{P}$ is a subset of an appropriate functional space and ϕ a loss function. Under appropriate assumptions, we know that

$$\int \psi(m, t, \mathbf{x})\, d\hat{F}(m, t, \mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \frac{\delta_i \psi(N_i, Y_i, \mathbf{X}_i)}{1 - \hat{G}(Y_i -)}, \tag{1}$$

is a consistent estimator of $E[\psi(M, T, \mathbf{X})]$, with $\hat{G}$ being a Kaplan-Meier estimator of the censoring distribution. Our approach is thus based on the IPCW method, see van der Laan and Robins (2003) chapter 3.3. It consists in determining a weighting scheme that compensates the lack of complete observations in the sample.

We first explain how the CART algorithm can be adapted to the presence of censoring. The building procedure of a regression tree is based on the definition of a *splitting criterion* that furnishes partition rules at each step of the algorithm. More precisely, at each step s, a tree with L_s leaves is constituted. These terminal nodes represent disjoint subpopulations of the initial n observed individuals. The rules used to create these populations are based on the values of T and $\mathbf{X}$. That is, the leaves provide us with a partition of the space $\mathcal{T} = \mathbb{R}^+ \times \mathcal{X}$ into L_s disjoint sets $\mathcal{T}_1^{(s)}, ..., \mathcal{T}_{L_s}^{(s)}$. The individual i belongs to the subpopulation of the leaf l if $\tilde{\mathbf{X}}_i := (T_i, \mathbf{X}_i) \in \mathcal{T}_l^{(s)}$.

At step $s+1$, each leaf is likely to become a new node of the tree by making use of the splitting criterion. Let $\tilde{X}^{(j)}$ denote the $j-$th component of $\tilde{X}$. In absence of censoring, to partition the subpopulation of the $l-$th leaf into two subpopulations, one determines, for each component $\tilde{X}^{(j)}$, the threshold $x_l^{(j)}$ that minimizes

$$\min_{\pi\in\mathbb{R}}\left\{\frac{\int \phi(m,\pi)\mathbf{1}_{\tilde{\mathbf{x}}\in\mathcal{T}_l^{(s)}}\mathbf{1}_{\tilde{x}^{(j)}\leq x_l^{(j)}}d\hat{F}_n(m,t,\mathbf{x})}{\int \mathbf{1}_{\tilde{\mathbf{x}}\in\mathcal{T}_l^{(s)}}\mathbf{1}_{\tilde{x}^{(j)}\leq x_l^{(j)}}d\hat{F}_n(m,t,\mathbf{x})}\right.$$
$$\left.+\frac{\int \phi(m,\pi)\mathbf{1}_{\tilde{\mathbf{x}}\in\mathcal{T}_l^{(s)}}\mathbf{1}_{\tilde{x}^{(j)}> x_l^{(j)}}d\hat{F}_n(m,t,\mathbf{x})}{\int \mathbf{1}_{\tilde{\mathbf{x}}\in\mathcal{T}_l^{(s)}}\mathbf{1}_{\tilde{x}^{(j)}> x_l^{(j)}}d\hat{F}_n(m,t,\mathbf{x})}\right\} =: L_l(j,x_l^{(j)}), \tag{2}$$

where $\hat{F}_n$ denotes the empirical distribution of $(M,T,\mathbf{X})$. Then one determines $j_0 = \arg\min_j L_l(j,x_l^{(j)})$. Next, the partition of the population of the $l-$th leaf is performed by separating the individuals having $\tilde{X}_i^{(j_0)} \leq x_l^{(j_0)}$, and those such that $\tilde{X}_i^{(j_0)} > x_l^{(j_0)}$. Here, the empirical distribution function $\hat{F}_n$ is unavailable but the idea is to replace $\hat{F}_n$ in (2) thanks to (1). To build the maximal tree in practice, the CART algorithm thus becomes:

Step 0: compute the estimator $\hat{G}$ from the dataset with n individuals.

Step 1: initialization. Consider the tree with only one leaf ($L_1 = 1$), corresponding to the whole population. Set $\mathcal{T}_1^{(1)} = \mathcal{T}$.

Step s: splitting. Consider the tree obtained at step $s-1$, with L_{s-1} leaves. Each leaf l corresponds to a set $T_l^{(s-1)}$ such that $\mathcal{T}_l^{(s-1)} \cap \mathcal{T}_{l'}^{(s-1)} = \emptyset$ and $\cup_l \mathcal{T}_l^{(s-1)} = \mathcal{T}$. The uncensored observations (denote by e_l their number) such that $\tilde{\mathbf{X}} \in \mathcal{T}_l^{(s-1)}$ are assigned to leaf l. For each leaf l, with $1 \leq l \leq L_{s-1}$:

s.1 if $e_l = 1$ or if all observations have the same values of $\tilde{\mathbf{X}}$, do not split;

s.2 else, determine j_0 and $x_l^{(j_0)}$ that minimizes $L_l(j,x_l^{(j)})$ given in (2) and define $\mathcal{L}_l = \mathcal{T}_l^{(s-1)} \cap \{\tilde{X}^{(j_0)} \leq x_l^{(j_0)}\}$, and $\mathcal{U}_l = \mathcal{T}_l^{(s-1)} \cap \{\tilde{X}^{(j_0)} > x_l^{(j_0)}\}$.

