
Mémoire Majeure Actuariat

Tarification en assurance IARD avec les GLM en intégrant les données issues de la sécurité routière

Encadrant	Christophe Dutang
Structure	Université Paris Dauphine
Courriel	dutang@ceremade.dauphine.fr
Données	Données réelles d'un assureur français et de l'ONISR

Contexte et objectif du projet

L'assurance est un contrat par lequel, moyennant le versement d'une prime dont le montant est fixé a priori (en début de période de couverture), l'assureur s'engage à indemniser l'assuré pendant toute la période de couverture (généralement un an). Cette prime doit refléter le risque associé au contrat. On renvoie à CHARPENTIER et DENUIT (2004a) sur la théorie du calcul des primes et à CHARPENTIER et DENUIT (2004b) pour les considérations économiques.

Pour chaque police d'assurance, la prime est fonction de variables dites de tarification permettant de segmenter la population en fonction de son risque. Il est usuel d'utiliser une approche fréquence/sévérité ou une approche indemnitaire pour modéliser le coût annuel d'une police d'assurance, voir cours Actuariat 1. Sur les données utilisées dans ce projet, on utilisera cette dernière approche car on ne dispose pas des montants individuels de sinistre.

Le but de ce projet est de proposer un tarificateur en se basant la méthodes des modèles linéaires généralisés (GLM proposé par MCCULLAGH et NELDER (1989)) en testant l'apport de données externes issues de la sécurité routière. Nous renvoyons aussi vers EFRON et HASTIE (2016) qui est disponible gratuitement en ligne mais moins complet.

Les données routières à utiliser sont les données gouvernementales de l'Observatoire national interministériel de la sécurité routière, ONISR (2018). Tout accident corporel survenant en France entre 2005 et 2017 est répertorié par l'observatoire qui met à disposition 4 fichiers annuels sur les caractéristiques de l'accident, le lieux des accidents, les véhicules impliqués et les usagers impliqués.

Pour tester la pertinence de l'apport de ces données nationales, on commencera d'abord par modéliser la prime pure avec un GLM sans inclure les données de la sécurité routière. Ensuite, on réalisera un zonier issu des données de l'ONISR qui permettra d'ajouter une ou plusieurs variables explicatives à chacune des données individuelles pour l'incorporer dans un second GLM.

Pour ce projet, les applications numériques seront à réaliser en R, (R CORE TEAM, 2020), et on encourage les étudiants à étudier le livre de CHARPENTIER (2014) sur ce sujet.

Guide pour le démarrage du projet

En utilisant les données du package R **CASdatasets**, (DUTANG et CHARPENTIER, 2019), les données `pgl7trainpol` et `pgl7trainclaim` d'assurance automobile vont permettre de calibrer un GLM sur la sévérité et la fréquence. Les données de test seront `pgl7testyear1`.

Les étapes suivantes devront être abordées :

1. Explorez les données `pgl7trainpol` et `pgl7trainclaim` à l'aide des techniques exploratoires usuelles (statistiques descriptives, ACP, AFC).
2. Présenter les modèles linéaires généralisés en attachant une attention particulière aux lois considérées.
3. Modéliser la fréquence des sinistres sur la base `pgl7trainpol` et la sévérité des sinistres sur la base `pgl7trainclaim` à l'aide des GLMs. Le choix des lois et la sélection des variables explicatives doivent être expliqués notamment à l'aide de statistique d'adéquation.
4. Sélectionner le meilleur modèle GLM en un sens qui sera à préciser.
5. En déduire une prime pure pour les polices étudiées.
6. Réaliser une analyse approfondie des données l'ONISR pour déterminer un zonier à une maille spatiale à déterminer.
7. Ajouter plusieurs variables explicatives issues du zonier. On notera les nouveaux jeux de données `pgl7trainpolONISR` et `pgl7trainclaimONISR`.
8. Modéliser la fréquence des sinistres sur la base `pgl7trainpolONISR` et la sévérité des sinistres sur la base `pgl7trainclaimONISR` à l'aide des GLMs.
9. Sélectionner le meilleur modèle GLM en un sens qui sera à préciser.
10. Comparer les tarifs obtenus sur la bases de test `pgl7testyear1`.

Références

- CHARPENTIER, A. et DENUIT, M. (2004a). Mathématiques de l'assurance non vie. T. 1. Economica.
- (2004b). Mathématiques de l'assurance non vie. T. 2. Economica.
- CHARPENTIER, A., éd. (2014). Computational Actuarial Science with R. Chapman et Hall-CRC.
- DUTANG, C. et CHARPENTIER, A. (2019). CASdatasets : Insurance datasets.
- EFRON, B. et HASTIE, T. J. (2016). Computer Age Statistical Inference. Cambridge University Press.
- MCCULLAGH, P. et NELDER, J. A. (1989). Generalized Linear Models. 2nd. Chapman et Hall.
- ONISR (2018). Description des bases de données annuelles des accidents corporels de la circulation routière. Rapp. tech. Ministère de l'intérieur. URL : <http://www.securite-routiere.gouv.fr/la-securite-routiere/l-observatoire-national-interministeriel-de-la-securite-routiere>.
- R CORE TEAM (2020). R : A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL : <https://www.R-project.org/>.