

## TP: PROVISIONS LIGNE À LIGNE PAR ALGORITHME CART REPONDÉRÉ

On se propose dans ce TP d'implémenter la technique de provisionnement ligne à ligne vue en cours. Il s'agit de déterminer une estimations des provisions IBNeR à partir des caractéristiques des sinistres de manière non-paramétrique.

Vous aurez besoin dans ce TP des librairies suivantes : `rpart`, `imputeYn`, `actuar`.

**Partie 1 : étude simulatoire** : vérification des propriétés de l'estimateur CART repondéré.

Pour cela,

- (1) simuler  $n + m$  réalisations iid  $(X_1, \dots, X_{n+m})$  de loi Uniforme,  $X_i \sim \mathcal{U}(0, 1)$  (remarque :  $n=100$  observations seront utilisées dans l'échantillon d'apprentissage, les  $m=50$  restantes seront utilisées pour la validation) ;
- (2) Simuler  $n + m$  durées de sinistres iid  $(T_1, \dots, T_{n+m})$ , de loi exponentielle telle que  $T_i \sim \mathcal{E}(\beta = 0,08 \times 1_{X_i \in [0,0,3[} + 0,05 \times 1_{X_i \in [0,3,0,6[} + 0,16 \times 1_{X_i \in [0,6,0,8[} + 0,5 \times 1_{X_i \in [0,8,1]})$

Quelle est la moyenne théorique de  $T$ , i.e.  $\mathbb{E}[T]$  ? Combien de partitions a-t-on dans la population ? Quelle est la moyenne théorique de  $T$  pour chacune de ces partitions, i.e.  $\pi_0(X) = \mathbb{E}[T|X]$  ?

- (3) Simuler  $n + m$  durées de censure  $(C_1, \dots, C_{n+m})$ , iid et indépendantes de  $T$ , de loi Pareto :  $C_i \sim \text{Pareto}(20; 1, 2)$ .
- (4) En déduire la durée de vie observée pour chaque sinistre :  $Y_i = \min(T_i, C_i)$ , ainsi que l'indicatrice de censure  $\delta_i = 1_{T_i \leq C_i}$ . Quel est le taux de censure que vous observez dans l'échantillon d'apprentissage ?
- (5) Calculer les poids Kaplan-Meier des  $n$  observations de l'échantillon d'apprentissage.
- (6) Construire un arbre CART repondéré sur cet échantillon. Visualiser cet estimateur, noté  $\hat{\pi}(X)$ . Les règles de segmentation sont-elles cohérentes ?
- (7) En notant  $w_i$  le poids Kaplan-Meier de l'observation  $i$ , calculer l'erreur quadratique pondérée (WMSE) de l'estimateur :

$$WMSE = \frac{1}{n} \sum_i w_i (\hat{\pi}(X_i) - \pi_0(X_i))^2$$

- (8) Reprenez le programme précédent en faisant varier  $(n, m)$ . Prenez  $n = 500$  et  $m = 250$ , puis  $n = 2000$  et  $m = 1000$ .
- (9) Faire un graphe de la WMSE en fonction de  $n$ . Qu'observez-vous ?

## Partie 2 : calcul de provisions.

On souhaite mettre en place un provisionnement ligne à ligne dans un contexte d'assurance incapacité-invalidité. Rappel : on se focalise ici sur l'évaluation de la provision IBNeR.

Quelques informations additionnelles concernant la construction des données :

- le pas de temps de la modélisation est trimestriel,
- nous considérerons l'évaluation de la provision globale par agrégation des provisions individuelles à 2 dates d'arrêtés distinctes,
- des échantillons d'apprentissage et de validation ont déjà été créés aléatoirement,
- nous procéderons par backtesting : seuls les sinistres clos à la fin de l'observation seront considérés afin de pouvoir comparer les prévisions à la réalité du terrain.

On se propose dans la suite d'effectuer les étapes suivantes

- (1) Charger les données contenues dans les fichiers RDS aux 2 dates d'arrêtés. Puis, pour chacune des dates d'arrêtés, poursuivre par les étapes suivantes.
- (2) Observez le contenu de l'objet chargé. Quels en sont les attributs ? Quelle est leur signification ?
- (3) Calculer la provision globale obtenue par Chain Ladder. Comparer à la provision réelle.
- (4) Calculer les provisions ligne à ligne des sinistres encore ouverts à la date d'arrêt : pour chaque durée atteinte  $k$  en considérant l'échantillon d'apprentissage ( $k$  est un élément d'une grille à définir par vous-même),
  - (a) sélectionner les sinistres (censurés ou non) tels que  $Y \geq k$  ;
  - (b) estimer les poids KM depuis les données ;
  - (c) construire l'arbre CART pondéré qui permet d'estimer  $E[T - k | T > k, X]$ , avec  $Y$  ici représentée par la variable 'EndObsW' et  $X$  est le vecteur constitué des informations 'Sex', 'SPC', 'ComNet', 'Cause' et 'BegAgeW' ;
  - (d) élaguer l'arbre. Puis sur l'échantillon de validation,
  - (e) prévoir la durée de vie résiduelle des sinistres qui ont déjà une ancienneté de  $k$  ;
  - (f) en déduire la provision pour ces sinistres non clos ayant cette ancienneté ;
  - (g) accroître  $k$  et revenir à l'étape (a) ;
- (5) Faire la somme des provisions obtenues par ancienneté de sinistres encore ouverts pour en déduire la provision globale.
- (6) Comparer la performance des méthodes Chain Ladder et CART pondéré pour l'estimation de la provision globale sur l'échantillon de validation.

Remarque : nous aurions pu considérer une implémentation différente. Au lieu de modéliser une espérance de vie résiduelle, nous aurions pu modéliser l'espérance de vie conditionnée par le fait d'avoir dépassé un certain seuil de durée. Cela aurait impliqué l'usage de la règle de Bayes, et aurait nécessité de construire 2 arbres CART pondérés pour chaque seuil d'ancienneté.