A tree-based algorithm for individual reserving, with reporting delays and long developments

7th SMIF Conference (03/02/2020)
Work conducted within the Research Chair DIALog under the aegis of the Risk Foundation, a joint initiative by CNP Assurances and ISFA (Universite de Lyon)

Xavier MILHAUD
ISFA, University of Lyon (Lyon)

Joint work with O. Lopez (Sorbonne University)
AIM : ESTIMATE NONPARAMETRICALLY THE COST OF RBNS CLAIMS

Estimate some individual claim amount $M$

given features

- $T \in \mathbb{R}^+$ : duration of the claim,
- $X \in \mathbb{R}^d$ : its characteristics.

However, for RBNS claims, we only observe the follow-up time $Y$ (censored) and features $X$. 

Next
**Main idea 1**: nonparametric models allows for flexibility.

→ Free relationship b/w response & risk factors.

**Main idea 2**: claim lifetime plays a key role to explain claim cost.

→ Well-known from claim handling experts, with positive correlation between duration and amount.

**Main idea 3**: significant impact of reporting on final claim amount.

→ And thus on reserves...

⇒ **Two last ideas are extensions** to [?]
DATA AT-HAND

We observe a sample of i.i.d. random variables \((Y_i, N_i, \delta_i, X_i)_{1 \leq i \leq n}\) with same distribution \((Y, N, \delta, X)\), where

\[
\begin{align*}
Y &= \min(T, C), \\
\delta &= 1_{T \leq C}, \\
N &= \delta M, \\
X &= \text{the vector of individual characteristics.}
\end{align*}
\]

\(C\) : censoring variable, coming from the censoring mechanism.

\(\Rightarrow C\) impacts both \(T\) and \(M\) at same time.
REPORTING DELAY AND LEFT-TRUNCATION

In the context of reserving, the *reporting delay* $\tau$ can sometimes be large, leading to *unknown* claims.

$\Rightarrow$ **The claim is observed conditionally to** $C \geq \tau$.

Usually, the phenomenon is truncated when $T \leq \tau$...

Here, $T \leq \tau$ means that claim was closed before being reported, but appears in the database!
REMIND OUR GOAL

We seek the best prediction for $M$ related to still open claims (from available data):

$$M^* = E \left[ M \mid \delta = 0, y, \tau, x \right],$$

Or equivalently

$$M^* = E \left[ M \mid T \geq y, \tau, x \right].$$
REGRESSION TREES (COMPLETE observations)

\[ \pi_0(x) = E_0[T \mid X = x] \]  \hspace{1cm} (1)

→ Most famous: linear relationship b/w \( T \) and \( X \) (restrictive class).

→ General solution: OLS, solve

\[ \pi_0(x) = \arg \min_{\pi(x)} E_0[\phi(T, \pi(x)) \mid X = x] \]  \hspace{1cm} (2)

where \( \phi(T, \pi(x)) = (T - \pi(x))^2 \).

→ CART: recursive partitioning of covariate space ⇒ minimizes intra-node variances at each step, maximum homogeneity on \( T \) following the segmentation rule ⇒ piecewise-constant estimator!
EXAMPLE OF TREE, CLASSIFICATION
MORTALITY [Olbricht, 2012])

SwissRe portfolio: ≃ 1.5M indiv. observed over 4y (gender, age).

\[
\begin{align*}
\text{AGE} < 56 \\
\text{AGE} < 46 \\
\text{AGE} < 43 &\quad \text{SEX} = \text{female} \\
\text{AGE} < 63 &\quad \text{SEX} = \text{female} \\
\text{AGE} < 63 &\quad \text{AGE} < 61 \\
\text{AGE} < 59 &\quad \text{AGE} < 62
\end{align*}
\]

\begin{align*}
286,298 &\quad 77,812 &\quad 78,792 &\quad 163,197 &\quad 32,293 &\quad 7,315 \\
0.479 &\quad 1.234 &\quad 1.498 &\quad 2.488 &\quad 2.849 &\quad 5.058
\end{align*}

\begin{align*}
36,921 &\quad 24,515 &\quad 9,835 &\quad 36,046 \\
4.767 &\quad 6.037 &\quad 6.914 &\quad 8.461
\end{align*}

(1) (2) (3) (4) (5) (6) (7) (8) (9) (10)
T is not fully observed... ⇒ Adapt CART, first to censoring

