**République Algérienne Démocratique et Populaire**
**Ministère de l'Enseignement Supérieur et de la Recherche Scientifique**


UNIVERSITÉ A. MIRA DE BÉJAIA

FACULTÉ DES SCIENCES EXACTES
DÉPARTEMENT DE RECHERCHE OPÉRATIONNELLE
Spécialité: Mathématiques Financières

# MÉMOIRE DE MASTER

THÈME

# Etude et estimation des sinistres en assurance automobile : cas de la SAA

*Préparé par :*
Walid OUCHERIF

Soutenu le **15/09/2021** devant le jury :

| | | | |
|---|---|---|---|
| Président | Mr BRAHMI Belkacem | MCA | Univ. A.Mira Béjaia |
| Encadreur | Mr TOUCHE Nassim | MCA | Univ. A.Mira Béjaia |
| Examinatrice | Mme BENOUARET Zina | MCB | Univ. A.Mira Béjaia |
| Examinatrice | Mme TAKHEDMIT Baya | MCB | Univ. A.Mira Béjaia |

Promotion : 2020/2021

*« L'économiste doit être mathématicien, historien, politicien et philosophe. Il doit aborder simultanément l'abstraction et la réalité et étudier le présent à la lumière du passé en vue de l'avenir sans qu'aucun aspect de la nature des institutions ne lui échappe. »*

John Maynard Keynes.

# Résumé

Ce mémoire à pour objet l'étude de la tarification de la garantie Tous Risques en assurance automobile dans le cas du- marché algérien. Le système tarifaire actuel de cette garantie n'est pas équitable car chaque assuré paie une prime proportionnelle à la valeur assurée du véhicule (valeur demandée par l'assuré) sans aucune autre distinction. Ceci est bien évidemment problématique car les assurés ne sont pas tous exposés au même risque. Le travail de ce mémoire a pour objectif de proposer une esquisse pour un système de tarification basé sur une individualisation de chaque prime d'assurance qui devrait être proportionnelle au risque de chaque assuré. Le principal outil mathématique ayant servi à l'élaboration de ces systèmes est le Modèle Linéaire Généralisé (MLG). En statistique, le GLM (Generalized Linear Model en anglais) est une généralisation souple de la régression linéaire. Le GLM généralise la régression linéaire en permettant au modèle linéaire d'être relié à la variable réponse via une fonction lien et en autorisant l'amplitude de la variance de chaque mesure d'être une fonction de sa valeur prévue. Dans ce document, il sera donné plusieurs modèles GLM basés sur des distributions différentes et une comparaison sera effectuée pour déterminer les modèles qui sont les plus performants dans la prédiction de la sinistralité en Algérie.

# Abstract

The purpose of this thesis is to study the pricing of the All Risks guarantee in automobile insurance in the case of the Algerian market. The current pricing system for this guarantee is not fair because each policyholder pays a premium proportional to the insured value of the vehicle (value requested by the policyholder) without any other distinction. This is obviously problematic because policyholders are not all exposed to the same risk. The purpose of this thesis is to provide a sketch for a pricing system based on an individualization of each insurance premium which should be proportional to the risk of each policyholder. The main mathematical tool used in the development of these systems is the Generalized Linear Model (MLG). In statistics, the GLM (Generalized Linear Model) is a flexible generalization of linear regression. GLM generalizes linear regression by allowing the linear model to be related to the response variable via a link function and allowing the magnitude of the variance of each measure to be a function of its predicted value. In this document, several GLM models based on different distributions will be given and a comparison will be made to determine which models are the most efficient in predicting claims in Algeria.

# Remerciements

Dans un premier temps, je tiens à remercier mes parents qui durant toute ma scolarité et ma vie, ont toujours cru en moi et en mes choix et je les remercie infiniment pour cela.

Ensuite mes remerciements vont à mon encadreur Mr TOUCHE Nassim qui m'a accordé son temps et surtout sa confiance dès le départ et qui m'a apporté de précieux indices sur la méthodologie à suivre pour mener ce travail dans la meilleure des manières.

Ce mémoire est bien évidemment un travail personnel mais il est certain que sans aide et conseil, il n'en serait qu'un travail de mauvaise qualité. C'est pour cette raison que je remercie infiniment les membres de la division automobile de la SAA, et particulièrement Mr ZERKAOUI Abdenour, directeur de la production automobile pour tous ses conseils et son aide précieuse. Merci également à Mr BOUFIDJELINE Hamid pour toutes les informations qui ont été très importantes pour la réalisation de ce mémoire, mais également pour avoir lui même préparé la base de données sur laquelle j'ai pu travailler. Merci à vous car tout ce que je sais aujourd'hui sur les assurances provient de vous.

Je tiens aussi à remercier Mr ARBANE Hamza, chef de la division automobile de la SAA pour son chaleureux accueil et pour m'avoir immédiatement intégré à son équipe dans les meilleures conditions possibles.

Enfin, mes remerciements vont aux membres du jury pour l'intérêt qu'ils ont porté à notre mémoire en acceptant d'examiner notre travail et de l'enrichir à travers leurs suggestions et propositions.

# Table des matières

# Table des figures

# Liste des tableaux

# Introduction générale

La définition la plus courante de l'opération d'assurance est celle de Joseph Hémard (1876-1932), juriste français et professeur en faculté de droit : « L'assurance est une opération par laquelle une partie, l'assuré, se fait promettre moyennant une rémunération (la prime ou cotisation) pour lui ou pour un tiers, en cas de réalisation d'un risque, une prestation par une autre partie, l'assureur, qui prenant en charge un ensemble de risques, les compense conformément aux lois de la statistique ».[29]

L'assurance automobile est une assurance qui couvre les dommages causés «avec» ou «à» un véhicule automobile. Elle est en général obligatoire et est régie par un code des assurances. Son objectif est d'apporter une compensation financière face aux pertes subies par un assuré ou une personne tierce, notamment lors d'un accident de la route où lors de la survenance d'un évènement où le véhicule a été endommagé dans certaines conditions. Les formes de contrats comme les garanties proposées par les compagnies d'assurances sont variées. Chaque contrat souscrit est propre à une situation. Que ce soit le véhicule, les garanties choisies, le souscripteur ou la compagnie d'assurance.

Ainsi dans ce mémoire, nous travaillerons sur une seule des garanties proposées par les assureurs algériens, la garantie Tous Risques. Elle ne concernera que les indemnisations concernant le véhicule lié à la police d'assurance et dans le cas où la non responsabilité de l'assuré n'est pas mise en jeu.

Ce mémoire est motivé par la question de la tarification automobile en Algérie. Le système tarifaire actuel repose sur des statistiques anciennes et n'a pas évolué durant des décennies et la principale problématique qui en sort est : pourquoi tous les individus paient la même prime même s'ils ne sont pas tous exposés au même risque ? En effet, en l'état actuel des choses, un bon conducteur paie exactement la même prime qu'un mauvais conducteur, ce qui est bien sûr injuste. Si le marché algérien était ultra compétitif, cela poserait évidemment un problème car les bons assurés (faibles risques) préfèreraient payer un assureur qui décide de prendre en considération l'hétérogénéité des risques (en faisant une segmentation). Ainsi, dans ce mémoire, nous allons essayer de donner une alternative au système actuel en étudiant la sinistralité au niveau individuel, ce qui pourra permettre de proposer une prime proportionnelle au risque de chaque assuré.

Pour ce faire, nous allons utiliser des outils mathématiques qui sont souvent adoptés dans la littérature actuarielle, à savoir les modèles linéaires généralisés (GLM : Generalized Linear Models). En effet, comme nous le verrons dans notre étude, ces outils sont très pratiques pour modéliser la fréquence des sinistres mais le sont malheureusement moins pour essayer de modéliser les coûts de ceux-ci. Évidemment, cela ne nous empêchera pas d'utiliser les résultats que nous obtiendrons des GLM pour essayer d'estimer des primes d'assurance car il faudrait savoir que notre étude est un travail inédit dans le marché d'assurance algérien et qu'il n'y a donc aucun point de référence sur lequel nous pouvons comparer nos résultats. C'est pour cela que même si notre travail n'est pas encore adapté pour une mise en place réelle, il reste cependant

révélateur qu'une révision du système tarifaire actuel s'impose.

Ainsi, dans ce mémoire, nous présenterons dans une première partie le contexte de l'assurance en Algérie car il est atypique et mérite une présentation dans le détail. Dans la foulée, nous présenterons l'organisme d'accueil qui est la Société Nationale d'Assurance SAA, sa position dans le marché algérien et la situation économique de l'assurance automobile. Dans une deuxième partie, il s'agira de donner tous les détails concernant la base de données utilisée pour faire l'étude du mémoire. Ensuite, nous allons poser les bases mathématiques sur lesquelles vont se reposer ensuite tous les résultats, qui eux seront présentés dans une quatrième partie.

Nous donnerons quelques conclusions en fin de mémoire qui pourraient aider la compagnie d'assurance à mieux modéliser sa sinistralité et à correctement répartir la charge des assurés en fonction du risque de chacun.

# Chapitre 1

# Paysage de l'assurance en Algérie

**Introduction** – Dans ce chapitre, nous allons présenter le secteur des assurances en Algérie en donnant d'abord un bref historique de l'activité d'assurance, puis les chiffres de l'assurance dans sa globalité et ensuite les chiffres qui nous intéressent, c'est-à-dire ceux de l'assurance automobile. Enfin, une présentation de l'organisme d'accueil, la société nationale d'assurance SAA, sera faite. Ses chiffres et son positionnement dans le marché des assurances seront également présentés.

## 1.1 Histoire de l'assurance en Algérie

Les besoins en assurance des Algériens durant la période coloniale sont considérés comme insignifiants compte tenu de leurs revenus et de leur situation socioculturelle. L'assurance en Algérie a commencé à se développer au lendemain de l'indépendance mais la majorité des opérations des assurances restaient aux mains des entreprises françaises. Pendant cette période, une grande partie des primes collectées par les compagnies d'assurance ont été transférées à l'étranger. Il aura fallu attendre 1966 pour que l'état reprenne le monopole sur l'assurance en Algérie. L'exploitation de cette activité était désormais réservée à l'état via les entreprises nationales qui avaient vues le jour durant les années 1963-1964.[6] En 1976, les quatre compagnies d'assurance nationales – CAAR, SAA, MAATEC et la CCRMA – se spécialisent chacune dans la couverture d'un certain type de risque. En 1995, une ordonnance supprime le monopole de l'état sur le marché d'assurance, autorisant ainsi la création de compagnies d'assurance privées. Grâce à ce changement, la responsabilité civile n'est plus la seule garantie proposée par les assureurs et permet ainsi une certaine rentabilité de ce secteur.

# Chapitre 2

# Analyse statistique des données

**Introduction** – Dans ce chapitre, nous allons présenter le contenu de la base de données et passer en revue toutes les manipulations qui ont été faites sur celle-ci afin de préparer au mieux les données pour une bonne modélisation de la sinistralité automobile du portefeuille de la SAA. Nous allons également proposer quelques variables (informations) supplémentaires qui pourraient s'avérer utiles pour mieux décrire la sinistralité en Algérie.

## 2.1 Objectif de l'étude

Le but de cette étude est d'essayer de modéliser au mieux la sinistralité automobile algérienne. Cette étude permettra une ébauche d'un système de tarification de la garantie Tous Risques mais pourra également servir de base pour les autres types de garanties dommages car le but sera de déterminer la prime pure [36] d'un contrat d'assurance. La prime pure correspond au montant moyen d'un sinistre que devra payer la compagnie d'assurance si le risque survenait. Le calcul de la prime pure a pour but d'évaluer, pour chaque assuré, le montant attendu des sinistres pour la période d'assurance étudiée. Cette évaluation se fait le plus souvent par des méthodes statistiques. La sinistralité est divisée en plusieurs composantes, chacune étant évaluée indépendamment :

— La probabilité d'un sinistre normal.

— Le coût d'un sinistre normal.

— La probabilité d'un sinistre grave.

— Le coût d'un sinistre grave.

S'il est possible de modéliser avec précision le potentiel futur sinistre d'un individu, il est également possible d'utiliser cette information pour le provisionnement d'un dossier sinistre lors de son ouverture. Cette information pourrait être utilisée plus généralement dans les PSAP (Provisions Sinistres A Payer).

## 2.2 Base de données et préparation des données

Les données qui ont été mises à notre disposition ont d'abord subi une consolidation pour n'en faire qu'une seule table, puis ont subi un nettoyage. Malheureusement, un fait inhérent aux systèmes d'informations algériens est la mauvaise construction des bases de données et des informations qu'elles contiennent. Il est clair et admis de tout le monde que la qualité de saisie des données n'est malheureusement pas optimale dans l'administration algérienne, et cela affectera grandement l'étude car les fausses données créent du biais et peuvent mener à de fausses conclusions. Il est donc primordial avant toute étude de mener une exploration des données pour chercher et traiter ces mauvaises données. Pour cette étape, il n'y a malheureusement pas de méthode universelle ayant une certaine efficacité dans le traitement des données manquantes,

mal saisies ou aberrantes. Cette partie de l'étude est en général la plus coûteuse en temps car il n'est pas possible de savoir vraiment ce que l'on doit chercher comme mauvaises informations. Ensuite, il s'agira de décider quoi faire de ces mauvaises données. La solution la plus facile serait bien sûr de les éliminer, mais il serait préférable d'essayer de trouver un moyen d'expliquer ou de corriger ces données car elles pourraient appartenir à une sous population qui ne possèdent pas beaucoup d'observations. Une autre option serait d'attribuer une valeur moyenne aux valeurs manquantes ou une valeur aléatoire de la variable en question. De plus la modification d'une donnée revient à supposer la survenue d'un évènement qui n'a pas forcément eu lieu et la suppression revient à supposer que l'évènement n'a pas eu lieu. Si ce travail de récupération est mené sur beaucoup de données, il serait nécessaire de procéder à un traitement automatique des données à réviser.

Dans notre étude, par défaut de temps, nous avons dû éliminer toute observation paraissant aberrante.

### 2.2.1 Contenu de la base de données et ses variables

La base de données contient les informations de toutes les polices d'assurances souscrites auprès de la SAA dans le cadre de la garantie Tous Risques durant les années 2014 à 2018. Avant nettoyage, il y avait au total quelques 850 000 observations. Le fichier qui nous a été remis en comporte 740 000 car toutes les observations où des informations importantes ont été omises ou mal saisies ont dues être retirées. La table contient la liste des polices d'assurances de deux types de catégories de véhicules : les véhicules particuliers ou les véhicules appartenant à des flottes. Il n'est bien entendu pas possible d'étudier ces deux catégories de la même façon car selon une analyse descriptive, ces deux types de catégories présentent une sinistralité très différente, en plus de ne pas contenir les mêmes variables explicatives ; les véhicules des flottes ne sont pas supposés appartenir à des individus à qui l'on pourrait demander un âge, un sexe, un âge de permis, ...

La base de données comporte 40 variables mais ne sont pas toutes significatives. Il sera question de faire une étude sur la fréquence des sinistres puis une étude sur la gravité des sinistres. Ces deux études portent donc sur la variable `sinistres` qui est le nombre de sinistres survenus durant la période d'exposition d'une police d'assurance, et la seconde variable `ChargeSin` qui est le montant que l'assureur a dû payer en totalité à son assuré pour les sinistres survenus durant la période d'exposition.

La liste des variables utiles sont les suivantes :

1. `id` : chaîne de caractères servant de clé d'identification de chaque police d'assurance. Elle est la concaténation du numéro d'agence, du numéro de police, du code d'appartenance à l'une des catégories citées plus haut.
2. `dr` : variable catégorielle. La direction régionale auquel appartient l'agence dans laquelle le contrat d'assurance a été souscrit. Cette variable est très importante car elle nous servira de variable spatiale, permettant une segmentation géographique du portefeuille. Cette variable contient 15 modalités mais on ne retiendra que 14. La modalité `annaba` a été retirée car elle ne contient que 700 observations, ce qui ne représente pas la quantité d'observations réelles qui se situe à 50 000. L'extraction des données de la direction régionale de Annaba est incorrecte. Ceci est problématique car beaucoup des régions du Nord-Est du pays dépendent de cette direction régionale.
3. `agence` : variable catégorielle. Code de l'agence dans laquelle la police d'assurance a été souscrite. Pour l'instant, nous n'avons pas utilisé cette variable car c'est une variable catégorielle et elle possède 592 modalités (592 agences SAA en Algérie en 2018) et cela représenterait un énorme coût en temps de calcul. Il serait judicieux d'utiliser cette variable

dans une étude future car on pourrait supposer que la plupart des individus souscrivent une assurance dans une agence près de leur domicile et on pourrait donc peut être associer la zone de circulation du véhicule à la localisation de l'agence (segmentation par Daïras-Communes). Une autre idée serait d'utiliser les codes d'agences des contrats d'assurances pour déduire les densités d'habitations de certaines régions car cette variable peut intuitivement et conceptuellement aider à expliquer la sinistralité.

4. `categorie` : variable catégorielle. Cette variable catégorielle possède deux modalités déjà citées précédemment. Elle différencie les contrats des flottes de ceux des véhicules particuliers. Cette variable permettra de créer deux bases de données sur lesquelles nous travaillerons : `basePartic` et `baseFlotte`
5. `annee` : valeur entière. Représente l'année de souscription du contrat. Etant donné l'instabilité de l'économie algérienne, il sera important d'utiliser cette variable pour actualiser les valeurs des montants des sinistres et des valeurs assurées des véhicules. Plus de détails sur ce point par la suite.
6. `duree` : variable catégorielle à deux niveaux. Période d'exposition au risque d'un contrat : 6 mois ou une année.
7. `police` : chaîne de caractères. Numéro de police du contrat.
8. `avenant` : valeur entière. Représente le numéro d'avenant du contrat. Par définition, un avenant est un : « acte par lequel on modifie les termes d'un contrat ». En assurance, plus précisément un avenant est une révision du contrat pour corriger, apporter des modifications, ajouter ou retirer des garanties ou bien renouveler celui-ci. La variable suivante a été créée à partir de `avenant`
9. `cutAvenant` : variable catégorielle. Variable déduite de la précédente possédant deux modalités : `new` et `old` permettant de différencier les nouvelles polices d'assurances des anciennes. Elle permettra de séparer les nouveaux assurés de l'agence de ceux qui sont à la SAA depuis au moins 1 an, en supposant que la proportion des avenants de modification liés à des rectifications des nouveaux contrats est faible. L'idée derrière la création de cette variable est de savoir s'il existe bien une tendance des nouveaux contrats à être plus sinistrés que ceux des assurés fidèles ou de savoir s'il y a un phénomène de souscription à la garantie Tous Risques dans un seul but : réparer son véhicule. La création de cette variable est intéressante mais mérite une étude à elle seule car la séparation en 2 modalités exclue le cas des contrats RC renouvelés pendant seulement une année pour une Tous Risques.
10. `code` : valeur entière. Code risque. Un contrat d'assurance peut porter sur plusieurs véhicules et le code risque liste ceux-ci, notamment pour les contrats des flottes.
11. `sex` : variable catégorielle. Sexe du souscripteur de la police d'assurance. Variable pas très intéressante comme on le verra plus tard dans l'étude. L'interprétation sera donnée plus tard. Variable indisponible dans `BaseFlotte`.
12. `age` : variable continue. Age du souscripteur du contrat. Variable indisponible dans `BaseFlotte`.
13. `permis` : variable continue. Ancienneté du permis en années. Variable indisponible dans `BaseFlotte`.
14. `nvPermis` : variable catégorielle. Variable permettant de séparer les assurés avec un nouveau permis ou non. Variable indisponible dans `BaseFlotte`. Cetta variable peut être déduite de la précédente en séparant les anciennetés de permis `0` des autres.
15. `marque` : variable catégorielle avec 113 modalités. Elle représente la marque du véhicule.
16. `brand` : variable catégorielle à 56 modalités représentant les marques mais regroupées selon une certaine logique. Par exemple les marques `RENAULT` et `DACIA` sont regroupées en un groupe `RENAULT-DACIA`.

