Premier rapport

Apprentissage statistique en actuariat
ACT-4114

Équipe 13

---

**Analyse exploratoire**

# NOTRE TITRE

---


| *Par* | *Numéro d'identification* |
|---|---|
| David Boulanger | 536 986 550 |
| Ann-Sophie Hamel | YZX YZX YZX |
| Léonie Allard | ZYX ZYX ZYX |
| Nathaniel Gaudreau | XYD XYD XYD |


*Travail présenté à*
*Madame*
Marie-Pier Côté

1 mars 2024


Université Laval | Faculté des sciences et de génie
École d'actuariat

# Table des Matières

# Introduction

L'objectif final de ce travail est de modéliser la fréquence des réclamations de la couverture de responsabilité civile (dommages matériels) en assurance automobile pour un portefeuille français. Pour le risque $j$, le nombre de réclamations sera noter par $N_j$. Pour procéder à la modélisation, on utilise les caractéristiques disponibles dans notre jeux de données. On notera, pour le risque $j$, le vecteur de variables explicatives par $X_j$. On tiendra aussi en compte l'exposition au risque dans notre modèle. Cette dernière se présente sous la forme de nombre de jours à risque avec un maximum de 1 an (365 jours). Cette variable sera transformer par une division de 365 afin d'obtenir une proportion d'année couverte. On notera l'exposition du risque $j$ par $t_j$. Dans le tableau de données original on retrouve :

- la variable de fréquence $N$ sous le nom `Numtppd` ;
- la variable d'exposition $365t$ (version nombre de jours) sous le nom `Exppdays`.

En ce qui concerne le jeu de données, il est disponible directement en `R` dans le paquetage `CASdatasets`. Plus précisément c'est le jeux de données `pg15training`. Les données ont été utilisées par l'institut française des actuaires dans un concours/jeux de tarification. Elles proviennent d'assureurs automobile privées inconnus. La matrice contient 2 ans d'observations (2009 et 2010) avec 50 000 observations dans chacune de ces deux années. Voici quelques informations pertinentes avant de débuter l'analyse :

1. l'âge minimal pour conduire en France est de 18 ans ;
2. la couverture étudiée est obligatoire ;
3. certaines variables catégorielles contiennent des groupes dont la signification demeure non spécifiée pour des raisons de confidentialité.

Pour plus d'information sur les données il est possible d'aller voir la documentation sur CRAN ou bien simplement de faire la commande suivante en `R` : `help(pg15training)`.

# Analyse exploratoire des données

Cette section est dédié à la détection erreurs ou anomalies dans notre jeux de données ainsi qu'à l'approche adopté pour la correction de ces erreurs. On regarde également les grandes lignes de l'analyse préliminaire effectuée.

## Doublons

D'abord, on remarque que certaines polices sont présentes en double dans le jeux de données. En fait, il s'agit des 21 premières lignes qui sont en surplus. Ces premières lignes sont exactement comme leur doublure à l'exception qu'aucun montant de réclamation (`Indtppd`) n'a été enregistré. La correction est assez directe, on retire simplement les 21 premières lignes et on retrouve maintenant un nombre exact de 100 000 observations tel que documenté dans la rubrique d'aide en `R`. On présente dans le tableau 1 ci-dessous un exemple de doublon avec la police numéro 200114978 (la première ligne).

Table 1: Exemple de doublon

| Ligne | PolNum | Numtppd | Indtppd |
|---|---|---|---|
| 1 | 200114978 | 1 | 0.0000 |
| 129 | 200114978 | 1 | 362.6195 |

## Variable endogène

## Variables explicatives

Pour les intervalles de confiance approximatifs, on utilise le théorème central limite (TCL).