“Classical KM weights” ⇒ additive version of $\hat{F}$ : let $\hat{G}$ KM estimator of $G(t) = \mathbb{P}(C \leq t)$, then we have

$$W_{i,n} = \frac{\delta_i}{n[1 - \hat{G}(Y_i^-)]} \Rightarrow \hat{F}(t) = \sum_{i=1}^{n} W_{i,n} 1_{Y_i \leq t} \rightarrow F(t)$$

IPCW with KM estimator of $G$ :

$$\sum_{i} W_{i,n}^* \psi(Y_i) = \frac{1}{n} \sum_{i} \frac{\delta_i \psi(Y_i)}{1 - G(Y_i^-)} \xrightarrow{LLN} E \left[ \frac{\delta \psi(Y)}{1 - G(Y^-)} \right] = E[\psi(T)]$$
Then to reporting delays $\tau$!

$\Rightarrow$ Necessary to modify the classical Kaplan-Meier weights!

Recall that we only get an observation when $C > \tau$...

Assume that $(T, \tau) \perp \perp C$, then one may consider

$$W_{i,n} = \frac{\delta_i 1_{\tau_i < Y_i}}{\sum_{j=1}^{n} 1_{\tau_j < Y_i \leq Y_j}} \prod_{Y_k < Y_i} \left(1 - \frac{\delta_k 1_{\tau_k < Y_k}}{\sum_{j=1}^{n} 1_{\tau_j < Y_k \leq Y_j}}\right).$$

Weighted CART, interpretation...
Recall that we wish to estimate $\mathbb{E}[M \mid T, X]$.

For RBNS claims, it amounts to estimate

$$M^* = \mathbb{E}[M \mid T \geq y, X]$$

Problem to use weighted CART (wCART) : $T$ is an incomplete (censored) explanatory variable when considering RBNS claims!
STRATEGIES TO PREDICT RBNS RESERVE

- Use Bayes formula:

\[ M^* = \mathbb{E}[M \mid T \geq Y, X = x] = \frac{\mathbb{E}[M \mathbb{1}_{T \geq Y} \mid X = x]}{\mathbb{E}[\mathbb{1}_{T \geq Y} \mid X = x]} \]

⇒ Yields to build 2 wCART trees (results presented hereafter)

- Use Plug-in principle:
  1. build \( \hat{\pi} \) estimator of \( \pi(t, x) = E[M \mid T = t, X = x] \) from wCART.
  2. then fit a model for \( T \mid T \geq y, X = x \), from which a prediction \( \hat{T}(y, x) \) can be computed (e.g. with Algorithm 1).
  3. finally, predict \( M^* \) by using plug-in type estimator \( \hat{\pi}(\hat{T}(y, x), x) \).

Once \( M^* \) predicted, easy to get the associated individual reserve!
Estimate the individual reserves at some given settlement dates.

- Use **backtesting**:
  1. consider only closed claims (known final claim amount);
  2. censoring/truncation variables updated at settlement date;
  3. define learning / test samples to build / validate estimators:
     - predictions compared to actual data **on test sample only**,
     - get indicators of the predictive power at individual level
  4. sum indiv. reserves to get an indicator of overall performance.

- Compare results to Chain Ladder ([Mack, 1993]) and Cox model ([Cox, 1972]);

- Bootstrap resampling to approximate variance of estimators.
Dataset: ausautoBI8999 (in R package CASdatasets).

Information:

- 22,036 claims in motor insurance over ten years (1989-1999),
- aggregated settled claim amount $\Rightarrow M$
- dates: accident, reporting, closing $\Rightarrow T, \tau$
- Individual claim features $\mathbf{X}$:
  - operational time (indicator for claim management difficulties),
  - type of injury,
  - # of injured people,
  - legal representation of the PH.

$\Rightarrow$ Introduce fictive administrative censoring with settlement dates!
PREDICTION ERRORS ON THE GLOBAL RESERVES

At ≠ settlement dates.
Censoring rate ≃ 50%, 1000 bootstrap samples.
VARIANCE OF THE ESTIMATORS
Plug-in strategy confirmed how crucial $T$ is to predict $M$:

- Our method allows to deal with risk heterogeneity,
- Well-suited to large reporting delays and long developments (without extrapolation).
Regression models and life tables (with discussion).

Tree-based methods : a useful tool for life insurance.