17. `ageVeh` : variable continue. Représente l'âge du véhicule en années.
18. `genre` : variable catégorielle avec 12 modalités. Elle représente le genre de véhicule ; exemples : Transport Publique de Marchandises, Véhicules de plus de 3,5 tonnes, Tracteurs Routiers, ...
19. `usage` : variable catégorielle avec 10 niveaux. L'usage auquel le véhicule est supposé être assuré ; exemples : affaire, fonctionnaire, location, taxi, ...
20. `Fiscal` : variable continue. Puissance fiscale du véhicule.
21. `zone` : variable catégorielle à deux modalités `nord` et `sud`. Variable permettant de séparer les véhicules sensés être utilisés dans le nord ou dans le sud du pays.
22. `inflam` : variable catégorielle à deux modalités permettant de séparer les véhicules transportant des marchandises inflammables ou non.
23. `tauxRed` : variable numérique. Représente un taux de réduction appliqué sur la prime demandée pendant la souscription.
24. `PleinTarifAct` : variable continue. Représente la prime Tous Risques avant application du taux de réduction. Ce montant est actualisé selon l'année de souscription du contrat.
25. `prime_trAct` : variable continue représentant le montant de la prime exact payé par l'assuré pour la garantie Tous Risques. Réduction appliquée et montant actualisé.
26. `sinistres` : valeur entière. Variable sur laquelle portera une des études dans ce mémoire. Représente le nombre de sinistres reportés durant la période d'exposition d'un contrat.
27. `VA` : variable continue. Représente la valeur assurée du véhicule. Ce montant est demandé par l'assuré pendant la souscription du contrat. Ce montant a été actualisé.
28. `ChargeSin` : variable continue. Deuxième variable sur laquelle portera l'étude. Représente le montant total des déboursements qu'a dû faire l'assureur sur un contrat concernant uniquement la garantie Tous Risques. Valeur actualisée.

### 2.2.2 Variables explicatives manquantes

La liste des variables contenues dans cette base de données est conséquente mais il manque malheureusement de précieuses informations qui à notre sens pourraient aider grandement à mieux prédire la fréquence et la gravite des sinistres. Une liste non exhaustive des variables qui pourraient apporter une amélioration à la modélisation de la sinistralité :

— `localisation` : Déduire la ville dans lequel est supposé circuler le véhicule en fonction de la localité de l'agence dans lequel a été souscrit le contrat. Pour améliorer la qualité de cette donnée, il faudrait faire correspondre également l'usage du véhicule. Car un véhicule assuré à un endroit peut être utilisé pour aller à un travail situé dans une autre ville. Cas fréquent dans le centre-nord du pays. Enfin l'ajout de cette donnée pourra déterminer la densité de population de la région dans laquelle le véhicule devra circuler.

— `profession` : cette variable existe dans les bases de données mais n'a pas été retenue car jugée de mauvaise qualité. Une idée serait d'utiliser la carte grise des véhicules pour déterminer correctement cette donnée.

— `dateSinistre` : la date pourrait renseigner sur une certaine saisonnalité des sinistres notamment sur la fréquence. On pourrait partitionner cette variable en mois ou en trimestre pour déterminer les périodes de l'année les plus sinistrées. Par exemple le mois de ramadan ou la période estivale.

— `energie` : le type de carburant ou énergie utilisée par les véhicules pour fonctionner. Cette donnée existe dans les bases de données de la SAA mais a été jugée inutile. Nous ne partageons pas ce point de vue car elle pourrait donner un indice sur l'utilisation du véhicule.

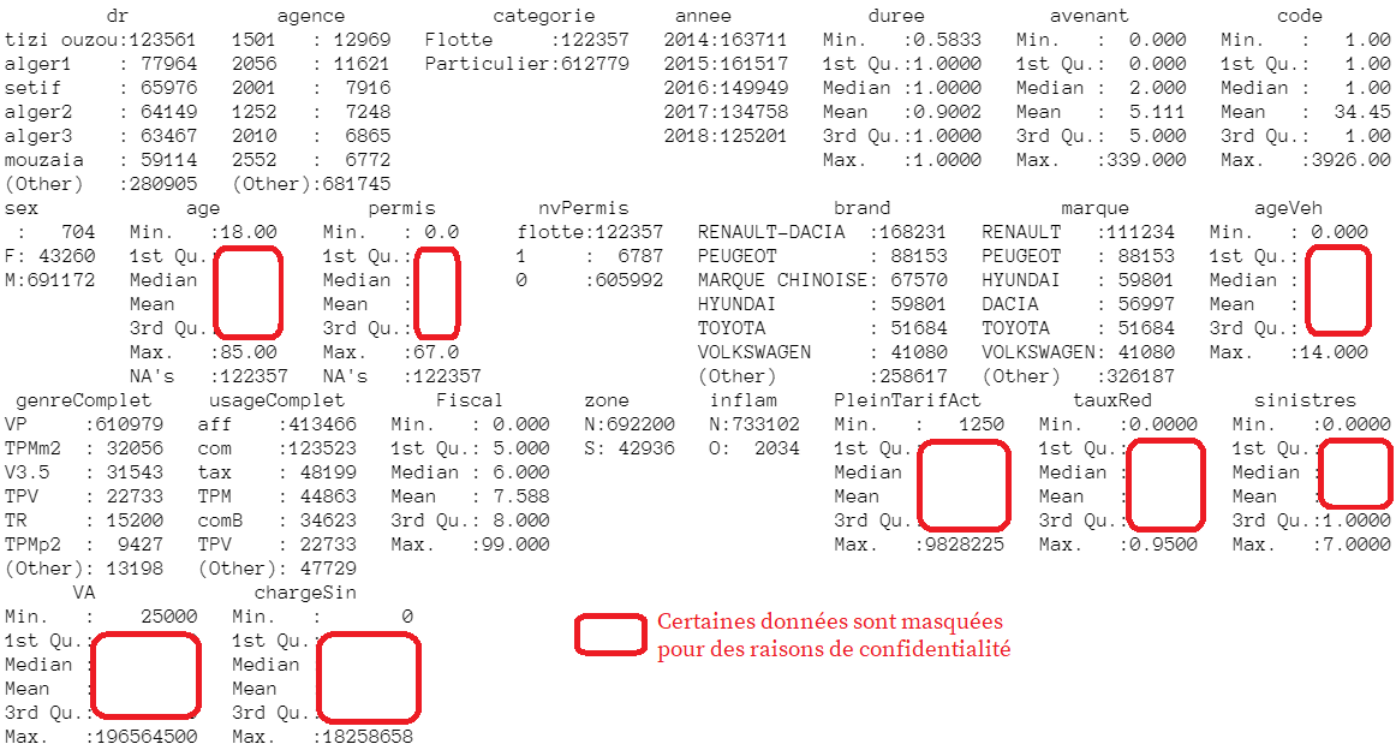



FIGURE 2.1 – Statistiques de la base de données.

En effet, en général, un véhicule acheté dans le but de rouler beaucoup est de type Diesel ou GPL. D'autres interprétations pourraient être faites en fonction des marques et genres de véhicules.

— `kilométrage` : le kilométrage du véhicule pourrait renseigner sur la manière dont est utilisé le véhicule et son ancienneté. De plus, si cette valeur est renseignée sur au moins deux années consécutives, il serait possible de déterminer le nombre de km que le véhicule parcourt sur une période d'exposition.

— `enfantPlusDe18` : variable intéressante permettant de savoir si un assuré a au moins un enfant de plus de 18 ans. Il est très rare qu'un jeune conducteur avec un nouveau permis souscrive un contrat d'assurance en son nom. Généralement, les jeunes conducteurs utilisent le véhicule des parents et aucune information sur la base actuelle ne permet de les repérer.

— `situationMaritale` : On pourrait penser qu'en Algérie, les véhicules sont assurés au nom du mari au sein d'un couple même si celui ci est utilisé par les deux membres du couple, d'où la non consistance de la variable sexe. De plus la variable sexe est également biaisée par le fait que le gouvernement ait accordé des crédits ANSEJ à des femmes qui n'utilisent pas le véhicule assuré en leur nom. C'est pour cette raison que la base de données contient anormalement des usages ou genres de véhicules liés à un assuré de sexe féminin. Il serait logique et non sexiste de supposer qu'en Algérie, il est rare de voir une femme au volant d'un semi remorque de plusieurs tonnes.

— `typeSinistre` : une variable qui pourrait être très importante selon moi pour bien modéliser la gravité des sinistres. En effet, dans les bases de données des assureurs étrangers, la séparation des sinistres en véhicules à réparer ou véhicules réformés (le montant du versement à l'assuré est la valeur assurée du véhicule moins quelques frais) est faite pour connaitre la proportion des véhicules à réparer ou totalement détruits. En Algérie, seule

une description du sinistre est faite mais chaque dossier en possède une différente. Cette variable pourrait servir à déterminer la proportion des petits sinistres, moyens sinistres et gros sinistres. La définition de la gravité de ces sinistres pourrait faire l'objet d'une étude à elle seule.

— `retraitPermis` : variable difficile à obtenir en Algérie mais pourrait aider à repérer les conducteurs à risque.

— `valeurVenale` : Selon moi, la variable `VA` correspondant à la valeur assurée du véhicule devrait être remplacée par une valeur vénale déterminée selon le marché. En effet, après une exploration des données, la base contient beaucoup d'observations telles que des véhicules identiques possèdent des valeurs assurées très différentes, ce qui est anormal. De plus, comme la prime est calculée en fonction de cette valeur assurée, un assuré pourrait de plein gré demander une police d'assurance avec une très faible valeur assurée afin de payer une petite prime. Il y a bien sûr des conséquences à cela mais en ce qui nous concerne pour l'étude de la gravité des sinistres, la qualité de cette donnée est importante et ne doit pas être biaisée par de tels phénomènes.

### 2.2.3 Préparation des données

Comme dit précédemment, le fichier qui nous a été remis a déjà subi un gros travail de consolidation et de nettoyage. Cependant il reste encore des observations aberrantes ou manquantes [1] qu'il faudrait traiter. Nous allons passer en revue le travail supplémentaire mené sur la base de données.

**Durée d'exposition** – la SAA permet aux assurés de souscrire des contrats Tous Risques pour des durées de 3, 6 ou 12 mois. Seuls les contrats de 6 et 12 mois ont été conservés car les contrats de 3 mois sont peu nombreux, mal saisis et très sinistrés, ce qui pourrait grandement affecter la qualité des modèles. De plus, si on suppose que la fréquence des sinistres de chaque individu est un processus de poisson d'intensité $\lambda$, alors un conducteur exposé au risque pendant 12 mois devrait avoir une sinistralité deux fois plus importante qu'un conducteur exposé au risque pendant 6 mois. Par exemple, si un assuré a subit 2 sinistres durant les 12 derniers mois, on pourrait dire qu'il subi un sinistre tous les 6 mois. Cependant, les données de notre base ne reflètent pas cette logique et nous voyons une fréquence de sinistralité plus élevée sur les contrats de 6 mois si on ramène ces contrats à 12 mois. En effet, les contrats de 6 mois sont en moyenne 9% plus sinistrés que ceux d'un an, ce qui correspond en réalité à une durée d'exposition au risque de 7 mois plutôt que 6. L'interprétation de cette remarque reste à faire mais nous avons pris l'initiative de transformer les contrats de 6 mois en 7 mois et il en sort une meilleure modélisation.

**Variable** `age` – l'étendu de la variable `age` est $[18; 114]$. Il est bien évidemment très rare si ce n'est impossible de rencontrer un individu de plus de 110 conduire un véhicule. Il serait donc préférable d'éliminer tout contrat dont l'âge de l'assuré est supérieur à 85 ans, car 98% des contrats sont souscrits par des assurés de moins de 85 ans.

**Variable** `sinistres` – correspondant au nombre de sinistres reportés pendant la période d'exposition. C'est une valeur entière allant de 0 à 18. Il y a 600 polices d'assurances avec plus de 7 sinistres. il est préférable de les éliminer.

**Variable** `genre` – pour faire une bonne segmentation, il faudrait s'appuyer sur un nombre conséquent d'observations pour que chaque sous population ait assez d'observations pour pouvoir établir une estimation des sinistres avec un bon niveau de certitude. Ainsi, si par exemple un genre de véhicule (tel que les camions d'ordures) possède moins de 100 observations, il serait illogique de conclure sur la sinistralité de ce genre de véhicule. Ainsi tous les genres ayant moins de 100 observations ont été retirés : 5 genres ont été retirés.

**Prise en considération des inflations de tous les montants** – la base de données contient

1. le package `R–mice` a été très utile pour repérer les valeurs manquantes

toutes les polices d'assurances établies par la SAA de 2014 à 2018. Durant ces 5 années, l'économie algérienne a fortement évolué et l'inflation a eu un fort impact sur le prix des véhicules et les prix des réparations. L'inflation a fluctué entre 3% et 8%. Pour éviter de complexifier cette tâche d'actualisation, il a été retenu un taux d'inflation annuel moyen de $4,5\%$. Par conséquent, les variables `VA, PleinTarifAct, prime_trAct` et `chargeSin` ont été actualisées avec ce taux et bien sûr en prenant en compte l'année de souscription du contrat.

**La valeur assurée des véhicules** `VA` – cette variable est très liée au montant de remboursement d'un sinistre. Premièrement, le montant de l'indemnisation ne pourrait dépasser la valeur assurée inscrite dans le contrat. Ensuite, cette valeur dépend du véhicule et représente supposément sa valeur vénale : si deux véhicules de marques différentes par exemple sont endommagés de la même manière, les montants des réparations ont de fortes chances d'être différents. Etant donnée l'étendue des valeurs assurées allant de 25 000 DA (absurde bien sûr) à plus de 18 millions de DA en valeur non actualisée, il serait préférable de séparer les grosses valeurs du reste des valeurs assurées sachant que ces dernières représentent 90% de tous les contrats si on fixe un seuil à 3 150 000 DA.

**Variable** `marque` – dans le même esprit que le traitement de la variable `genre`, il faudrait éliminer les marques qui ne possèdent pas beaucoup d'observations. Pour cette variable, les 60 premières marques en termes de nombres ont été retenues pour un minimum de 1000 observations par marque.

**Variables** `inflam` et `Fiscal` – 8995 observations dans toute la base ont le champ `Fiscal` égal à 0, ce qui correspond à une puissance fiscale nulle, ce qui est aberrant et non admissible. Ces observations ont été retirées. 2042 observations sont déclarées comme des véhicules transportant des marchandises inflammables, ce qui est insuffisant pour conclure sur ce type de véhicule car ces 2000 observations font partie de plusieurs genres, marques, région, ...
Pour effectuer l'étude, il reste 735 136 observations à notre disposition, ce qui est suffisant pour avoir des résultats convaincants.

### 2.2.4 Contenu de la base de données et ses variables

La base de données contient les informations de toutes les polices d'assurances souscrites auprès de la SAA dans le cadre de la garantie Tous Risques durant les années 2014 à 2018. Avant nettoyage, il y avait au total quelques 850 000 observations. Le fichier qui nous a été remis en comporte 740 000 car toutes les observations où des informations importantes ont été omises ou mal saisies ont dues être retirées. La table contient la liste des polices d'assurances de deux types de catégories de véhicules : les véhicules particuliers ou les véhicules appartenant à des flottes. Il n'est bien entendu pas possible d'étudier ces deux catégories de la même façon car selon une analyse descriptive, ces deux types de catégories présentent une sinistralité très différente, en plus de ne pas contenir les mêmes variables explicatives ; les véhicules des flottes ne sont pas supposés appartenir à des individus à qui l'on pourrait demander un âge, un sexe, un âge de permis, ...

La base de données comporte 40 variables mais ne sont pas toutes significatives. Il sera question de faire une étude sur la fréquence des sinistres puis une étude sur la gravité des sinistres. Ces deux études portent donc sur la variable `sinistres` qui est le nombre de sinistres survenus durant la période d'exposition d'une police d'assurance, et la seconde variable `ChargeSin` qui est le montant que l'assureur a dû payer en totalité à son assuré pour les sinistres survenus durant la période d'exposition.

La liste des variables utiles sont les suivantes :

1. `id` : chaîne de caractères servant de clé d'identification de chaque police d'assurance. Elle est la concaténation du numéro d'agence, du numéro de police, du code d'appartenance à l'une des catégories citées plus haut.
2. `dr` : variable catégorielle. La direction régionale auquel appartient l'agence dans laquelle le contrat d'assurance a été souscrit. Cette variable est très importante car elle nous servira de variable spatiale, permettant une segmentation géographique du portefeuille. Cette variable contient 15 modalités mais on ne retiendra que 14. La modalité `annaba` a été retirée car elle ne contient que 700 observations, ce qui ne représente pas la quantité d'observations réelles qui se situe à 50 000. L'extraction des données de la direction régionale de Annaba est incorrecte. Ceci est problématique car beaucoup des régions du Nord-Est du pays dépendent de cette direction régionale.
3. `agence` : variable catégorielle. Code de l'agence dans laquelle la police d'assurance a été souscrite. Pour l'instant, nous n'avons pas utilisé cette variable car c'est une variable catégorielle et elle possède 592 modalités (592 agences SAA en Algérie en 2018) et cela représenterait un énorme coût en temps de calcul. Il serait judicieux d'utiliser cette variable dans une étude future car on pourrait supposer que la plupart des individus souscrivent une assurance dans une agence près de leur domicile et on pourrait donc peut être associer la zone de circulation du véhicule à la localisation de l'agence (segmentation par Daïras-Communes). Une autre idée serait d'utiliser les codes d'agences des contrats d'assurances pour déduire les densités d'habitations de certaines régions car cette variable peut intuitivement et conceptuellement aider à expliquer la sinistralité.
4. `categorie` : variable catégorielle. Cette variable catégorielle possède deux modalités déjà citées précédemment. Elle différencie les contrats des flottes de ceux des véhicules particuliers. Cette variable permettra de créer deux bases de données sur lesquelles nous travaillerons : `basePartic` et `baseFlotte`
5. `annee` : valeur entière. Représente l'année de souscription du contrat. Etant donné l'instabilité de l'économie algérienne, il sera important d'utiliser cette variable pour actualiser les valeurs des montants des sinistres et des valeurs assurées des véhicules. Plus de détails sur ce point par la suite.
6. `duree` : variable catégorielle à deux niveaux. Période d'exposition au risque d'un contrat : 6 mois ou une année.
7. `police` : chaîne de caractères. Numéro de police du contrat.
8. `avenant` : valeur entière. Représente le numéro d'avenant du contrat. Par définition, un avenant est un : « acte par lequel on modifie les termes d'un contrat ». En assurance, plus précisément un avenant est une révision du contrat pour corriger, apporter des modifications, ajouter ou retirer des garanties ou bien renouveler celui-ci. La variable suivante a été créée à partir de `avenant`
9. `cutAvenant` : variable catégorielle. Variable déduite de la précédente possédant deux modalités : `new` et `old` permettant de différencier les nouvelles polices d'assurances des anciennes. Elle permettra de séparer les nouveaux assurés de l'agence de ceux qui sont à la SAA depuis au moins 1 an, en supposant que la proportion des avenants de modification liés à des rectifications des nouveaux contrats est faible. L'idée derrière la création de cette variable est de savoir s'il existe bien une tendance des nouveaux contrats à être plus sinistrés que ceux des assurés fidèles ou de savoir s'il y a un phénomène de souscription à la garantie Tous Risques dans un seul but : réparer son véhicule. La création de cette variable est intéressante mais mérite une étude à elle seule car la séparation en 2 modalités exclue le cas des contrats RC renouvelés pendant seulement une année pour une Tous Risques.
10. `code` : valeur entière. Code risque. Un contrat d'assurance peut porter sur plusieurs véhicules et le code risque liste ceux-ci, notamment pour les contrats des flottes.