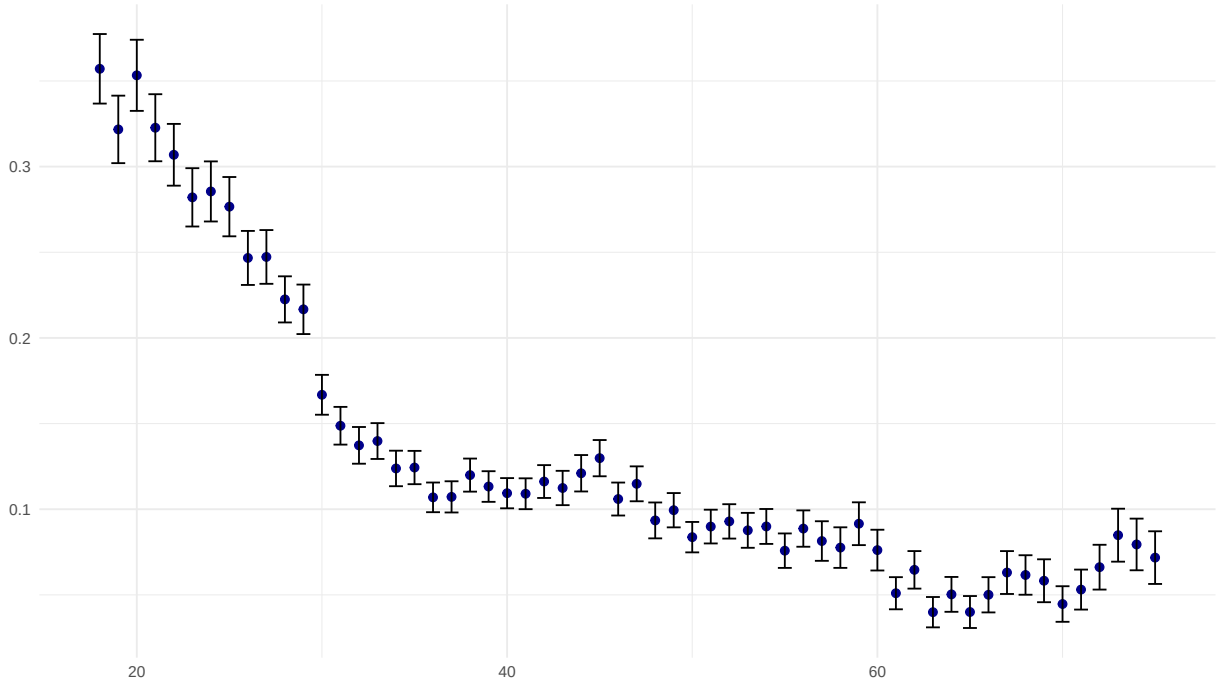


Figure 1: Moyenne de réclamation par age

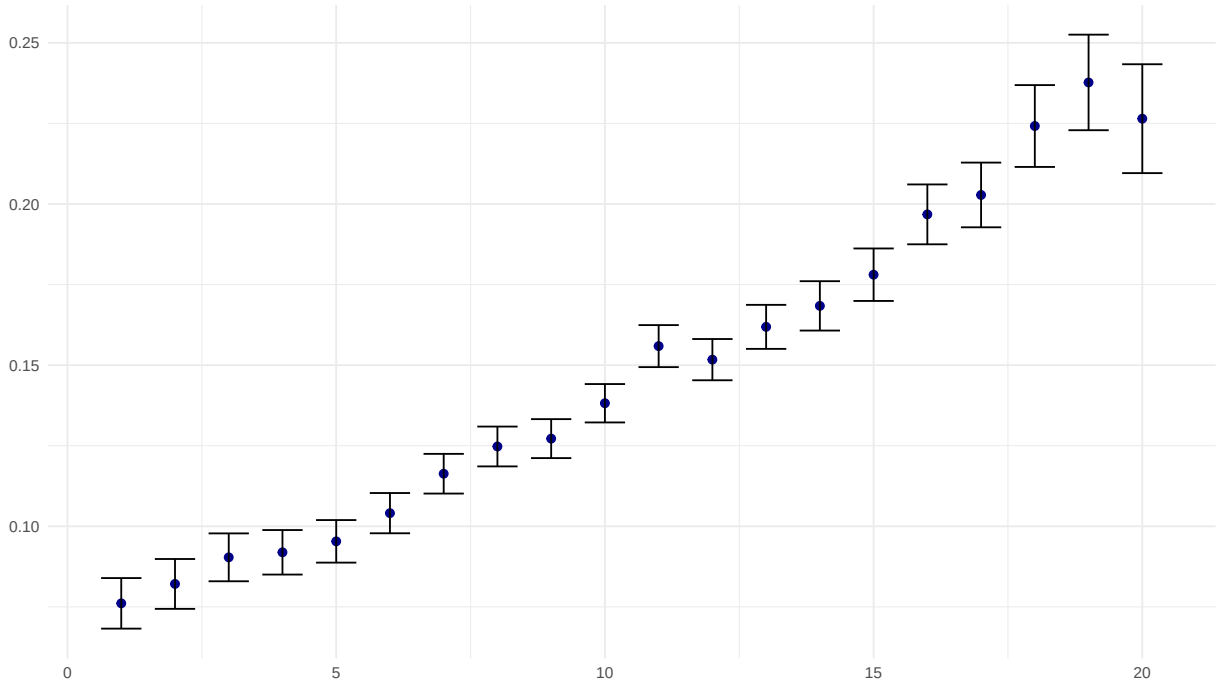


Figure 2: Moyenne de réclamation par Group1

## Traitement des valeurs manquantes

## Analyse en composantes principales

## Création de nouvelles variables explicatives

## Classification hiérarchique

## Algorithme des k-moyennes

## Conclusion

## Bibliographie

## Annexe

### Description du jeu de données

Comme sur le forum, mais sans fautes d'orthographes.

### Déclaration de l'utilisation de l'intelligence artificielle

générative

Il faut insérer la déclaration complétée ici.