Define a collection of disjoints sets $\mathcal{T}_{l'}^{(s)}$ which consists of the sets $\mathcal{L}_l, \mathcal{U}_l$ for $1 \leq l \leq L_{s-1}$ (or $\mathcal{T}_l^{(s-1)}$ if the $l-$th leaf satisfied the condition s.1). Set L_s the new number of leaves. Go to step $s+1$, unless $L_s = L_{s-1}$.

At the end of this process, we thus get the piecewise constant tree-based estimator

$$\hat{\pi}^K(\mathbf{x},t) = \sum_{l=1}^{K} \hat{\gamma}_l\, R_l(t,\mathbf{x}),$$

where K denote the total number of leaves, with each leaf l associated to a set $\mathcal{T}_l$ and a rule $R_l(\tilde{\mathbf{x}}) = \mathbf{1}_{\tilde{\mathbf{x}}\in\mathcal{T}_l}$, and

$$\hat{\gamma}_l = \arg\min_{\pi\in\mathbb{R}} \frac{\int \phi(m,\pi)\, R_l(\tilde{\mathbf{x}})\, d\hat{F}(m,t,\mathbf{x})}{\int R_l(\tilde{\mathbf{x}})\, d\hat{F}(m,t,\mathbf{x})}.$$

The coefficient $\hat{\gamma}_l$ can be seen as an estimator of $\gamma_l = \arg\min_{\pi \in \mathbb{R}} E[\phi(M, \pi) \,|\, \tilde{\mathbf{X}} \in \mathcal{T}_l]$.

In the rest of the paper, we show the consistency of using such a procedure for building the estimator of π_0. A part of the paper is dedicated to the pruning step, and the study of the penalization to be applied to this estimator in order to integrate the model complexity.

Finally, a simulation study as well as two real-life applications are proposed. The former allows us to confirm the convergence of the estimator towards to quantity of interest, illustrated by Table 1. In this simulation study, a mixture of four components is artificially generated and the goal is to see whether the regression tree uncovers this structure. In the real-life examples, we first compare the performance of Cox predictions and tree-based ones for income protection insurance purposes, and check that the tree-based estimators globally gives better results than the other one. Lastly, another example from medical cares shows how to adapt the regression tree method to the prediction of unobserved claim amounts, which is the special scheme that was described in introduction. This particular situation requires to fit two different regression trees in order to make the ratio of their predictions.

Bibliographie

[1] Breiman, L. and Friedman, J. and Olshen, R. A. and Stone, C. J. (1984), *Classification and Regression Trees*, Chapman and Hall.

[2] van der Laan, Mark J. and Robins, James M. (2003), *Unified methods for censored longitudinal data and causality*, Springer Series in Statistics, New-York.

% of	Sample	Group-specific MWSE				Global
censored	size	Group 1	Group 2	Group 3	Group 4	MWSE
observations	n	MWSE	MWSE	MWSE	MWSE	
10%	100	0.19516	0.42008	0.17937	0.30992	***1.10454***
	500	0.03058	0.07523	0.03183	0.06029	***0.19796***
	1 000	0.01509	0.03650	0.01517	0.02619	***0.09306***
	5 000	0.00295	0.00714	0.00289	0.00530	***0.01804***
	10 000	0.00105	0.00378	0.00117	0.00292	***0.00910***
30%	100	0.20060	0.43664	0.17448	0.29022	***1.10765***
	500	0.03736	0.07604	0.04301	0.06584	***0.22217***
	1 000	0.01748	0.04095	0.01535	0.02674	***0.10043***
	5 000	0.00319	0.00758	0.00291	0.00547	***0.01904***
	10 000	0.00117	0.00372	0.00125	0.00292	***0.00930***
50%	100	0.19784	0.45945	0.17387	0.28363	***1.11476***
	500	0.04906	0.08993	0.05301	0.06466	***0.25668***
	1 000	0.02481	0.05115	0.01788	0.03004	***0.12387***
	5 000	0.00520	0.00867	0.00389	0.00516	***0.02299***
	10 000	0.00153	0.00407	0.00162	0.00308	***0.01057***

Table 1: Mean weighted squared errors depending on the censoring rate and sample size.