11. `sex` : variable catégorielle. Sexe du souscripteur de la police d'assurance. Variable pas très intéressante comme on le verra plus tard dans l'étude. L'interprétation sera donnée plus tard. Variable indisponible dans `BaseFlotte`.
12. `age` : variable continue. Age du souscripteur du contrat. Variable indisponible dans `BaseFlotte`.
13. `permis` : variable continue. Ancienneté du permis en années. Variable indisponible dans `BaseFlotte`.
14. `nvPermis` : variable catégorielle. Variable permettant de séparer les assurés avec un nouveau permis ou non. Variable indisponible dans `BaseFlotte`. Cetta variable peut être déduite de la précédente en séparant les anciennetés de permis `0` des autres.
15. `marque` : variable catégorielle avec 113 modalités. Elle représente la marque du véhicule.
16. `brand` : variable catégorielle à 56 modalités représentant les marques mais regroupées selon une certaine logique. Par exemple les marques `RENAULT` et `DACIA` sont regroupées en un groupe `RENAULT-DACIA`.
17. `ageVeh` : variable continue. Représente l'âge du véhicule en années.
18. `genre` : variable catégorielle avec 12 modalités. Elle représente le genre de véhicule ; exemples : Transport Publique de Marchandises, Véhicules de plus de 3, 5 tonnes, Tracteurs Routiers, ...
19. `usage` : variable catégorielle avec 10 niveaux. L'usage auquel le véhicule est supposé être assuré ; exemples : affaire, fonctionnaire, location, taxi, ...
20. `Fiscal` : variable continue. Puissance fiscale du véhicule.
21. `zone` : variable catégorielle à deux modalités `nord` et `sud`. Variable permettant de séparer les véhicules sensés être utilisés dans le nord ou dans le sud du pays.
22. `inflam` : variable catégorielle à deux modalités permettant de séparer les véhicules transportant des marchandises inflammables ou non.
23. `tauxRed` : variable numérique. Représente un taux de réduction appliqué sur la prime demandée pendant la souscription.
24. `PleinTarifAct` : variable continue. Représente la prime Tous Risques avant application du taux de réduction. Ce montant est actualisé selon l'année de souscription du contrat.
25. `prime_trAct` : variable continue représentant le montant de la prime exact payé par l'assuré pour la garantie Tous Risques. Réduction appliquée et montant actualisé.
26. `sinistres` : valeur entière. Variable sur laquelle portera une des études dans ce mémoire. Représente le nombre de sinistres reportés durant la période d'exposition d'un contrat.
27. `VA` : variable continue. Représente la valeur assurée du véhicule. Ce montant est demandé par l'assuré pendant la souscription du contrat. Ce montant a été actualisé.
28. `ChargeSin` : variable continue. Deuxième variable sur laquelle portera l'étude. Représente le montant total des déboursements qu'a dû faire l'assureur sur un contrat concernant uniquement la garantie Tous Risques. Valeur actualisée.

**Conclusion** – Après avoir passé en revue toutes les informations contenues dans la base de données, les modifications qui y ont été apportées ainsi que les potentielles variables qui pourraient aider à mieux expliquer la sinistralité, il serait temps maintenant de commencer à traiter le problème lui même en commençant par la présentation de quelques outils théoriques.

## 2.3 Les chiffres de l'assurance en Algérie

En 2009, le marché algérien de l'assurance était à 65% détenu par l'état. Cependant, malgré l'ouverture au marché de ce secteur d'activité aux autres compagnies privées, ce chiffre semble

stagner. En 2020, la proportion est de 72% et représente un montant conséquent de 95,7 milliards de dinars sur un total de 132 milliards de dinars.

En 2019, Swiss RE, une compagnie de réassurance suisse a présenté son rapport SIGMA 4/2019 [30] sur les chiffres de l'assurance à l'échelle mondiale. Ainsi, le chiffre d'affaire du marché mondial de l'assurance se porte à 6 300 milliards de dollars US, ce qui représente 7,2% du PIB mondial. On définit le taux de pénétration comme le ration Primes / PIB qui constitue un indicateur de développement de l'assurance dans un pays. Dans certains pays il dépasse les 15%. Cependant, celui ci n'est que de 0,72% en Algérie, ce qui représente un taux très faible pour ce secteur.

| Pays | Taux de pénétration | Densité de l'assurance |
|---|---|---|
| Taïwan | 20,88% | 5161 $ |
| Royaume-Uni | 10,61% | 4503 $ |
| France | 8,89% | 3667 $ |
| Etats-Unis | 7,14% | 4481 $ |
| Maroc | 3,88% | 127 $ |
| Tunisie | 2,14% | 75 $ |
| Algérie | 0,68% | 28 $ |

TABLE 2.1 – Taux de pénétrations et densités d'assurance par habitant de certains pays en 2008.[1]

En algérie, les nouvelles garanties ne sont que très récentes et la majorité des chiffres d'affaires des assureurs étant toujours en majorité réalisés par les branches automobiles et IARD (Incendie, Accidents et Risques Divers).

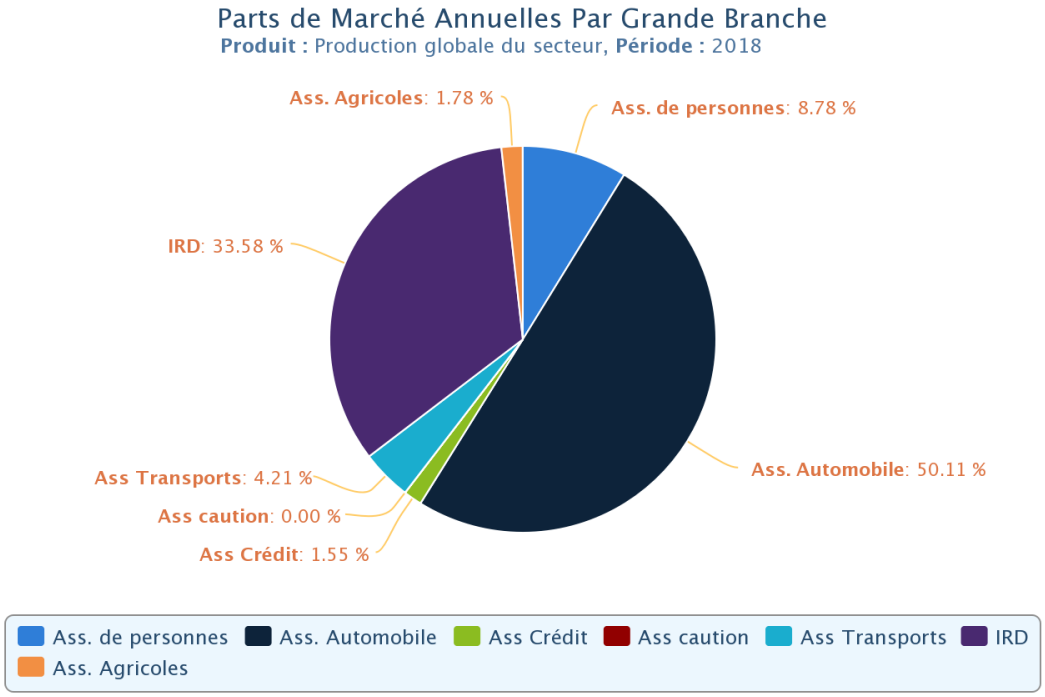



FIGURE 2.2 – Parts de marché annuelles par grande branche en 2018.[12]

De plus, en conséquence des évènements ayant conduit à la nationalisation de l'activité assurance en Algérie après son indépendance, les compagnies nationales ont des portefeuilles spécifiques en lien avec la raison de leur création originale. En effet, le portefeuille de la Société Nationale d'Assurance SAA par exemple est en grande partie composé de l'assurance automobile, tandis que celui de la CAAR est majoritairement composé des assurances gros risques et de transport. Les risques agricoles sont en grande partie assurés par la CNMA (Caisse Nationale de Mutualité Agricole). Cette tendance s'est maintenue depuis la création de ces compagnies d'assurance mais du fait de la nécessité de diversification des garanties, cette prédisposition des compagnies à des marchés spécifiques tend à disparaître.

## 2.4 Organisme d'accueil : la SAA et sa place dans le marché

La rédaction de ce mémoire a été faite durant un passage à la Direction Générale de la SAA dans le contexte d'un stage pratique. Nous avons pu découvrir comment fonctionne les assurances algériennes et également beaucoup appris sur la mise en place et le suivi des garanties, spécialement dans la branche automobile. Nous tenons à remercier infiniment les membres de la division automobile qui nous ont donné l'accès à leur base de données Tous Risques pour nous permettre de mener à bien l'étude sur la tarification des assureurs algériens.

### 2.4.1 Histoire de la SAA

En 1963, la Société Nationale d'Assurance [32] voit le jour en tant que compagnie générale d'assurance sous la marque SAA et le premier point de vente ouvre ses portes à Alger-Centre. En Mai 1966, le monopole de l'état algérien sur les opérations d'assurance conduit à la nationalisation de la SAA. En janvier 1976, la SAA se spécialise dans la branche des risques simples en développant des offres adaptées aux particuliers, aux professionnels, aux collectivités locales et institutions relevant du secteur de la santé. En 1989, la SAA transforme son mode de gouvernance et devient une entreprise publique économique (EPE) avec un capital de 80 millions de DA. La SAA élargit son champ d'activités en 1990 aux risques industriels, du transport, risques agricoles et assurances de personnes. Le marché de l'assurance algérien en pleine expansion a nécessité une certaine modernisation. Ainsi, la SAA a revu l'organisation de son réseau d'agences en partant du principe performance/meilleure rémunération. En 2004, une réorganisation structurelle est réalisée au sein de la compagnie en créant une division par segment de marché afin de booster la productivité. En 2011, son capital social atteint 20 milliards de DA puis à 30 milliards en 2017.

### 2.4.2 Parts de marché de la SAA

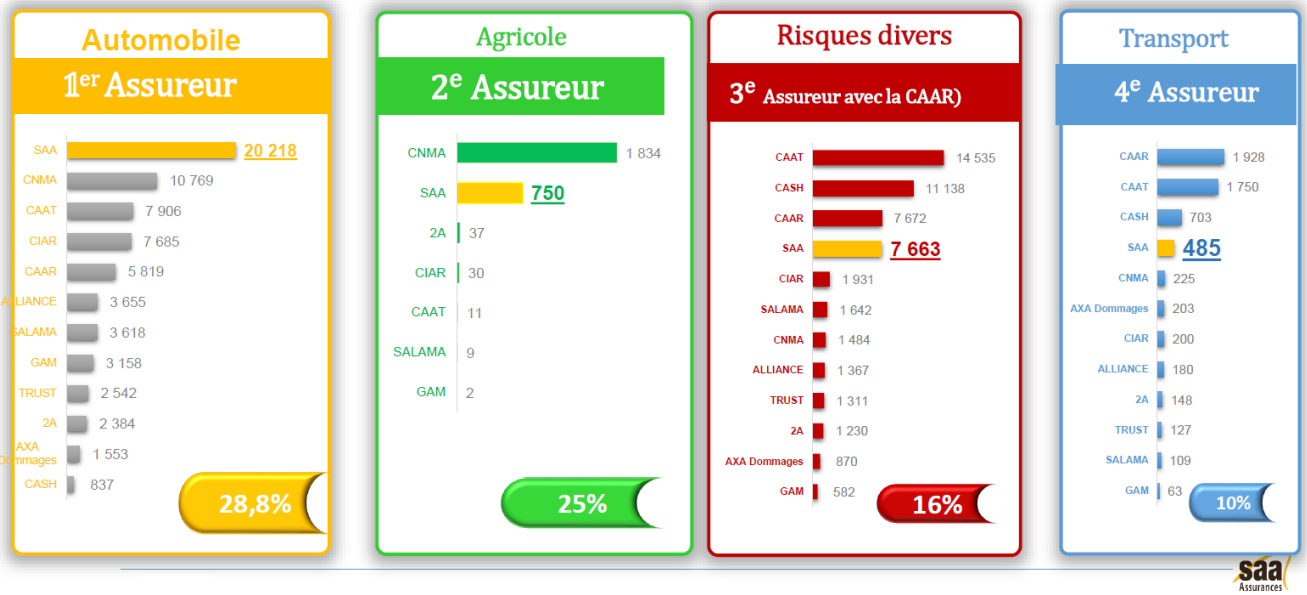



FIGURE 2.3 – Positionnement de la SAA sur le marché d'assurance en 2018. Montants en millions de DA.[34]

| Branche | SAA | Croissance | Secteur | Part de la SAA |
|---|---|---|---|---|
| Automobile | 20 038 | +2,07% | 69 021 | 29,03% |
| IRD | 6 453 | +7,67% | 45 867 | 13,07% |
| Assurance Agricole | 614 | +24,78% | 2 624 | 24,82% |
| Assurance Transport | 489 | +20,53% | 5 828 | 8,38% |
| Assurance Crédit | 85 | +1 597% | 2 144 | 3,96% |
| Total Ass. Dommages | 27 679 | +4,34% | 126 095 | 21,95% |

TABLE 2.2 – Parts de marché de la SAA en 2018 par branche en millions de DA et croissance par rapport à 2017.[33]

Comme on peut le voir sur la figure 4.5, la SAA est le premier assureur automobile en Algérie avec une assez grande marge d'avance sur le second. De plus, grâce à ses stratégies de diversification, la SAA a réussi à s'imposer également sur les autres secteurs.

### 2.4.3 Organigramme de la SAA

J'ai eu le plaisir et l'honneur de faire mon stage dans la branche production de la division automobile au sein de la SAA. Le direction de la production automobile se charge de mettre en place de nouveaux produits d'assurance auto et de porter un suivi sur ceux-ci pour évaluer leur rentabilité. De plus, la tarification et les conventions sont élaborées dans ce service.

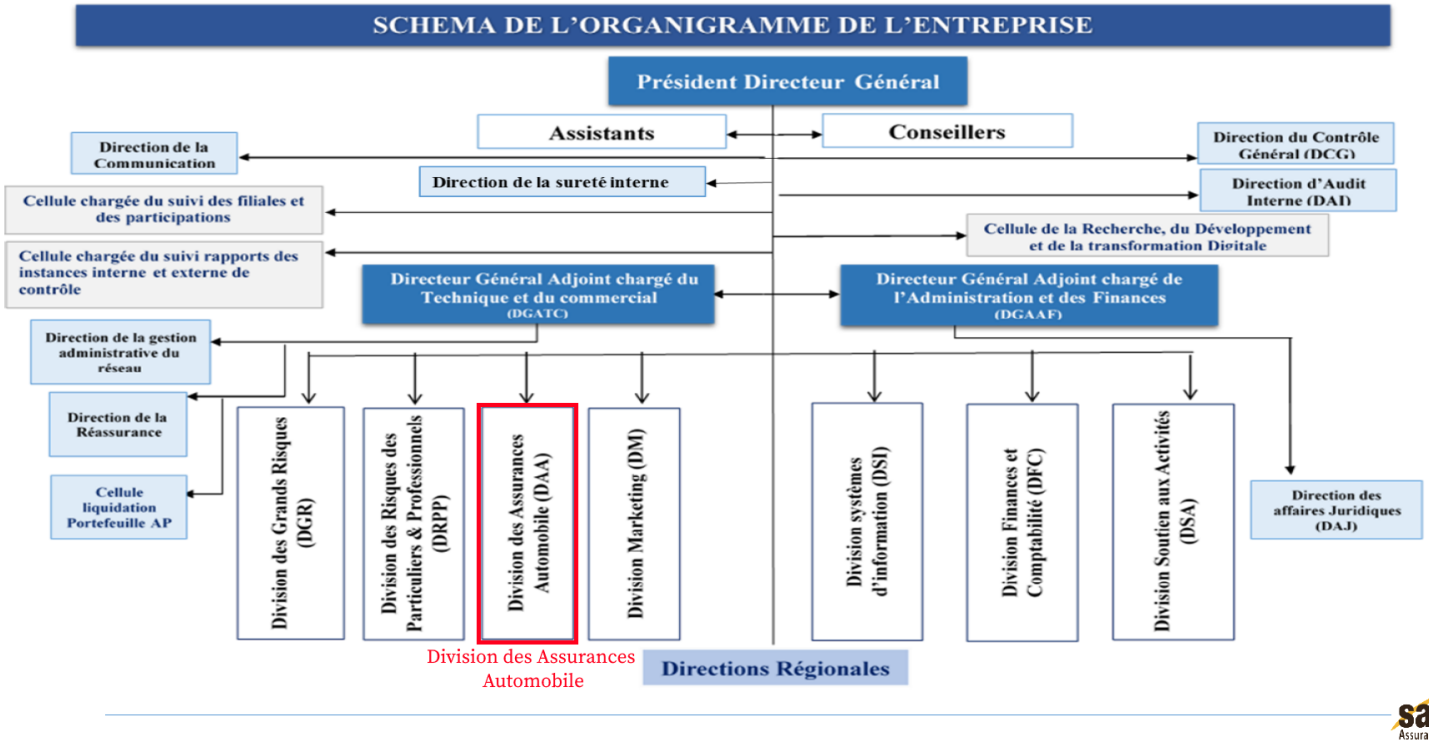



FIGURE 2.4 – Schéma de l'organigramme de la SAA.[34]

## 2.5 Contexte actuel économique de l'assurance automobile en Algérie

### 2.5.1 Situation des assurances pour l'année 2020

Si un secteur d'activité n'est pas aussi rentable dans un pays comparé au reste du monde, il ne relève pas toujours entièrement de la responsabilité des acteurs de ce secteur d'activité. A l'instar des chiffres précédemment cités, l'assurance en Algérie ne représente qu'une faible proportion du PIB algérien mais cela est d'abord dû à la culture du pays, à la façon dont est menée cette activité mais elle est également étroitement liée à l'économie du pays. En effet, les hydrocarbures représentent la principale rentrée en devises du pays et il s'en suit que l'économie algérienne dépend entièrement de ses ventes en ressources fossiles. Ainsi, lorsque les cours de ces dites ressources subissent d'importantes baisses, tous les secteurs d'activités en Algérie en ressentent les effets sur leurs finances. De la macroéconomie à la microéconomie, tout est imbriqué.[13]
Le marché des assurances ne déroge pas à cette règle et aurait pu lui aussi subir une forte régression. Fort heureusement, les assureurs ont eu l'initiative de diversifier leurs produits pour maintenir une certaine rentabilité de l'activité. Cependant, la situation exceptionnelle née du mouvement social et politique de février 2019 et de la crise sanitaire en 2020 dont les retombées n'ont pas fini de faire des dégâts sur l'économie algérienne ne fera qu'empirer l'état fragile de l'assurance en Algérie.
Les statistiques des notes de conjoncture de la CNA [9] sont là pour confirmer ces faits.

| EN DA | CHIFFRE D'AFFAIRES | | STRUCTURE DU MARCHÉ | | ÉVOLUTION | |
|---|---|---|---|---|---|---|
| | 31/12/2019 | 31/12/2020 | 2019 | 2020 | En % | En valeur |
| Assurance Automobile | 69 195 082 014 | 62 805 521 360 | 52,3% | 50,0% | -9,2% | -6 389 560 654 |
| IRD | 51 698 674 198 | 52 368 904 142 | 39,1% | 41,7% | 1,3% | 670 229 944 |
| Assurances Agricoles | 2 684 518 677 | 2 207 908 660 | 2,0% | 1,8% | -17,8% | -476 610 017 |
| Assurance Transport | 6 374 762 750 | 6 047 824 193 | 4,8% | 4,8% | -5,1% | -326 938 557 |
| Assurance Crédit | 2 286 152 672 | 2 079 731 873 | 1,7% | 1,7% | -9,0% | -206 420 799 |
| Total | 132 239 190 311 | 125 509 890 228 | 100% | 100% | -5,1% | -6 729 300 083 |

FIGURE 2.5 – Production des assurances de dommages en fin 2020 par branche

La branche «automobile» enregistre, à fin 2020, un chiffre d'affaires de 62,8 milliards de DA. Comparativement aux 69,2 milliards de DA de 2019, il y a eu donc une régression de 9,2 %.

### 2.5.2 Rentabilité de l'assurance automobile

Le diagramme suivant [10] montre la répartition des revenues de la branche automobile par garantie.

On peut clairement voir que la majeur partie des rentrées de liquidités en automobile proviennent des différentes garanties facultatives (hors RC) proposées par les assureurs – 76,1 %

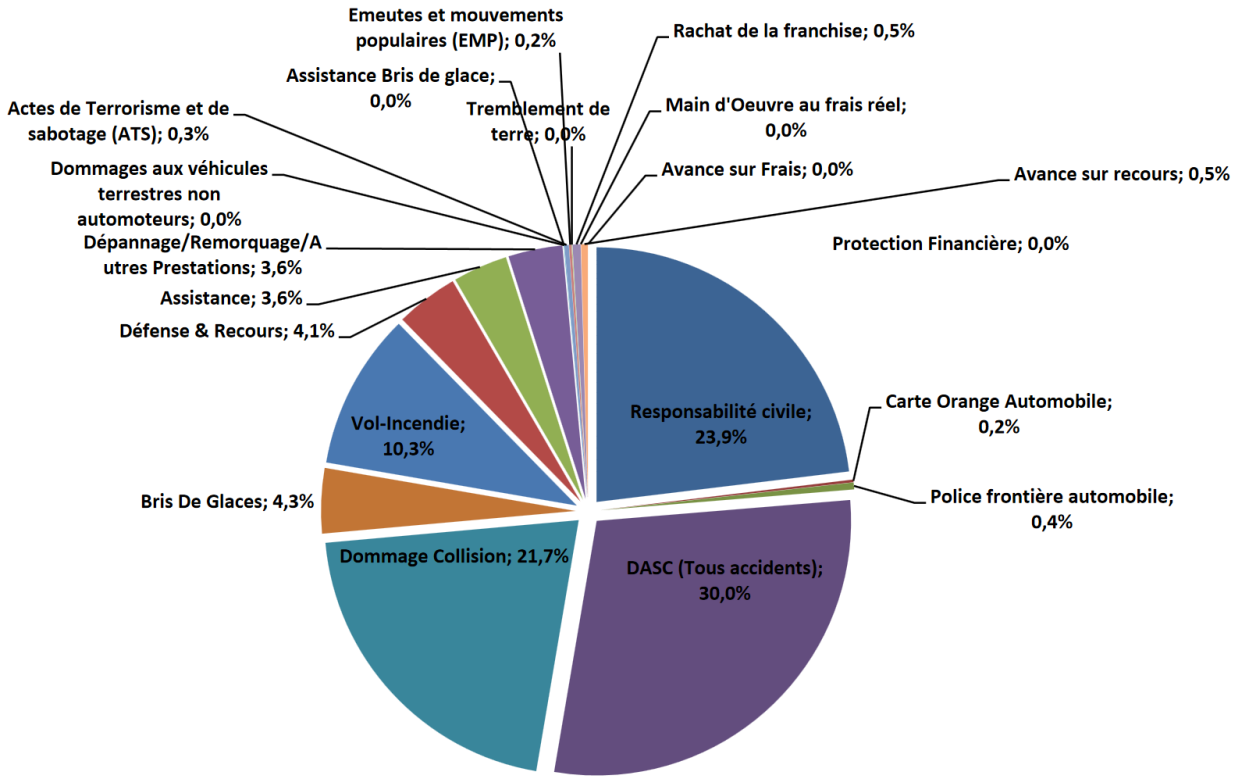



FIGURE 2.6 – Production de la branche automobile en 2019 par garantie

d'entre elles. Et plus particulièrement, 51,7 % proviennent des garanties Tous Risques ou Dommage Collision.
En Algérie, la tarification de la RC ne dépend pas des assureurs. La réglementation oblige les assureurs à tous pratiquer le même système tarifaire pour la responsabilité civile. Il serait intéressant d'étudier les chiffres de la RC mais pas dans le but de rentabiliser cette garantie.
Cependant, les assureurs peuvent pratiquer des prix différents pour les garanties alternatives et c'est sur cela qu'il est possible d'améliorer la situation économique d'un assureur et surtout rester compétitif dans un marché clairement déloyal.

**Remarque** – Il est à noter que depuis la nationalisation de l'activité d'assurance, le gouvernement a tenté de garder l'esprit social en tarifant la garantie obligatoire RC à partir du salaire minimum algérien. Ce tarif a été instauré il y a plus d'une quarantaine d'années et aujourd'hui, la prime moyenne est de 1500 da et sachant que l'Algérie enregistre un très fort taux de sinistralité, les charges sinistres en RC à payer dépassent de loin les primes récoltées : **23,9 % de production contre 56,1 % des sinistres à payer en 2019 [10]** Pour compenser la garantie RC, les assureurs utilisent donc les garanties plus rentables et c'est là où la diversification des garanties instaurées par les compagnies d'assurance algériennes a induit des profits et résultats intéressants.

## 2.6 Problématique et motivations de ce mémoire

**Point Important 1** – L'étude menée dans ce mémoire concerne la garantie Tous Risques mais il est clairement possible d'appliquer la même méthodologie pour estimer les sinistres de chaque garantie et déterminer un tarif adéquat pour chacune d'elle. Pour la majorité des individus voulant souscrire un contrat d'assurance, un des facteurs les plus importants de décision pour le choix de contracter ou non une garantie dommage est le tarif de celle-ci. Le système actuel ne

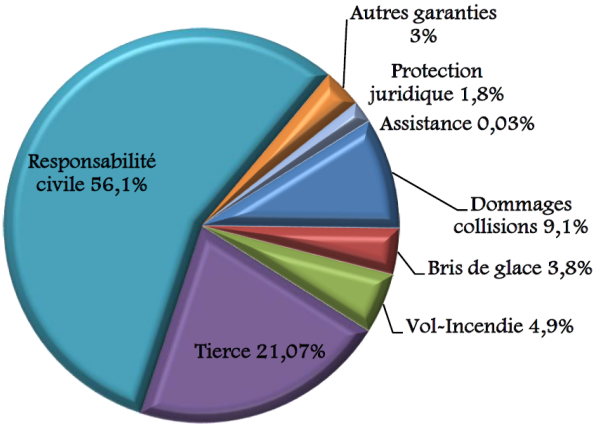



FIGURE 2.7 – Répartition des sinistres à payer en 2019

tient compte que de la valeur assurée du véhicule pour déterminer la prime d'un contrat mais il serait logique de penser que cela pourrait énormément affecter l'intention d'un assuré à prendre une garantie Tous Risques. Le tableau 4.6 simule et compare quelques primes de garanties Tous Risques en fonction du véhicule et de sa valeur assurée.

| Véhicule | Valeur du véhicule | Prime Tous Risques | **Prime totale demandée** |
|---|---|---|---|
| Dacia Sandero | 2 950 000 DA | 147 500 DA | **221 965 DA** |
| Seat Leon | 7 000 000 DA | 350 000 DA | **579 931 DA** |
| Renault Clio | 4 200 000 DA | 210 000 DA | **313 046 DA** |
| Peugeot Expert | 4 800 000 DA | 240 000 DA | **356 606 DA** |
| Kia Rio | 3 500 000 DA | 175 000 DA | **261 895 DA** |

TABLE 2.3 – Tableau des primes qu'auraient payé des assurés particuliers. Les véhicules choisis sont des modèles fréquents en Algérie. [2]

En 2021, la valeur d'un véhicule neuf est d'au moins 2,5 millions de dinars, et la prime Tous Risques correspondante à cette VA est de 125 000, sans compter les différentes taxes dont la TVA. Avec un salaire moyen en Algérie de 40 000 da en 2020 [28], et un niveau de vie qui ne cesse de diminuer à cause de la situation économique algérienne, ce genre de tarification même pour un véhicule de moyenne gamme pourrait en repousser plus d'un, en plus de devoir être payé en une seule fois lors de la souscription d'un contrat et en liquide (facteur ayant tendance à repousser lorsqu'il s'agit de payer un service non matériel – on s'imagine mal se rendre dans une agence d'assurance avec une importante somme pour payer une assurance – Donc solutions : diminuer **stratégiquement** les primes aux individus à bas risques **et** proposer le paiement en ligne ou en plusieurs fois.).
Certes, pour remédier à la cherté de cette garantie, les assureurs ont eu recours à l'attribution de réductions mais cette stratégie ne se base sur aucune étude statistique pour l'affectation des différents taux de réduction et il est fort probable que beaucoup d'assurés qui ne bénéficient d'aucune réduction paient une prime complète alors qu'elle n'est pas proportionnelle au risque auquel ces assurés sont exposés. Inversement, il y a une forte probabilité qu'il y ait attribution de réductions à des groupes d'assurés qui sont très sinistrés et donc ne devraient pas bénéficier de forts taux de réduction. La concurrence déloyal dans le marché algérien joue sur ce taux de

2. Les données des valeurs du véhicules ont été tirées de Ouedkniss et il semble légitime de croire que c'est une référence convaincante concernant la valeur des véhicules.

réduction pour attirer de nouveaux assurés, même si aucune étude n'est menée sur les risques encourus et les pertes potentielles dues à l'attribution de fortes réductions.

**Point Important 2** – En plus du caractère couteux des garanties dommages, il est important de considérer l'évolution du marché automobile algérien car les garanties facultatives en dépendent fortement.
En effet, d'après les données mises à ma disposition, 87 % des véhicules assurés en Tous Risques sont des véhicules de 5 ans ou moins [3].

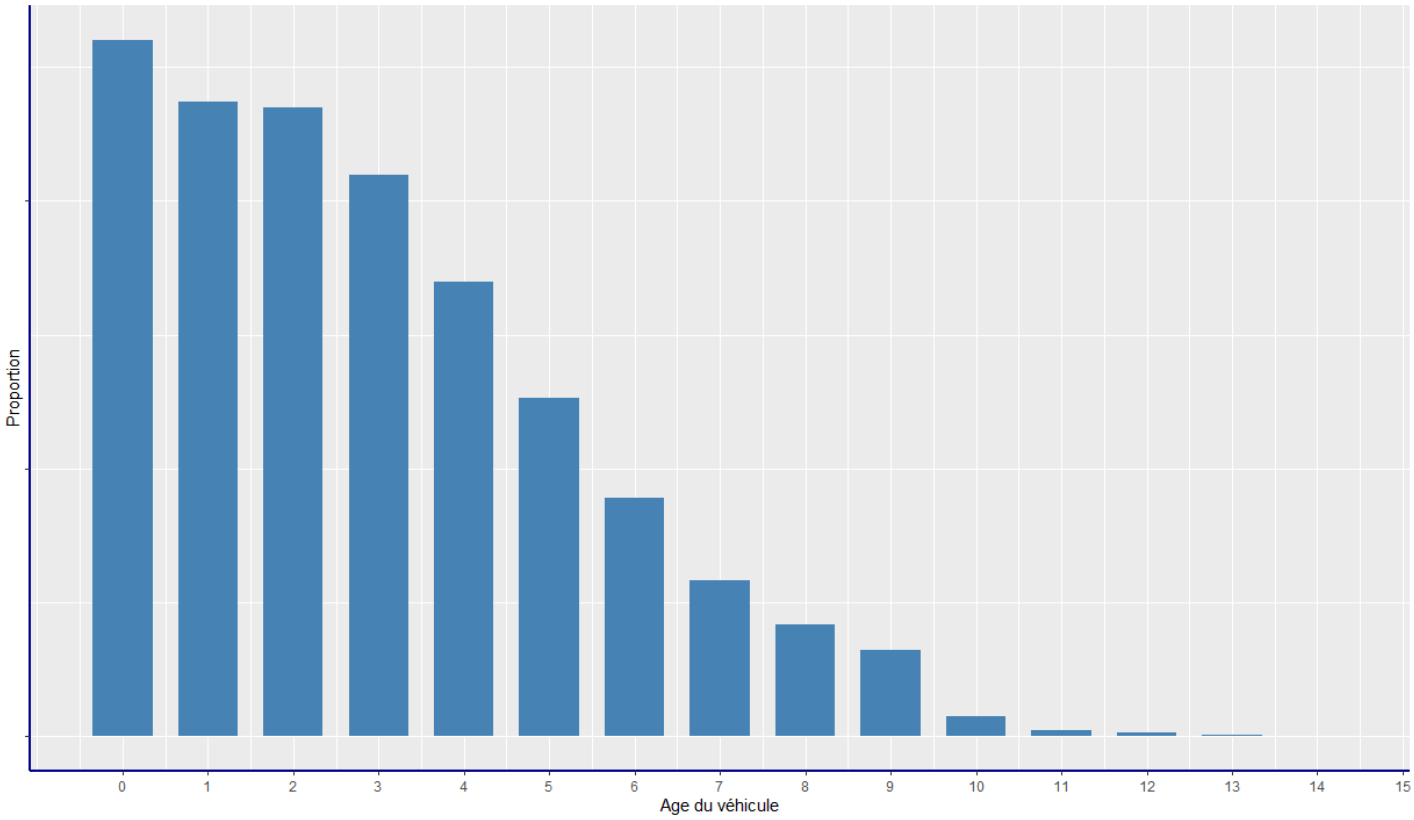



FIGURE 2.8 – Distribution des contrats Tous Risques en fonction de l'âge du véhicule

On en déduit que la rentabilité de cette garantie dépend fortement du marché du véhicule neuf en Algérie. Et comme soulevé précédemment à maintes reprises, la situation économique en Algérie se détériore et le marché des véhicules neufs algériens a subit un énorme déclin. L'importation des véhicules neufs est à l'arrêt depuis 2015. Le soit disant montage des véhicules a débuté peu après et le marché a gardé sa forme pendant quelques années. Depuis février 2019, toutes les usines de montages sont à l'arrêt et il devient très difficile pour les algériens de se procurer un véhicule neuf et cela a eu une répercussion direct sur l'économie des assurances. Seuls les véhicules importés par les licences Moudjahidines sont aujourd'hui en vente mais ceux-ci sont rares et très couteux pour un algérien moyen (plus de 5 millions de DA en général). De plus, le portefeuille des assureurs auto et surtout la rentabilité de ce portefeuille dépend fortement des véhicules de petites et moyennes gammes (80 % des véhicules assurés entre 2014 et 2018 ont des valeurs assurées de moins de 2 millions de DA [4]).

3. Dans la catégorie des véhicules particuliers
4. Selon les données mises à ma disposition

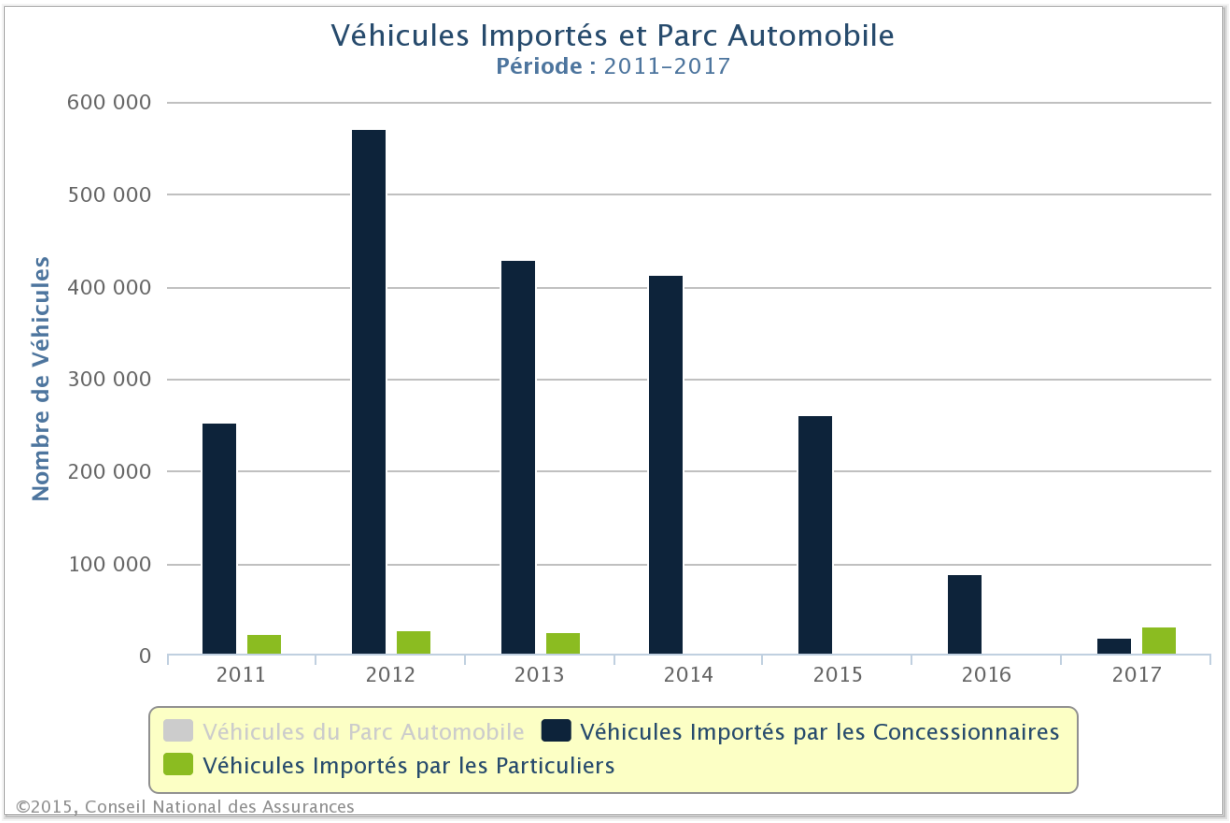



FIGURE 2.9 – Evolution du nombre de véhicules importés par les concessionnaires et par les particuliers Moudjahidines.[11]

Les données disponibles dans la Base de Données Centralisée des Statistiques (BDCS) s'arrêtent à l'année 2017 pour les informations concernant l'importation des véhicules, mais selon un article de Algérie-eco, environ 8000 véhicules ont été importés par les particuliers durant le premier trimestre 2021, et si on ramène ce chiffre à l'année, cela donne environ 32 000 véhicules par an. Si on utopise le fait que tous ces véhicules souscrivent une assurance Tous Risques, cela reste difficile de concevoir que cela suffise à garder la rentabilité des portefeuilles automobiles des assureurs. Pour aller plus loin, il serait intéressant de recueillir les données sur les nombres de sinistres (qui sont très élevés en Algérie) et les recouper avec le faible nombre de véhicules importés afin de suivre l'évolution du patrimoine automobile assurable.

**Conclusion** – Si la situation ne s'améliore pas et que le marché des véhicules neufs ne redécolle pas, les assureurs vont voir les rentabilités de leurs garanties dommages fortement baisser à l'horizon 2023. Etant donné que les actions que peuvent mener les assureurs sont limitées pour remédier à la crise du marché automobile, il ne leur reste qu'à trouver des stratégies afin de fidéliser leurs assurés et essayer de les garder dans leur portefeuille Tous Risques ou Dommages Collision plus longtemps. A notre sens, la révision de la tarification aura un impact majeur sur la stratégie de fidélisation de l'assuré à la garantie facultative. D'où l'objet de ce mémoire.

# Chapitre 3

# Analyse statistique des données

**Introduction** – Dans ce chapitre, nous allons présenter le contenu de la base de données et passer en revue toutes les manipulations qui ont été faites sur celle-ci afin de préparer au mieux les données pour une bonne modélisation de la sinistralité automobile du portefeuille de la SAA. Nous allons également proposer quelques variables (informations) supplémentaires qui pourraient s'avérer utiles pour mieux décrire la sinistralité en Algérie.

## 3.1 Objectif de l'étude

Le but de cette étude est d'essayer de modéliser au mieux la sinistralité automobile algérienne. Cette étude permettra une ébauche d'un système de tarification de la garantie Tous Risques mais pourra également servir de base pour les autres types de garanties dommages car le but sera de déterminer la prime pure [36] d'un contrat d'assurance. La prime pure correspond au montant moyen d'un sinistre que devra payer la compagnie d'assurance si le risque survenait. Le calcul de la prime pure a pour but d'évaluer, pour chaque assuré, le montant attendu des sinistres pour la période d'assurance étudiée. Cette évaluation se fait le plus souvent par des méthodes statistiques. La sinistralité est divisée en plusieurs composantes, chacune étant évaluée indépendamment :

— La probabilité d'un sinistre normal.

— Le coût d'un sinistre normal.

— La probabilité d'un sinistre grave.

— Le coût d'un sinistre grave.

S'il est possible de modéliser avec précision le potentiel futur sinistre d'un individu, il est également possible d'utiliser cette information pour le provisionnement d'un dossier sinistre lors de son ouverture. Cette information pourrait être utilisée plus généralement dans les PSAP (Provisions Sinistres A Payer).

## 3.2 Base de données et préparation des données

Les données qui ont été mises à notre disposition ont d'abord subi une consolidation pour n'en faire qu'une seule table, puis ont subi un nettoyage. Malheureusement, un fait inhérent aux systèmes d'informations algériens est la mauvaise construction des bases de données et des informations qu'elles contiennent. Il est clair et admis de tout le monde que la qualité de saisie des données n'est malheureusement pas optimale dans l'administration algérienne, et cela affectera grandement l'étude car les fausses données créent du biais et peuvent mener à de fausses conclusions. Il est donc primordial avant toute étude de mener une exploration des données pour chercher et traiter ces mauvaises données. Pour cette étape, il n'y a malheureusement pas de méthode universelle ayant une certaine efficacité dans le traitement des données manquantes,

mal saisies ou aberrantes. Cette partie de l'étude est en général la plus coûteuse en temps car il n'est pas possible de savoir vraiment ce que l'on doit chercher comme mauvaises informations. Ensuite, il s'agira de décider quoi faire de ces mauvaises données. La solution la plus facile serait bien sûr de les éliminer, mais il serait préférable d'essayer de trouver un moyen d'expliquer ou de corriger ces données car elles pourraient appartenir à une sous population qui ne possèdent pas beaucoup d'observations. Une autre option serait d'attribuer une valeur moyenne aux valeurs manquantes ou une valeur aléatoire de la variable en question. De plus la modification d'une donnée revient à supposer la survenue d'un évènement qui n'a pas forcément eu lieu et la suppression revient à supposer que l'évènement n'a pas eu lieu. Si ce travail de récupération est mené sur beaucoup de données, il serait nécessaire de procéder à un traitement automatique des données à réviser.

Dans notre étude, par défaut de temps, nous avons dû éliminer toute observation paraissant aberrante.

# Chapitre 4

# Modélisation linéaire

Nous allons commencer par donner brièvement quelques notions sur la modélisation linéaire avant de parler de modèles généralisés.
La régression linéaire permet le traitement de la fluctuation d'une variable par celui d'une ou plusieurs autres variables. Le modèle linéaire classique, ou modèle linéaire simple, constitue la base des Modèles Linéaires Généralisés (ou GLM : Generalized Linear Models en anglais) et une compréhension approfondie est essentielle pour maîtriser les GLM. Beaucoup de concepts de régression trouvés dans les GLM ont leur genèse dans le modèle linéaire simple. Les distributions des variables réponses rencontrées dans le monde des assurances sont généralement non normales et c'est pour cette raison que les GLM existent. Mais la compréhension de ceux-ci nécessitent des outils que nous verrons dans ce chapitre mais qui ne sont pas utilisables directement en assurance.

## 4.1 Définition

On appelle **modèle linéaire** un modèle statistique qui peut s'écrire sous la forme [8] :

$$Y = \sum_{j=1}^{k} \beta_j X^j + \epsilon.$$

On définit les quantités qui interviennent dans ce modèle :

- $Y$ est une variable aléatoire réelle (v.a.r) que l'on observe et que l'on souhaite expliquer et/ou prédire ; on l'appelle **variable à expliquer** ou **variable réponse** ; On suppose que la variance de $Y$ est constante : c'est ce qu'on appelle l'hypothèse d'homoscédasticité.
- les $k$ variables $X^1, \ldots, X^k$ sont des variables réelles ou discrètes, non aléatoires et également observées ; l'écriture de ce modèle suppose que l'ensemble des $X^j$ est sensé expliquer $Y$ par une relation de cause à effet ; les variables $X^j$ sont appelées **variables explicatives, prédicteurs** ou **régresseurs**.
- les $\beta_j, j = 1, \ldots, k$ sont les paramètres du modèle, non observés, et donc à estimer par des techniques statistiques appropriées.
- $\epsilon$ est le terme d'erreur dans le modèle ; c'est une v.a.r non observée pour laquelle on pose les hypothèses suivantes :
$$\mathbb{E}(\epsilon) = 0, \quad Var(\epsilon) = \sigma^2 > 0.$$
où $\sigma^2$ est un paramètre inconnu à estimer. Une hypothèse sera faite plus tard sur ce point.
- Les hypothèses posées sur $\epsilon$ impliquent les caractéristiques suivantes sur $Y$ :

$$\mathbb{E}(Y) = \sum_{j=1}^{k} \beta_j X^j, \quad Var(Y) = \sigma^2.$$

En moyenne, $Y$ s'écrit donc comme une combinaison linéaire des $X^j$ : la liaison entre les $X^k$ et $Y$ est linéaire. C'est la raison pour laquelle ce modèle est appelé **modèle linéaire.**

## 4.2 Critère de qualité

En réalité, lorsqu'on fait une régression linéaire, on cherche une certaine fonction $f$ telle que :

$$Y \approx f(X).$$

Pour faire une régression, il faudrait définir un critère quantifiant la qualité de l'ajustement de $Y$ par la fonction $f$ sur les données. On suppose que $f$ est une fonction inconnue que l'on cherche et qui est dans une classe de fonctions $\mathcal{G}$. Le problème mathématique s'écrit alors sous cette forme :

$$\underset{f \in \mathcal{G}}{\operatorname{argmin}} \sum_{i=1}^{n} l(y_i - f(x_i)),$$

avec :

— $n$ représente le nombre de données

— $l$ est la **fonction de coût** ou la **fonction de perte**.

La fonction de coût la plus utilisée est la fonction **coût quadratique** : $l(u) = u^2$.

## 4.3 Modèle de régression multiple

### 4.3.1 Généralités

Un modèle de régression multiple sera dorénavant noté sous forme vectorielle [14] :

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1}.$$

avec :

— $Y$ est un vecteur aléatoire de dimension $n$

— $n$ nombre d'observation de la variable aléatoire $Y$

— $X$ est une matrice de taille $n \times p$ connue, appelée matrice du **plan d'expérience** ou matrice **design**. $X$ est la concaténation des $p$ variables $X_j$ : $X = (X_1, X_2 \ldots, X_p)$. Nous noterons la $i^e$ ligne de la matrice $X$ par le vecteur ligne $x_i^{'} = (x_{i1}, \ldots, x_{ip})$

— $\beta$ est le vecteur de dimension $p$ des paramètres inconnus du modèle

— $\epsilon$ est le vecteur centré, de dimension $n$, des erreurs.

Pour pouvoir modéliser une variable aléatoire $Y$, il va falloir faire des suppositions sur la matrice $X$. La première consistera à supposer que la matrice $X$ est de plein rang. On notera cette hypothèse $\mathcal{H}_1$. Dans la plupart des situations et spécialement dans la notre, le nombre d'observations $n$ est supérieur au nombre des paramètres à estimer $p$, alors le rang de la matrice $X$ sera égal à $p$.

Il arrive que parmi les variables explicatives $X_j$, certaines peuvent interagir entre elles. Il est par exemple naturel de penser que la variable `densité d'habitation` soit liée à la variable `localisation`. Pour représenter mathématiquement cette interaction, nous écrivons le produit entre les variables explicatives qui interagissent.

Ce type d'écriture reste un cas de régression linéaire. En effet, nous pouvons considérer que n'importe quelle transformation connue et fixée des variables explicatives (log, exp, produit,

...) rentre dans le modèle de régression linéaire. Ceci est important car pour le traitement des variables continues ou numériques plus généralement, il est souvent question de transformer les variables pour obtenir de meilleures approximations de $\mathbb{E}(Y)$.

### 4.3.2 Estimation des paramètres - Moindres Carrés Ordinaires MCO

On appelle estimateur des (**MCO**) $\widehat{\beta}$ de $\beta$ la valeur suivante :

$$\widehat{\beta} = \underset{\beta_1,\ldots,\beta_p}{\operatorname{argmin}} \sum_{i=1}^{n} \left( y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \|Y - X\beta\|^2.$$

**Théorème 1** (Estimateur des MCO [14]). *Si l'hypothèse $\mathcal{H}_1$ est vérifiée, alors l'estimateur des moindres carrés ordinaires $\widehat{\beta}$ de $\beta$ vaut :*

$$\widehat{\beta} = (X'X)^{-1}X'Y.$$

### 4.3.3 Approche géométrique

Il est intéressant de considérer les vecteurs des observations dans l'espace des variables $\mathbb{R}^n$. Géométriquement, le vecteur $Y$ définit dans $\mathbb{R}^n$ un vecteur $\overrightarrow{OY}$ d'origine $O$ et d'extrémité $Y$ avec pour coordonnées $(y_1, \ldots, y_n)$. La matrice $X$ du plan d'expérience est formée de $p$ vecteurs colonnes. Chaque vecteur $X_j$ définit dans $\mathbb{R}^n$ un vecteur $\overrightarrow{OX_j}$ d'origine $O$ et d'extrémité $X_j$. Ce vecteur a pour coordonnées $(x_{1j}, \ldots, x_{nj})$. Ces $p$ vecteurs linéairement indépendants (hypothèse $\mathcal{H}_1$) engendrent un sous-espace vectoriel de $\mathbb{R}^n$ que l'on notera $\mathfrak{F}(X)$, de dimension $p$.

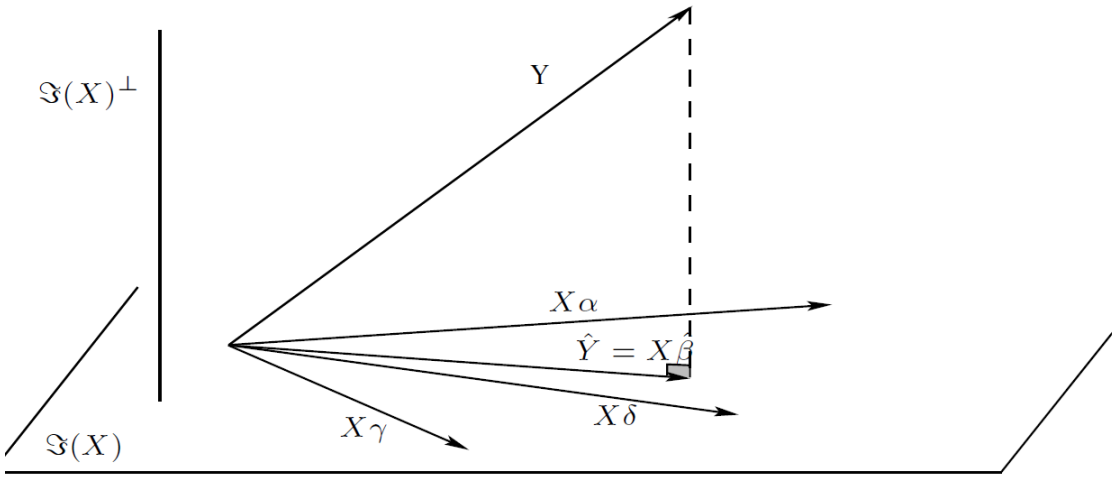



FIGURE 4.1 – Représentation graphique dans l'espace des variables.[14]

L'espace $\mathfrak{F}(X)$ est engendré par les colonnes de $X$. L'espace orthogonal à $\mathfrak{F}(X)$ est noté $\mathfrak{F}(X)^{\perp}$ est quant à lui appelé espace des résidus.

Minimiser $S(\beta) = \sum_{i=1}^{n} \left( y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$ revient à chercher un élément de $\mathfrak{F}(X)$ qui soit le plus proche de $Y$, au sens de la norme euclidienne classique. Par définition, cet unique élément noté $\widehat{Y}$ est appelé projection orthogonale de $Y$ sur $\mathfrak{F}(X)$ noté :

$$\widehat{Y} = P_X Y = X\widehat{\beta}.$$

La matrice $P_X$ est la matrice de projection orthogonale sur $\mathfrak{F}(X)$ ou **hat matrix** et $\widehat{\beta}$ est l'estimateur des moindres carrés de $\beta$. Le vecteur $\widehat{Y}$ contient les valeurs ajustées de $Y$ par le modèle.

**Proposition 1.** *[14] L'estimateur des moindres carrés $\widehat{\beta}$ est un estimateur sans biais de $\beta$ et sa variance vaut $Var(\widehat{\beta}) = \sigma^2(X'X)^{-1}$*

**Théorème 2** (Gauss-Markov [14]). *L'estimateur des MCO est optimal parmi les estimateurs linéaires sans biais de $\beta$.*

### 4.3.4 Résidus

Les résidus sont définis par la relation suivante :

$$\widehat{\epsilon} = Y - \widehat{Y} = (I - P_X)Y = P_{X^\perp}Y.$$

Les résidus appartiennent donc à $\mathfrak{F}(X)^\perp$. On peut réécrire les résidus sous la forme suivante :

$$\widehat{\epsilon} = P_{X^\perp}Y = P_{X^\perp}(X\beta + \epsilon) = P_{X^\perp}\epsilon.$$

Ici $\epsilon$ est appelé le terme d'erreur ou résidus. La terminologie **erreur** est trompeuse : il n'y a aucune incidence que la déviation de Y du modèle théorique est de quelque manière que ce soit erronée. En économétrie, le terme **perturbation** est utilisé.

D'autres hypothèses sur la modélisation sont imposées pour pouvoir faire une régression linéaire. On admettra donc que :

- **Homoscédasticité** – Hypothèse $\mathcal{H}_2$ : la variance de $\epsilon$ est finie et ne varie pas avec les variables $x$, $Var(\epsilon) = \sigma^2$
- **Normalité et indépendance** – Hypothèse $\mathcal{H}_3$ : La distribution de $\epsilon$ est normale. De plus, les observations $y_i$ sont indépendantes. On écrira $\epsilon \sim \mathcal{N}(0, \sigma^2)$

**Proposition 2.** *[14] Sous les hypothèses $\mathcal{H}_1$ et $\mathcal{H}_2$, nous avons :*

$$\begin{array}{ll} \mathbb{E}(\widehat{\epsilon}) = P_{X^\perp}(\epsilon) = 0, & Var(\widehat{\epsilon}) = \sigma^2 P_{X^\perp}, \\ \mathbb{E}(\widehat{Y}) = X\beta, & Var(\widehat{Y}) = \sigma^2 P_X, \\ Cov(\widehat{\epsilon}, \widehat{Y}) = 0. & \end{array}$$

Il est également possible de compléter la figure 5.1 pour visualiser les résidus géométriquement.

Le théorème de Pythagore donne directement l'égalité suivante :

$$\|Y - \overline{y}\,\mathbb{1}\|^2 = \|\widehat{Y}\|^2 + \|\widehat{\epsilon}\|^2.$$

On définit le coefficient de détermination multiple $R^2$ comme suit :

$$R^2 = \frac{\|\widehat{Y} - \overline{y}\,\mathbb{1}\|^2}{\|Y - \overline{y}\,\mathbb{1}\|^2} = \cos^2\theta = \frac{\text{Variation expliquée par le modèle}}{\text{Variation totale}}.$$

Le coefficient de détermination est une mesure permettant de quantifier la qualité d'un modèle. Plus elle se rapproche de 1, mieux le modèle modélise la variable à expliquer $Y$. Ce coefficient ne tient pas compte de la dimension de $\mathfrak{F}(X)$, ce qui est problématique car l'ajout de variables augmente le $R^2$ mais augmente également la complexité du modèle. (*cf Coefficient de détermination ajusté*). Nous verrons plus tard d'autres mesures de la qualité d'un modèle.

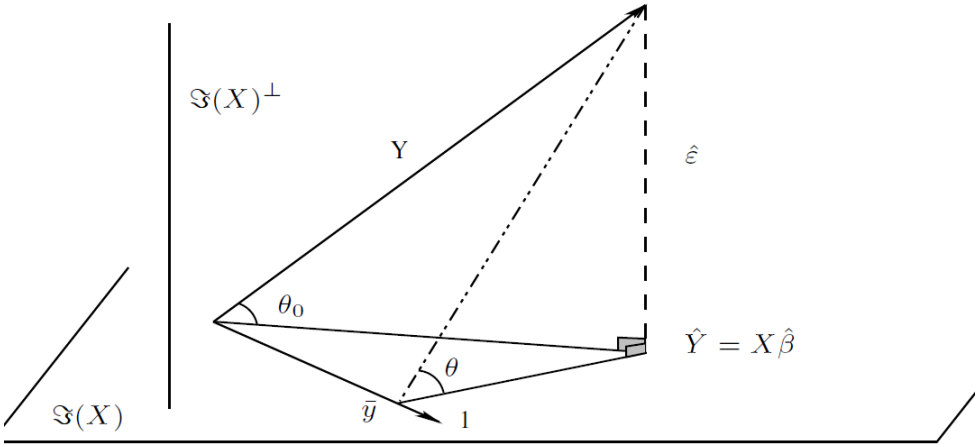



FIGURE 4.2 – Représentation graphique des résidus dans l'espace des variables.[14]

**Proposition 3.** *[14] La statistique que l'on appelle* ***somme des carrés résiduels*** *ou* ***SCR*** *définit un estimateur sans biais de* $\sigma^2$

$$\widehat{\sigma}^2 = \frac{\|\widehat{\epsilon}\|^2}{n-p} = \frac{SCR}{n-p}.$$

Ce dernier estimateur nous permet alors de créer un estimateur de la variance de $\widehat{\beta}$

$$\widehat{\sigma}^2_{\widehat{\beta}} = \widehat{\sigma}^2 (X'X)^{-1} \Rightarrow \widehat{\sigma}_{\widehat{\beta}_j} = \sqrt{\widehat{\sigma}^2 [(X'X)^{-1}]_{jj}}.$$

### 4.3.5 Prévision

Un des plus importants buts de la régression est de pouvoir proposer des prévisions pour la variable à expliquer $y$ lorsque nous avons de nouvelles valeurs de $x$. Soit une nouvelle valeur que l'on notera $x'_{n+1} = (x'_{n+1,1}, x'_{n+1,2}, \ldots, x'_{n+1,p})$. Nous souhaitons alors prédire la valeur de $y$. On a alors alors :

$$y_{n+1} = x'_{n+1}\beta + \epsilon_{n+1},$$

avec $\mathbb{E}(\epsilon_{n+1}) = 0, \quad V(\epsilon_{n+1}) = \sigma^2$ et $Cov(\epsilon_{n+1}, \epsilon_i) = 0$ pour $i = 1, \ldots, n$.

Il est possible de prédire la valeur correspondante grâce au modèle ajusté

$$\widehat{y} = x'_{n+1}\widehat{\beta},$$

Cependant, la variance de la prévision n'est pas égale à celle des $\widehat{y}$. En effet, deux erreurs viennent perturber l'exactitude de la prévision, la première due à l'incertitude sur $\epsilon_{n+1}$ et l'autre à l'incertitude due à l'estimation. Calculons la variance de l'erreur de prévision :

$$\mathrm{V}\left(y_{n+1} - \hat{y}^p_{n+1}\right) = \mathrm{V}\left(x'_{n+1}\beta + \varepsilon_{n+1} - x'_{n+1}\hat{\beta}\right) = \sigma^2 + x'_{n+1} V(\hat{\beta}) x_{n+1} = \sigma^2 \left(1 + x'_{n+1}\left(X'X\right)^{-1} x_{n+1}\right).$$

Ce résultat permet clairement de voir l'erreur due à $\sigma^2$ à laquelle vient s'ajouter l'incertitude d'estimation.

### 4.3.6 Inférence statistique : tests

Grâce à l'hypothèse gaussienne $\mathcal{H}_3$ , il est possible d'établir les résultats suivants [14].

**Proposition 4.** *[14] $(\widehat{\beta}, \widehat{\sigma}^2)$ est une statistique complète et $(\widehat{\beta}, \widehat{\sigma}^2)$ est de variance minimale dans la classe des estimateurs sans biais.*

**Proposition 5** (variance connue [14])**.** *Sous l'hypothèse $\mathcal{H}_1$ et $\mathcal{H}_3$, nous avons :*

- *$\widehat{\beta}$ est un vecteur gaussien de moyenne $\beta$ et de variance $\sigma^2(X'X)^{-1}$,*
- *$(n-p)\widehat{\sigma}^2/\sigma^2$ suit un $\chi^2_{n-p}$ à $n-p$ ddl,*
- *$\widehat{\beta}$ et $\widehat{\sigma}^2$ sont indépendants.*

Or dans la réalité, la variance est rarement connue, on utilise alors son estimateur sans biais $\widehat{\sigma^2}$. On a alors la proposition suivante.

**Proposition 6** (variance inconnue [14])**.** *Sous l'hypothèse $\mathcal{H}_1$ et $\mathcal{H}_3$, nous avons :*

- *pour $j = 1, \ldots, p$, on a : $T_j = \dfrac{\widehat{\beta}_j - \beta_j}{\sqrt{\widehat{\sigma}^2 \left[(X'X)^{-1}\right]_{jj}}} \sim \mathcal{T}(n-p)$,*
- *Soit $R$ une matrice de taille $q \times p$ de rang $q, q \leq p$, alors*

$$\frac{1}{q\widehat{\sigma}^2}(R(\widehat{\beta}-\beta))' \left[R(X'X)^{-1}R'\right]^{-1} R(\widehat{\beta}-\beta) \sim \mathcal{F}_{q,n-p},$$

*$\mathcal{T}$ étant la loi de Student et $\mathcal{F}$ celle de Fischer.*

**Utilisation du test de Student** – En pratique, dans la modélisation [16], il est souvent question de faire les tests d'hypothèses suivants : $H_0 : \beta_j = \beta_j^0$ constituant l'hypothèse nulle contre $H_1 : \beta_j \neq \beta_j^0$ avec $\beta_j^0$ est une valeur supposé de la valeur de $\beta_j$. Alors la statistique

$$T_j = \frac{\widehat{\beta}_j - \beta_j}{\sqrt{\widehat{\sigma}^2 \left[(X'X)^{-1}\right]_{jj}}}.$$

est utilisée pour faire ce test d'hypothèse.
Chaque test de ce genre sur les $j = 1, \ldots, p$ permet de déterminer si réellement la valeur de $\beta_j$ est statistiquement et de façon significative différente de $\beta_j^0$.

Supposons que nous voulons savoir si la variable explicative $x_j$ est significative et est bien corrélée à la variable réponse $y$, alors il suffit de poser $\beta_j^0 = 0$, calculer la statistique de Student et conclure sur la significativité du coefficient $\beta_j$.

**Utilisation du test de Fischer** – Supposons que nous souhaitons tester la nullité simultanée des $q$ derniers coefficients du modèle avec $q \leq p$, le problème s'écrit alors de la façon suivante :

$$H_0 : \beta_{q-p+1} = \cdots = \beta_p = 0 \qquad \text{contre} \qquad H_1 : \exists j \in \{p-q+1, \ldots, p\} : \beta_j \neq 0.$$

Ce test est en réalité un moyen de déterminer s'il n'existe pas un modèle emboité avec $q$ variables explicatives et qui possède le même pouvoir prédictif et le même ajustement que le modèle complet à $p$ régresseurs. On a le théorème suivant pour faire ce test utilisant la statistique de Fischer.

**Théorème 3.** *[14] Soit un modèle de régression à $p$ variables $Y = X\beta + \epsilon$ satisfaisant $\mathcal{H}_1$ et $\mathcal{H}_3$. Nous souhaitons tester la validité d'un sous-modèle (ou modèle emboîté) où un ou plusieurs coefficients sont nuls. Le plan d'expérience privé de ces variables sera noté $X_0$, les $p_0$ colonnes de $X_0$ engendreront un sous-espace noté $\mathfrak{F}_0$ et le sous-modèle sera $Y = X_0\beta_0 + \epsilon$. Notons l'hypothèse nulle (modèle restreint) $H_0 : \mathbb{E}(Y) \in \mathfrak{F}_0$ et l'hypothèse alternative (modèle complet) $H_1 : \mathbb{E}(Y) \in \mathfrak{F}(X)$. Pour tester ces deux hypothèses, nous utilisons la statistique de test F ci-dessous qui possède comme loi sous $H_0$ :*

$$F = \frac{\|\widehat{Y}_0 - \widehat{Y}\|^2/(p - p_0)}{\|Y - \widehat{Y}\|^2/(n - p)} \sim \mathcal{F}_{p-p_0,n-p}.$$

*Il est également possible d'écrire de façon équivalente cette statistique :*

$$F = \frac{n - p}{p - p_0}\frac{SCR_0 - SCR}{SCR} \sim \mathcal{F}_{p-p_0,n-p}.$$

*L'hypothèse $H_0$ sera repoussée en faveur de $H_1$ si l'observation de la statistique F est supérieure à $f_{p-p_0,n-p}(1 - \alpha)$. La valeur $\alpha$ est le niveau du test et $f$ la répartition de la loi de Fischer.*

## 4.4 Choix des variables

Cette étape de la régression est très importante pour tout utilisateur ayant à manipuler plusieurs variables explicatives. En assurance automobile par exemple, la base de données des assureurs est pleine d'informations qui sont peut être inutiles et il est primordial pour tout utilisateur de régression de pouvoir identifier les seules variables qui rentrent dans l'explicabilité de la réponse $y$ afin de réduire la complexité des modèles et peut être diminuer drastiquement les temps de calculs de ceux-ci. Dans notre cas malheureusement, c'est le manque de variables qui affectera la qualité du modèle. En effet, manquer de variables explicatives peut s'avérer plus couteux en termes d'erreurs que d'en prendre certaines qui n'apportent rien au modèle. C'est pour cela que nous avons consacré la section 5.5.2 à proposer de nouvelles informations à demander aux assurés lors de la souscription des contrats.

L'objectif de la sélection des variables est de déterminer au mieux l'ensemble des variables explicatives pertinentes $X_j$ tel que leurs coefficients $j$ soient non nuls dans le modèle.

### 4.4.1 Choix incorrect de variables

Il faut savoir que faire un mauvais choix de variables explicatives peut avoir des conséquences sur la qualité des modèles. Par « mauvais choix », il faut comprendre soit en prendre trop peu, soit en prendre le bon nombre mais pas les bonnes, soit en prendre trop. Admettons que nous ayons trois variables explicatives potentielles $X_1$, $X_2$ et $X_3$ et que le vrai modèle soit :

$$Y = \beta_1 X_1 + \beta_2 X_2 + \epsilon = X_{12}\beta_{12} + \epsilon.$$

La variable $X_3$ n'intervient pas dans le modèle pour prédire $Y$, mais cette information n'est pas disponible. Avec trois variables, nous pouvons donc produire $(2^3 - 1)$ modèles différents, trois modèles à une variable, trois modèles à deux variables et un modèle à trois variables.

### 4.4.2 Biais des estimateurs

L'analyse des sept modèles produit les résultats sur les biais des estimateurs suivants :

| modèle | estimations | propriétés |
|---|---|---|
| $Y_1 = X_1\beta_1 + \varepsilon$ | $\hat{Y}_1 = X_1\hat{\beta}_1$ | $B(\hat{Y}_1) = -P_{X_1^\perp}X_2\beta_2$ |
| | $\hat{\sigma}_1^2 = \frac{\|P_{X_1^\perp}Y\|^2}{n-1}$ | $B(\hat{\sigma}_1^2) = \frac{1}{n-1}\beta_2^2\|P_{X_1^\perp}X_2\|^2$ |
| $Y = X_2\beta_2 + \varepsilon$ | $\hat{Y}_2 = X_2\hat{\beta}_2$ | $B(\hat{Y}_2) = -P_{X_2^\perp}X_1\beta_1$ |
| | $\hat{\sigma}_2^2 = \frac{\|P_{X_2^\perp}Y\|^2}{n-1}$ | $B(\hat{\sigma}_2^2) = \frac{1}{n-1}\beta_1^2\|P_{X_2^\perp}X_1\|^2$ |
| $Y = X_3\beta_3 + \varepsilon$ | $\hat{Y}_3 = X_3\hat{\beta}_3$ | $B(\hat{Y}_3) = -P_{X_3^\perp}X_{12}\beta_{12}$ |
| | $\hat{\sigma}_3^2 = \frac{\|P_{X_3^\perp}Y\|^2}{n-1}$ | $B(\hat{\sigma}_3^2) = \frac{1}{n-1}\beta_{12}'X_{12}'P_{X_{12}^\perp}X_{12}\beta_{12}$ |
| $Y = X_{12}\beta_{12} + \varepsilon$ | $\hat{Y}_{12} = X_{12}\beta_{12}$ | $B(\hat{Y}_{12}) = 0$ |
| | $\hat{\sigma}_{12}^2 = \frac{\|P_{X_{12}^\perp}Y\|^2}{n-2}$ | $B(\hat{\sigma}_{12}^2) = 0$ |
| $Y = X_{13}\beta_{13} + \varepsilon$ | $\hat{Y}_{13} = X_{13}\hat{\beta}_{13}$ | $B(\hat{Y}_{13}) = -P_{X_{13}^\perp}X_{12}\beta_{12}$ |
| | $\hat{\sigma}_{13}^2 = \frac{\|P_{X_{13}^\perp}Y\|^2}{n-2}$ | $B(\hat{\sigma}_{13}^2) = \frac{1}{n-2}\beta_{12}'X_{12}'P_{X_{13}^\perp}X_{12}\beta_{12}$ |
| $Y = X_{23}\beta_{23} + \varepsilon$ | $\hat{Y}_{23} = X_{23}\hat{\beta}_{23}$ | $B(\hat{Y}_{23}) = -P_{X_{23}^\perp}X_{12}\beta_{12}$ |
| | $\hat{\sigma}_{23}^2 = \frac{\|P_{X_{23}^\perp}Y\|^2}{n-2}$ | $B(\hat{\sigma}_{23}^2) = \frac{1}{n-2}\beta_{12}'X_{12}'P_{X_{23}^\perp}X_{12}\beta_{12}$ |
| $Y = X_{123}\beta_{123} + \varepsilon$ | $\hat{Y}_{123} = X_{123}\hat{\beta}_{123}$ | $B(\hat{Y}_{123}) = 0$ |
| | $\hat{\sigma}_{123}^2 = \frac{\|P_{X_{123}^\perp}Y\|^2}{n-3}$ | $B(\hat{\sigma}_{123}^2) = 0$ |

FIGURE 4.3 – Biais des différents estimateurs. [14]

Nous constatons alors que dans les modèles «trop petits» à une variable, c'est-à-dire admettant moins de variables que le modèle « correct » inconnu du statisticien, les estimateurs obtenus sont biaisés. A l'inverse, lorsque les modèles sont « trop grands » (ici à 3 variables), les estimateurs ne sont pas biaisés. Il semblerait donc qu'il vaille mieux travailler avec des modèles « trop grands » afin d'éviter de créer du biais en enlevant des variables explicatives qu'on estimerait inutiles.

### 4.4.3 Variance des estimateurs

Cependant, ce dont nous venons de parler ne concerne que les biais des estimateurs. L'autre aspect de la qualité d'une régression est de mesurer la variance des estimateurs. Le calcul des variances selon les différents modèles donne :

| Modèle | Variance |
|---|---|
| $Y = X_1\beta_1 + \epsilon$ | $V(\widehat{Y}_1) = P_{X_1}\sigma^2$ |
| $Y = X_{12}\beta_{12} + \epsilon$ | $V(\widehat{Y}_{12}) = P_{X_{12}}\sigma^2 = P_{X_1}\sigma^2 + P_{X_2\cap X_1^\perp}\sigma^2$ |
| $Y = X_{123}\beta_{13} + \epsilon$ | $V(\widehat{Y}_{123}) = P_{X_{123}}\sigma^2 = P_{X_1}\sigma^2 + P_{X_{23}\cap X_1^\perp}$ |

**Conclusion** – La variance des données ajustées dans le modèle le plus petit est plus faible que celle des données ajustées dans le modèle le plus grand. C'est pour cette raison que l'utilisateur

de la régression doit "rechercher" le meilleur modèle en évitant de passer à coté de variables explicatives pour éviter de créer du biais, et aussi l'insertion de variables inutiles pour éviter d'augmenter la variance des données ajustées par le modèle.

### 4.4.4 Régression linéaire pondérée

La régression pondérée est une méthode pouvant être utilisée lorsque l'hypothèse de variance constante dans les valeurs résiduelles pour les moindres carrés est contredite (hétéroscédasticité). Avec une pondération adaptée, cette procédure minimise la somme des carrés des valeurs résiduelles pondérés, de manière à générer des valeurs résiduelles présentant une variance constante (on parle aussi d'homoscédasticité).
Ainsi dans la suite $\omega_i$ sera la pondération de chaque observation et $\mathbf{\Omega} = \text{diag}(\omega)$.

Au lieu de $\sum_{i=1}^{n} \left(y_i - \mathbf{x}_i^\top \beta\right)^2$, on considère $\sum_{i=1}^{n} \omega_i \left(y_i - \mathbf{x}_i^\top \beta\right)^2$,

$$W(\boldsymbol{\beta}) = (\mathbf{Y} - \mathbf{X}\beta)^\top \mathbf{\Omega}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}), \quad \mathbf{\Omega} = \text{diag}(\omega),$$

$$\frac{\partial W(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2\mathbf{X}^\top \mathbf{\Omega}\mathbf{Y} + 2\mathbf{X}^\top \mathbf{\Omega}\mathbf{X}\beta,$$

et

$$\frac{\partial^2 W(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} = 2\mathbf{X}^\top \mathbf{\Omega}\mathbf{X}.$$

Aussi

$$\frac{\partial W(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2\mathbf{X}^\top \mathbf{\Omega}\mathbf{Y} + 2\mathbf{X}^\top \Omega\mathbf{X}\boldsymbol{\beta} = \mathbf{0} \text{ si } \widehat{\beta} = \left(\mathbf{X}^\top \Omega\mathbf{X}\right)^{-1} \mathbf{X}^\top \Omega\mathbf{Y},$$

De plus, si on pose maintenant :

$$\Omega^{1/2} = \text{diag}(\sqrt{\omega}),$$
$$\tilde{\mathbf{X}} = \Omega^{1/2}\mathbf{X} \quad \text{et} \quad \tilde{\mathbf{Y}} = \Omega^{1/2}\mathbf{Y},$$

nous avons alors :

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^\top \mathbf{\Omega}\mathbf{X}\right)^{-1} \mathbf{X}^\top \Omega\mathbf{Y} = \left(\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}}\right)^{-1} \tilde{\mathbf{X}}^\top \tilde{\mathbf{Y}}.$$

### 4.4.5 Contenu de la base de données et ses variables

La base de données contient les informations de toutes les polices d'assurances souscrites auprès de la SAA dans le cadre de la garantie Tous Risques durant les années 2014 à 2018. Avant nettoyage, il y avait au total quelques 850 000 observations. Le fichier qui nous a été remis en comporte 740 000 car toutes les observations où des informations importantes ont été omises ou mal saisies ont dues être retirées. La table contient la liste des polices d'assurances de deux types de catégories de véhicules : les véhicules particuliers ou les véhicules appartenant à des flottes. Il n'est bien entendu pas possible d'étudier ces deux catégories de la même façon car selon une analyse descriptive, ces deux types de catégories présentent une sinistralité très différente, en plus de ne pas contenir les mêmes variables explicatives ; les véhicules des flottes ne sont pas supposés appartenir à des individus à qui l'on pourrait demander un âge, un sexe, un âge de permis, ...

La base de données comporte 40 variables mais ne sont pas toutes significatives. Il sera question de faire une étude sur la fréquence des sinistres puis une étude sur la gravité des sinistres. Ces deux études portent donc sur la variable `sinistres` qui est le nombre de sinistres survenus durant la période d'exposition d'une police d'assurance, et la seconde variable `ChargeSin` qui est

le montant que l'assureur a dû payer en totalité à son assuré pour les sinistres survenus durant la période d'exposition.

La liste des variables utiles sont les suivantes :

1. `id` : chaîne de caractères servant de clé d'identification de chaque police d'assurance. Elle est la concaténation du numéro d'agence, du numéro de police, du code d'appartenance à l'une des catégories citées plus haut.
2. `dr` : variable catégorielle. La direction régionale auquel appartient l'agence dans laquelle le contrat d'assurance a été souscrit. Cette variable est très importante car elle nous servira de variable spatiale, permettant une segmentation géographique du portefeuille. Cette variable contient 15 modalités mais on ne retiendra que 14. La modalité `annaba` a été retirée car elle ne contient que 700 observations, ce qui ne représente pas la quantité d'observations réelles qui se situe à 50 000. L'extraction des données de la direction régionale de Annaba est incorrecte. Ceci est problématique car beaucoup des régions du Nord-Est du pays dépendent de cette direction régionale.
3. `agence` : variable catégorielle. Code de l'agence dans laquelle la police d'assurance a été souscrite. Pour l'instant, nous n'avons pas utilisé cette variable car c'est une variable catégorielle et elle possède 592 modalités (592 agences SAA en Algérie en 2018) et cela représenterait un énorme coût en temps de calcul. Il serait judicieux d'utiliser cette variable dans une étude future car on pourrait supposer que la plupart des individus souscrivent une assurance dans une agence près de leur domicile et on pourrait donc peut être associer la zone de circulation du véhicule à la localisation de l'agence (segmentation par Daïras-Communes). Une autre idée serait d'utiliser les codes d'agences des contrats d'assurances pour déduire les densités d'habitations de certaines régions car cette variable peut intuitivement et conceptuellement aider à expliquer la sinistralité.
4. `categorie` : variable catégorielle. Cette variable catégorielle possède deux modalités déjà citées précédemment. Elle différencie les contrats des flottes de ceux des véhicules particuliers. Cette variable permettra de créer deux bases de données sur lesquelles nous travaillerons : `basePartic` et `baseFlotte`
5. `annee` : valeur entière. Représente l'année de souscription du contrat. Etant donné l'instabilité de l'économie algérienne, il sera important d'utiliser cette variable pour actualiser les valeurs des montants des sinistres et des valeurs assurées des véhicules. Plus de détails sur ce point par la suite.
6. `duree` : variable catégorielle à deux niveaux. Période d'exposition au risque d'un contrat : 6 mois ou une année.
7. `police` : chaîne de caractères. Numéro de police du contrat.
8. `avenant` : valeur entière. Représente le numéro d'avenant du contrat. Par définition, un avenant est un : « acte par lequel on modifie les termes d'un contrat ». En assurance, plus précisément un avenant est une révision du contrat pour corriger, apporter des modifications, ajouter ou retirer des garanties ou bien renouveler celui-ci. La variable suivante a été créée à partir de `avenant`
9. `cutAvenant` : variable catégorielle. Variable déduite de la précédente possédant deux modalités : `new` et `old` permettant de différencier les nouvelles polices d'assurances des anciennes. Elle permettra de séparer les nouveaux assurés de l'agence de ceux qui sont à la SAA depuis au moins 1 an, en supposant que la proportion des avenants de modification liés à des rectifications des nouveaux contrats est faible. L'idée derrière la création de cette variable est de savoir s'il existe bien une tendance des nouveaux contrats à être plus sinistrés que ceux des assurés fidèles ou de savoir s'il y a un phénomène de souscription à la

garantie Tous Risques dans un seul but : réparer son véhicule. La création de cette variable est intéressante mais mérite une étude à elle seule car la séparation en 2 modalités exclue le cas des contrats RC renouvelés pendant seulement une année pour une Tous Risques.

10. `code` : valeur entière. Code risque. Un contrat d'assurance peut porter sur plusieurs véhicules et le code risque liste ceux-ci, notamment pour les contrats des flottes.
11. `sex` : variable catégorielle. Sexe du souscripteur de la police d'assurance. Variable pas très intéressante comme on le verra plus tard dans l'étude. L'interprétation sera donnée plus tard. Variable indisponible dans `BaseFlotte`.
12. `age` : variable continue. Age du souscripteur du contrat. Variable indisponible dans `BaseFlotte`.
13. `permis` : variable continue. Ancienneté du permis en années. Variable indisponible dans `BaseFlotte`.
14. `nvPermis` : variable catégorielle. Variable permettant de séparer les assurés avec un nouveau permis ou non. Variable indisponible dans `BaseFlotte`. Cetta variable peut être déduite de la précédente en séparant les anciennetés de permis `0` des autres.
15. `marque` : variable catégorielle avec 113 modalités. Elle représente la marque du véhicule.
16. `brand` : variable catégorielle à 56 modalités représentant les marques mais regroupées selon une certaine logique. Par exemple les marques `RENAULT` et `DACIA` sont regroupées en un groupe `RENAULT-DACIA`.
17. `ageVeh` : variable continue. Représente l'âge du véhicule en années.
18. `genre` : variable catégorielle avec 12 modalités. Elle représente le genre de véhicule ; exemples : Transport Publique de Marchandises, Véhicules de plus de 3, 5 tonnes, Tracteurs Routiers, ...
19. `usage` : variable catégorielle avec 10 niveaux. L'usage auquel le véhicule est supposé être assuré ; exemples : affaire, fonctionnaire, location, taxi, ...
20. `Fiscal` : variable continue. Puissance fiscale du véhicule.
21. `zone` : variable catégorielle à deux modalités `nord` et `sud`. Variable permettant de séparer les véhicules sensés être utilisés dans le nord ou dans le sud du pays.
22. `inflam` : variable catégorielle à deux modalités permettant de séparer les véhicules transportant des marchandises inflammables ou non.
23. `tauxRed` : variable numérique. Représente un taux de réduction appliqué sur la prime demandée pendant la souscription.
24. `PleinTarifAct` : variable continue. Représente la prime Tous Risques avant application du taux de réduction. Ce montant est actualisé selon l'année de souscription du contrat.
25. `prime_trAct` : variable continue représentant le montant de la prime exact payé par l'assuré pour la garantie Tous Risques. Réduction appliquée et montant actualisé.
26. `sinistres` : valeur entière. Variable sur laquelle portera une des études dans ce mémoire. Représente le nombre de sinistres reportés durant la période d'exposition d'un contrat.
27. `VA` : variable continue. Représente la valeur assurée du véhicule. Ce montant est demandé par l'assuré pendant la souscription du contrat. Ce montant a été actualisé.
28. `ChargeSin` : variable continue. Deuxième variable sur laquelle portera l'étude. Représente le montant total des déboursements qu'a dû faire l'assureur sur un contrat concernant uniquement la garantie Tous Risques. Valeur actualisée.

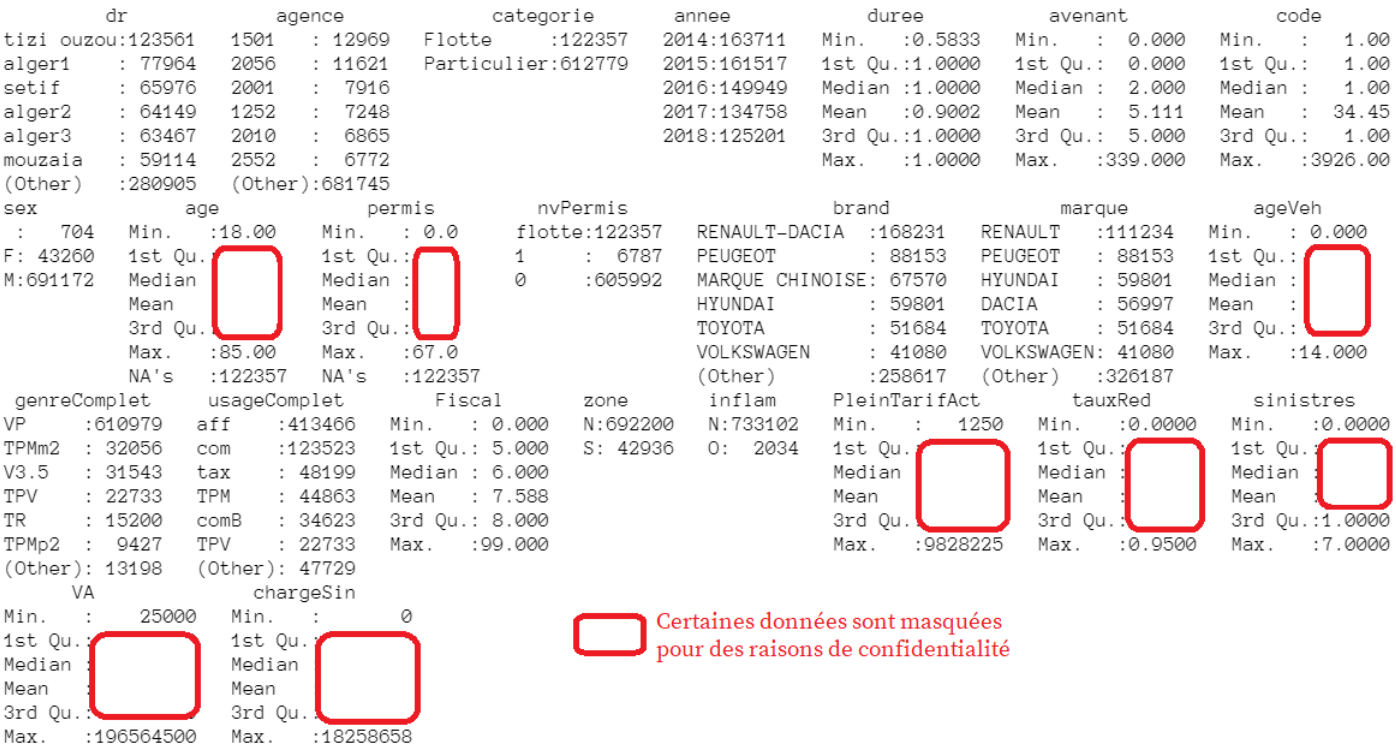



FIGURE 4.4 – Statistiques de la base de données.

### 4.4.6 Variables explicatives manquantes

La liste des variables contenues dans cette base de données est conséquente mais il manque malheureusement de précieuses informations qui à notre sens pourraient aider grandement à mieux prédire la fréquence et la gravite des sinistres. Une liste non exhaustive des variables qui pourraient apporter une amélioration à la modélisation de la sinistralité :

— `localisation` : Déduire la ville dans lequel est supposé circuler le véhicule en fonction de la localité de l'agence dans lequel a été souscrit le contrat. Pour améliorer la qualité de cette donnée, il faudrait faire correspondre également l'usage du véhicule. Car un véhicule assuré à un endroit peut être utilisé pour aller à un travail situé dans une autre ville. Cas fréquent dans le centre-nord du pays. Enfin l'ajout de cette donnée pourra déterminer la densité de population de la région dans laquelle le véhicule devra circuler.

— `profession` : cette variable existe dans les bases de données mais n'a pas été retenue car jugée de mauvaise qualité. Une idée serait d'utiliser la carte grise des véhicules pour déterminer correctement cette donnée.

— `dateSinistre` : la date pourrait renseigner sur une certaine saisonnalité des sinistres notamment sur la fréquence. On pourrait partitionner cette variable en mois ou en trimestre pour déterminer les périodes de l'année les plus sinistrées. Par exemple le mois de ramadan ou la période estivale.

— `energie` : le type de carburant ou énergie utilisée par les véhicules pour fonctionner. Cette donnée existe dans les bases de données de la SAA mais a été jugée inutile. Nous ne partageons pas ce point de vue car elle pourrait donner un indice sur l'utilisation du véhicule. En effet, en général, un véhicule acheté dans le but de rouler beaucoup est de type Diesel ou GPL. D'autres interprétations pourraient être faites en fonction des marques et genres de véhicules.

— `kilométrage` : le kilométrage du véhicule pourrait renseigner sur la manière dont est utilisé le véhicule et son ancienneté. De plus, si cette valeur est renseignée sur au moins deux années consécutives, il serait possible de déterminer le nombre de km que le véhicule parcourt sur une période d'exposition.

— `enfantPlusDe18` : variable intéressante permettant de savoir si un assuré a au moins un enfant de plus de 18 ans. Il est très rare qu'un jeune conducteur avec un nouveau permis souscrive un contrat d'assurance en son nom. Généralement, les jeunes conducteurs utilisent le véhicule des parents et aucune information sur la base actuelle ne permet de les repérer.

— `situationMaritale` : On pourrait penser qu'en Algérie, les véhicules sont assurés au nom du mari au sein d'un couple même si celui ci est utilisé par les deux membres du couple, d'où la non consistance de la variable sexe. De plus la variable sexe est également biaisée par le fait que le gouvernement ait accordé des crédits ANSEJ à des femmes qui n'utilisent pas le véhicule assuré en leur nom. C'est pour cette raison que la base de données contient anormalement des usages ou genres de véhicules liés à un assuré de sexe féminin. Il serait logique et non sexiste de supposer qu'en Algérie, il est rare de voir une femme au volant d'un semi remorque de plusieurs tonnes.

— `typeSinistre` : une variable qui pourrait être très importante selon moi pour bien modéliser la gravité des sinistres. En effet, dans les bases de données des assureurs étrangers, la séparation des sinistres en véhicules à réparer ou véhicules réformés (le montant du versement à l'assuré est la valeur assurée du véhicule moins quelques frais) est faite pour connaitre la proportion des véhicules à réparer ou totalement détruits. En Algérie, seule une description du sinistre est faite mais chaque dossier en possède une différente. Cette variable pourrait servir à déterminer la proportion des petits sinistres, moyens sinistres et gros sinistres. La définition de la gravité de ces sinistres pourrait faire l'objet d'une étude à elle seule.

— `retraitPermis` : variable difficile à obtenir en Algérie mais pourrait aider à repérer les conducteurs à risque.

— `valeurVenale` : Selon moi, la variable `VA` correspondant à la valeur assurée du véhicule devrait être remplacée par une valeur vénale déterminée selon le marché. En effet, après une exploration des données, la base contient beaucoup d'observations telles que des véhicules identiques possèdent des valeurs assurées très différentes, ce qui est anormal. De plus, comme la prime est calculée en fonction de cette valeur assurée, un assuré pourrait de plein gré demander une police d'assurance avec une très faible valeur assurée afin de payer une petite prime. Il y a bien sûr des conséquences à cela mais en ce qui nous concerne pour l'étude de la gravité des sinistres, la qualité de cette donnée est importante et ne doit pas être biaisée par de tels phénomènes.

### 4.4.7 Préparation des données

Comme dit précédemment, le fichier qui nous a été remis a déjà subi un gros travail de consolidation et de nettoyage. Cependant il reste encore des observations aberrantes ou manquantes [1] qu'il faudrait traiter. Nous allons passer en revue le travail supplémentaire mené sur la base de données.

**Durée d'exposition** – la SAA permet aux assurés de souscrire des contrats Tous Risques pour des durées de 3, 6 ou 12 mois. Seuls les contrats de 6 et 12 mois ont été conservés car les contrats de 3 mois sont peu nombreux, mal saisis et très sinistrés, ce qui pourrait grandement affecter la qualité des modèles. De plus, si on suppose que la fréquence des sinistres de chaque individu

1. le package `R–mice` a été très utile pour repérer les valeurs manquantes

est un processus de poisson d'intensité $\lambda$, alors un conducteur exposé au risque pendant 12 mois devrait avoir une sinistralité deux fois plus importante qu'un conducteur exposé au risque pendant 6 mois. Par exemple, si un assuré a subit 2 sinistres durant les 12 derniers mois, on pourrait dire qu'il subi un sinistre tous les 6 mois. Cependant, les données de notre base ne reflètent pas cette logique et nous voyons une fréquence de sinistralité plus élevée sur les contrats de 6 mois si on ramène ces contrats à 12 mois. En effet, les contrats de 6 mois sont en moyenne 9% plus sinistrés que ceux d'un an, ce qui correspond en réalité à une durée d'exposition au risque de 7 mois plutôt que 6. L'interprétation de cette remarque reste à faire mais nous avons pris l'initiative de transformer les contrats de 6 mois en 7 mois et il en sort une meilleure modélisation.

**Variable** `age` – l'étendu de la variable `age` est $[18; 114]$. Il est bien évidemment très rare si ce n'est impossible de rencontrer un individu de plus de 110 conduire un véhicule. Il serait donc préférable d'éliminer tout contrat dont l'âge de l'assuré est supérieur à 85 ans, car 98% des contrats sont souscrits par des assurés de moins de 85 ans.

**Variable** `sinistres` – correspondant au nombre de sinistres reportés pendant la période d'exposition. C'est une valeur entière allant de 0 à 18. Il y a 600 polices d'assurances avec plus de 7 sinistres. il est préférable de les éliminer.

**Variable** `genre` – pour faire une bonne segmentation, il faudrait s'appuyer sur un nombre conséquent d'observations pour que chaque sous population ait assez d'observations pour pouvoir établir une estimation des sinistres avec un bon niveau de certitude. Ainsi, si par exemple un genre de véhicule (tel que les camions d'ordures) possède moins de 100 observations, il serait illogique de conclure sur la sinistralité de ce genre de véhicule. Ainsi tous les genres ayant moins de 100 observations ont été retirés : 5 genres ont été retirés.

**Prise en considération des inflations de tous les montants** – la base de données contient toutes les polices d'assurances établies par la SAA de 2014 à 2018. Durant ces 5 années, l'économie algérienne a fortement évolué et l'inflation a eu un fort impact sur le prix des véhicules et les prix des réparations. L'inflation a fluctué entre 3% et 8%. Pour éviter de complexifier cette tâche d'actualisation, il a été retenu un taux d'inflation annuel moyen de $4,5\%$. Par conséquent, les variables `VA, PleinTarifAct, prime_trAct` et `chargeSin` ont été actualisées avec ce taux et bien sûr en prenant en compte l'année de souscription du contrat.

**La valeur assurée des véhicules** `VA` – cette variable est très liée au montant de remboursement d'un sinistre. Premièrement, le montant de l'indemnisation ne pourrait dépasser la valeur assurée inscrite dans le contrat. Ensuite, cette valeur dépend du véhicule et représente supposément sa valeur vénale : si deux véhicules de marques différentes par exemple sont endommagés de la même manière, les montants des réparations ont de fortes chances d'être différents. Etant donnée l'étendue des valeurs assurées allant de 25 000 DA (absurde bien sûr) à plus de 18 millions de DA en valeur non actualisée, il serait préférable de séparer les grosses valeurs du reste des valeurs assurées sachant que ces dernières représentent 90% de tous les contrats si on fixe un seuil à 3 150 000 DA.

**Variable** `marque` – dans le même esprit que le traitement de la variable `genre`, il faudrait éliminer les marques qui ne possèdent pas beaucoup d'observations. Pour cette variable, les 60 premières marques en termes de nombres ont été retenues pour un minimum de 1000 observations par marque.

**Variables** `inflam` et `Fiscal` – 8995 observations dans toute la base ont le champ `Fiscal` égal à 0, ce qui correspond à une puissance fiscale nulle, ce qui est aberrant et non admissible. Ces observations ont été retirées. 2042 observations sont déclarées comme des véhicules transportant des marchandises inflammables, ce qui est insuffisant pour conclure sur ce type de véhicule car ces 2000 observations font partie de plusieurs genres, marques, région, ...
Pour effectuer l'étude, il reste 735 136 observations à notre disposition, ce qui est suffisant pour avoir des résultats convaincants.

**Conclusion** – Après avoir passé en revue toutes les informations contenues dans la base de données, les modifications qui y ont été apportées ainsi que les potentielles variables qui pourraient aider à mieux expliquer la sinistralité, il serait temps maintenant de commencer à traiter le problème lui même en commençant par la présentation de quelques outils théoriques.

**Introduction** – Dans ce chapitre, nous allons présenter le secteur des assurances en Algérie en donnant d'abord un bref historique de l'activité d'assurance, puis les chiffres de l'assurance dans sa globalité et ensuite les chiffres qui nous intéressent, c'est-à-dire ceux de l'assurance automobile. Enfin, une présentation de l'organisme d'accueil, la société nationale d'assurance SAA, sera faite. Ses chiffres et son positionnement dans le marché des assurances seront également présentés.

## 4.5 Histoire de l'assurance en Algérie

Les besoins en assurance des Algériens durant la période coloniale sont considérés comme insignifiants compte tenu de leurs revenus et de leur situation socioculturelle. L'assurance en Algérie a commencé à se développer au lendemain de l'indépendance mais la majorité des opérations des assurances restaient aux mains des entreprises françaises. Pendant cette période, une grande partie des primes collectées par les compagnies d'assurance ont été transférées à l'étranger. Il aura fallu attendre 1966 pour que l'état reprenne le monopole sur l'assurance en Algérie. L'exploitation de cette activité était désormais réservée à l'état via les entreprises nationales qui avaient vues le jour durant les années 1963-1964.[6] En 1976, les quatre compagnies d'assurance nationales – CAAR, SAA, MAATEC et la CCRMA – se spécialisent chacune dans la couverture d'un certain type de risque. En 1995, une ordonnance supprime le monopole de l'état sur le marché d'assurance, autorisant ainsi la création de compagnies d'assurance privées. Grâce à ce changement, la responsabilité civile n'est plus la seule garantie proposée par les assureurs et permet ainsi une certaine rentabilité de ce secteur.

## 4.6 Les chiffres de l'assurance en Algérie

En 2009, le marché algérien de l'assurance était à 65% détenu par l'état. Cependant, malgré l'ouverture au marché de ce secteur d'activité aux autres compagnies privées, ce chiffre semble stagner. En 2020, la proportion est de 72% et représente un montant conséquent de 95,7 milliards de dinars sur un total de 132 milliards de dinars.

En 2019, Swiss RE, une compagnie de réassurance suisse a présenté son rapport SIGMA 4/2019 [30] sur les chiffres de l'assurance à l'échelle mondiale. Ainsi, le chiffre d'affaire du marché mondial de l'assurance se porte à 6 300 milliards de dollars US, ce qui représente 7,2% du PIB mondial. On définit le taux de pénétration comme le ration Primes / PIB qui constitue un indicateur de développement de l'assurance dans un pays. Dans certains pays il dépasse les 15%. Cependant, celui ci n'est que de 0,72% en Algérie, ce qui représente un taux très faible pour ce secteur.

| Pays | Taux de pénétration | Densité de l'assurance |
|---|---|---|
| Taïwan | 20,88% | 5161 $ |
| Royaume-Uni | 10,61% | 4503 $ |
| France | 8,89% | 3667 $ |
| Etats-Unis | 7,14% | 4481 $ |
| Maroc | 3,88% | 127 $ |
| Tunisie | 2,14% | 75 $ |
| Algérie | 0,68% | 28 $ |

TABLE 4.1 – Taux de pénétrations et densités d'assurance par habitant de certains pays en 2008.[1]

### 4.6.1 Contenu de la base de données et ses variables

La base de données contient les informations de toutes les polices d'assurances souscrites auprès de la SAA dans le cadre de la garantie Tous Risques durant les années 2014 à 2018. Avant nettoyage, il y avait au total quelques 850 000 observations. Le fichier qui nous a été remis en comporte 740 000 car toutes les observations où des informations importantes ont été omises ou mal saisies ont dues être retirées. La table contient la liste des polices d'assurances de deux types de catégories de véhicules : les véhicules particuliers ou les véhicules appartenant à des flottes. Il n'est bien entendu pas possible d'étudier ces deux catégories de la même façon car selon une analyse descriptive, ces deux types de catégories présentent une sinistralité très différente, en plus de ne pas contenir les mêmes variables explicatives ; les véhicules des flottes ne sont pas supposés appartenir à des individus à qui l'on pourrait demander un âge, un sexe, un âge de permis, ...

La base de données comporte 40 variables mais ne sont pas toutes significatives. Il sera question de faire une étude sur la fréquence des sinistres puis une étude sur la gravité des sinistres. Ces deux études portent donc sur la variable `sinistres` qui est le nombre de sinistres survenus durant la période d'exposition d'une police d'assurance, et la seconde variable `ChargeSin` qui est le montant que l'assureur a dû payer en totalité à son assuré pour les sinistres survenus durant la période d'exposition.

La liste des variables utiles sont les suivantes :

1. `id` : chaîne de caractères servant de clé d'identification de chaque police d'assurance. Elle est la concaténation du numéro d'agence, du numéro de police, du code d'appartenance à l'une des catégories citées plus haut.
2. `dr` : variable catégorielle. La direction régionale auquel appartient l'agence dans laquelle le contrat d'assurance a été souscrit. Cette variable est très importante car elle nous servira de variable spatiale, permettant une segmentation géographique du portefeuille. Cette variable contient 15 modalités mais on ne retiendra que 14. La modalité `annaba` a été retirée car elle ne contient que 700 observations, ce qui ne représente pas la quantité d'observations réelles qui se situe à 50 000. L'extraction des données de la direction régionale de Annaba est incorrecte. Ceci est problématique car beaucoup des régions du Nord-Est du pays dépendent de cette direction régionale.
3. `agence` : variable catégorielle. Code de l'agence dans laquelle la police d'assurance a été souscrite. Pour l'instant, nous n'avons pas utilisé cette variable car c'est une variable catégorielle et elle possède 592 modalités (592 agences SAA en Algérie en 2018) et cela représenterait un énorme coût en temps de calcul. Il serait judicieux d'utiliser cette variable dans une étude future car on pourrait supposer que la plupart des individus souscrivent une assurance dans une agence près de leur domicile et on pourrait donc peut être associer

la zone de circulation du véhicule à la localisation de l'agence (segmentation par Daïras-Communes). Une autre idée serait d'utiliser les codes d'agences des contrats d'assurances pour déduire les densités d'habitations de certaines régions car cette variable peut intuitivement et conceptuellement aider à expliquer la sinistralité.

4. `categorie` : variable catégorielle. Cette variable catégorielle possède deux modalités déjà citées précédemment. Elle différencie les contrats des flottes de ceux des véhicules particuliers. Cette variable permettra de créer deux bases de données sur lesquelles nous travaillerons : `basePartic` et `baseFlotte`
5. `annee` : valeur entière. Représente l'année de souscription du contrat. Etant donné l'instabilité de l'économie algérienne, il sera important d'utiliser cette variable pour actualiser les valeurs des montants des sinistres et des valeurs assurées des véhicules. Plus de détails sur ce point par la suite.
6. `duree` : variable catégorielle à deux niveaux. Période d'exposition au risque d'un contrat : 6 mois ou une année.
7. `police` : chaîne de caractères. Numéro de police du contrat.
8. `avenant` : valeur entière. Représente le numéro d'avenant du contrat. Par définition, un avenant est un : « acte par lequel on modifie les termes d'un contrat ». En assurance, plus précisément un avenant est une révision du contrat pour corriger, apporter des modifications, ajouter ou retirer des garanties ou bien renouveler celui-ci. La variable suivante a été créée à partir de `avenant`
9. `cutAvenant` : variable catégorielle. Variable déduite de la précédente possédant deux modalités : `new` et `old` permettant de différencier les nouvelles polices d'assurances des anciennes. Elle permettra de séparer les nouveaux assurés de l'agence de ceux qui sont à la SAA depuis au moins 1 an, en supposant que la proportion des avenants de modification liés à des rectifications des nouveaux contrats est faible. L'idée derrière la création de cette variable est de savoir s'il existe bien une tendance des nouveaux contrats à être plus sinistrés que ceux des assurés fidèles ou de savoir s'il y a un phénomène de souscription à la garantie Tous Risques dans un seul but : réparer son véhicule. La création de cette variable est intéressante mais mérite une étude à elle seule car la séparation en 2 modalités exclue le cas des contrats RC renouvelés pendant seulement une année pour une Tous Risques.
10. `code` : valeur entière. Code risque. Un contrat d'assurance peut porter sur plusieurs véhicules et le code risque liste ceux-ci, notamment pour les contrats des flottes.
11. `sex` : variable catégorielle. Sexe du souscripteur de la police d'assurance. Variable pas très intéressante comme on le verra plus tard dans l'étude. L'interprétation sera donnée plus tard. Variable indisponible dans `BaseFlotte`.
12. `age` : variable continue. Age du souscripteur du contrat. Variable indisponible dans `BaseFlotte`.
13. `permis` : variable continue. Ancienneté du permis en années. Variable indisponible dans `BaseFlotte`.
14. `nvPermis` : variable catégorielle. Variable permettant de séparer les assurés avec un nouveau permis ou non. Variable indisponible dans `BaseFlotte`. Cetta variable peut être déduite de la précédente en séparant les anciennetés de permis `0` des autres.
15. `marque` : variable catégorielle avec 113 modalités. Elle représente la marque du véhicule.
16. `brand` : variable catégorielle à 56 modalités représentant les marques mais regroupées selon une certaine logique. Par exemple les marques `RENAULT` et `DACIA` sont regroupées en un groupe `RENAULT-DACIA`.
17. `ageVeh` : variable continue. Représente l'âge du véhicule en années.

18. `genre` : variable catégorielle avec 12 modalités. Elle représente le genre de véhicule ; exemples : Transport Publique de Marchandises, Véhicules de plus de 3, 5 tonnes, Tracteurs Routiers, ...
19. `usage` : variable catégorielle avec 10 niveaux. L'usage auquel le véhicule est supposé être assuré ; exemples : affaire, fonctionnaire, location, taxi, ...
20. `Fiscal` : variable continue. Puissance fiscale du véhicule.
21. `zone` : variable catégorielle à deux modalités `nord` et `sud`. Variable permettant de séparer les véhicules sensés être utilisés dans le nord ou dans le sud du pays.
22. `inflam` : variable catégorielle à deux modalités permettant de séparer les véhicules transportant des marchandises inflammables ou non.
23. `tauxRed` : variable numérique. Représente un taux de réduction appliqué sur la prime demandée pendant la souscription.
24. `PleinTarifAct` : variable continue. Représente la prime Tous Risques avant application du taux de réduction. Ce montant est actualisé selon l'année de souscription du contrat.
25. `prime_trAct` : variable continue représentant le montant de la prime exact payé par l'assuré pour la garantie Tous Risques. Réduction appliquée et montant actualisé.
26. `sinistres` : valeur entière. Variable sur laquelle portera une des études dans ce mémoire. Représente le nombre de sinistres reportés durant la période d'exposition d'un contrat.
27. `VA` : variable continue. Représente la valeur assurée du véhicule. Ce montant est demandé par l'assuré pendant la souscription du contrat. Ce montant a été actualisé.
28. `ChargeSin` : variable continue. Deuxième variable sur laquelle portera l'étude. Représente le montant total des déboursements qu'a dû faire l'assureur sur un contrat concernant uniquement la garantie Tous Risques. Valeur actualisée.

En algérie, les nouvelles garanties ne sont que très récentes et la majorité des chiffres d'affaires des assureurs étant toujours en majorité réalisés par les branches automobiles et IARD (Incendie, Accidents et Risques Divers).

De plus, en conséquence des évènements ayant conduit à la nationalisation de l'activité assurance en Algérie après son indépendance, les compagnies nationales ont des portefeuilles spécifiques en lien avec la raison de leur création originale. En effet, le portefeuille de la Société Nationale d'Assurance SAA par exemple est en grande partie composé de l'assurance automobile, tandis que celui de la CAAR est majoritairement composé des assurances gros risques et de transport. Les risques agricoles sont en grande partie assurés par la CNMA (Caisse Nationale de Mutualité Agricole). Cette tendance s'est maintenue depuis la création de ces compagnies d'assurance mais du fait de la nécessité de diversification des garanties, cette prédisposition des compagnies à des marchés spécifiques tend à disparaître.

## 4.7 Organisme d'accueil : la SAA et sa place dans le marché

La rédaction de ce mémoire a été faite durant un passage à la Direction Générale de la SAA dans le contexte d'un stage pratique. Nous avons pu découvrir comment fonctionne les assurances algériennes et également beaucoup appris sur la mise en place et le suivi des garanties, spécialement dans la branche automobile. Nous tenons à remercier infiniment les membres de la division automobile qui nous ont donné l'accès à leur base de données Tous Risques pour nous permettre de mener à bien l'étude sur la tarification des assureurs algériens.

| Branche | SAA | Croissance | Secteur | Part de la SAA |
|---|---|---|---|---|
| Automobile | 20 038 | +2,07% | 69 021 | 29,03% |
| IRD | 6 453 | +7,67% | 45 867 | 13,07% |
| Assurance Agricole | 614 | +24,78% | 2 624 | 24,82% |
| Assurance Transport | 489 | +20,53% | 5 828 | 8,38% |
| Assurance Crédit | 85 | +1 597% | 2 144 | 3,96% |
| Total Ass. Dommages | 27 679 | +4,34% | 126 095 | 21,95% |

TABLE 4.2 – Parts de marché de la SAA en 2018 par branche en millions de DA et croissance par rapport à 2017.[33]

### 4.7.1 Histoire de la SAA

En 1963, la Société Nationale d'Assurance [32] voit le jour en tant que compagnie générale d'assurance sous la marque SAA et le premier point de vente ouvre ses portes à Alger-Centre. En Mai 1966, le monopole de l'état algérien sur les opérations d'assurance conduit à la nationalisation de la SAA. En janvier 1976, la SAA se spécialise dans la branche des risques simples en développant des offres adaptées aux particuliers, aux professionnels, aux collectivités locales et institutions relevant du secteur de la santé. En 1989, la SAA transforme son mode de gouvernance et devient une entreprise publique économique (EPE) avec un capital de 80 millions de DA. La SAA élargit son champ d'activités en 1990 aux risques industriels, du transport, risques agricoles et assurances de personnes. Le marché de l'assurance algérien en pleine expansion a nécessité une certaine modernisation. Ainsi, la SAA a revu l'organisation de son réseau d'agences en partant du principe performance/meilleure rémunération. En 2004, une réorganisation structurelle est réalisée au sein de la compagnie en créant une division par segment de marché afin de booster la productivité. En 2011, son capital social atteint 20 milliards de DA puis à 30 milliards en 2017.

### 4.7.2 Parts de marché de la SAA

Comme on peut le voir sur la figure 4.5, la SAA est le premier assureur automobile en Algérie avec une assez grande marge d'avance sur le second. De plus, grâce à ses stratégies de diversification, la SAA a réussi à s'imposer également sur les autres secteurs.

### 4.7.3 Organigramme de la SAA

J'ai eu le plaisir et l'honneur de faire mon stage dans la branche production de la division automobile au sein de la SAA. Le direction de la production automobile se charge de mettre en place de nouveaux produits d'assurance auto et de porter un suivi sur ceux-ci pour évaluer leur rentabilité. De plus, la tarification et les conventions sont élaborées dans ce service.

**Introduction** – Dans ce chapitre, nous allons présenter le secteur des assurances en Algérie en donnant d'abord un bref historique de l'activité d'assurance, puis les chiffres de l'assurance dans sa globalité et ensuite les chiffres qui nous intéressent, c'est-à-dire ceux de l'assurance automobile. Enfin, une présentation de l'organisme d'accueil, la société nationale d'assurance SAA, sera faite. Ses chiffres et son positionnement dans le marché des assurances seront également présentés.

## 4.8 Histoire de l'assurance en Algérie

Les besoins en assurance des Algériens durant la période coloniale sont considérés comme insignifiants compte tenu de leurs revenus et de leur situation socioculturelle. L'assurance en Algérie a commencé à se développer au lendemain de l'indépendance mais la majorité des opérations des assurances restaient aux mains des entreprises françaises. Pendant cette période, une grande partie des primes collectées par les compagnies d'assurance ont été transférées à l'étranger. Il aura fallu attendre 1966 pour que l'état reprenne le monopole sur l'assurance en Algérie. L'exploitation de cette activité était désormais réservée à l'état via les entreprises nationales qui avaient vues le jour durant les années 1963-1964.[6] En 1976, les quatre compagnies d'assurance nationales – CAAR, SAA, MAATEC et la CCRMA – se spécialisent chacune dans la couverture d'un certain type de risque. En 1995, une ordonnance supprime le monopole de l'état sur le marché d'assurance, autorisant ainsi la création de compagnies d'assurance privées. Grâce à ce changement, la responsabilité civile n'est plus la seule garantie proposée par les assureurs et permet ainsi une certaine rentabilité de ce secteur.

## 4.9 Les chiffres de l'assurance en Algérie

En 2009, le marché algérien de l'assurance était à 65% détenu par l'état. Cependant, malgré l'ouverture au marché de ce secteur d'activité aux autres compagnies privées, ce chiffre semble stagner. En 2020, la proportion est de 72% et représente un montant conséquent de 95,7 milliards de dinars sur un total de 132 milliards de dinars.

En 2019, Swiss RE, une compagnie de réassurance suisse a présenté son rapport SIGMA 4/2019 [30] sur les chiffres de l'assurance à l'échelle mondiale. Ainsi, le chiffre d'affaire du marché mondial de l'assurance se porte à 6 300 milliards de dollars US, ce qui représente 7,2% du PIB mondial. On définit le taux de pénétration comme le ration Primes / PIB qui constitue un indicateur de développement de l'assurance dans un pays. Dans certains pays il dépasse les 15%. Cependant, celui ci n'est que de 0,72% en Algérie, ce qui représente un taux très faible pour ce secteur.

| Pays | Taux de pénétration | Densité de l'assurance |
|---|---|---|
| Taïwan | 20,88% | 5161 $ |
| Royaume-Uni | 10,61% | 4503 $ |
| France | 8,89% | 3667 $ |
| Etats-Unis | 7,14% | 4481 $ |
| Maroc | 3,88% | 127 $ |
| Tunisie | 2,14% | 75 $ |
| Algérie | 0,68% | 28 $ |

TABLE 4.3 – Taux de pénétrations et densités d'assurance par habitant de certains pays en 2008.[1]

En algérie, les nouvelles garanties ne sont que très récentes et la majorité des chiffres d'affaires des assureurs étant toujours en majorité réalisés par les branches automobiles et IARD (Incendie, Accidents et Risques Divers).

De plus, en conséquence des évènements ayant conduit à la nationalisation de l'activité assurance en Algérie après son indépendance, les compagnies nationales ont des portefeuilles spécifiques en lien avec la raison de leur création originale. En effet, le portefeuille de la Société Nationale d'Assurance SAA par exemple est en grande partie composé de l'assurance automobile, tandis que celui de la CAAR est majoritairement composé des assurances gros risques et de transport. Les risques agricoles sont en grande partie assurés par la CNMA (Caisse Nationale de Mutualité Agricole). Cette tendance s'est maintenue depuis la création de ces compagnies d'assurance mais du fait de la nécessité de diversification des garanties, cette prédisposition des compagnies à des marchés spécifiques tend à disparaître.

## 4.10 Organisme d'accueil : la SAA et sa place dans le marché

La rédaction de ce mémoire a été faite durant un passage à la Direction Générale de la SAA dans le contexte d'un stage pratique. Nous avons pu découvrir comment fonctionne les assurances algériennes et également beaucoup appris sur la mise en place et le suivi des garanties, spécialement dans la branche automobile. Nous tenons à remercier infiniment les membres de la division automobile qui nous ont donné l'accès à leur base de données Tous Risques pour nous permettre de mener à bien l'étude sur la tarification des assureurs algériens.

### 4.10.1 Histoire de la SAA

En 1963, la Société Nationale d'Assurance [32] voit le jour en tant que compagnie générale d'assurance sous la marque SAA et le premier point de vente ouvre ses portes à Alger-Centre. En Mai 1966, le monopole de l'état algérien sur les opérations d'assurance conduit à la nationalisation de la SAA. En janvier 1976, la SAA se spécialise dans la branche des risques simples en développant des offres adaptées aux particuliers, aux professionnels, aux collectivités locales et institutions relevant du secteur de la santé. En 1989, la SAA transforme son mode de gouvernance et devient une entreprise publique économique (EPE) avec un capital de 80 millions de DA. La SAA élargit son champ d'activités en 1990 aux risques industriels, du transport, risques agricoles et assurances de personnes. Le marché de l'assurance algérien en pleine expansion a nécessité une certaine modernisation. Ainsi, la SAA a revu l'organisation de son réseau d'agences en partant du principe performance/meilleure rémunération. En 2004, une réorganisation structurelle est réalisée au sein de la compagnie en créant une division par segment de marché afin de booster la productivité. En 2011, son capital social atteint 20 milliards de DA puis à 30 milliards en 2017.

| Branche | SAA | Croissance | Secteur | Part de la SAA |
|---|---|---|---|---|
| Automobile | 20 038 | +2,07% | 69 021 | 29,03% |
| IRD | 6 453 | +7,67% | 45 867 | 13,07% |
| Assurance Agricole | 614 | +24,78% | 2 624 | 24,82% |
| Assurance Transport | 489 | +20,53% | 5 828 | 8,38% |
| Assurance Crédit | 85 | +1 597% | 2 144 | 3,96% |
| Total Ass. Dommages | 27 679 | +4,34% | 126 095 | 21,95% |

TABLE 4.4 – Parts de marché de la SAA en 2018 par branche en millions de DA et croissance par rapport à 2017.[33]

### 4.10.2 Parts de marché de la SAA

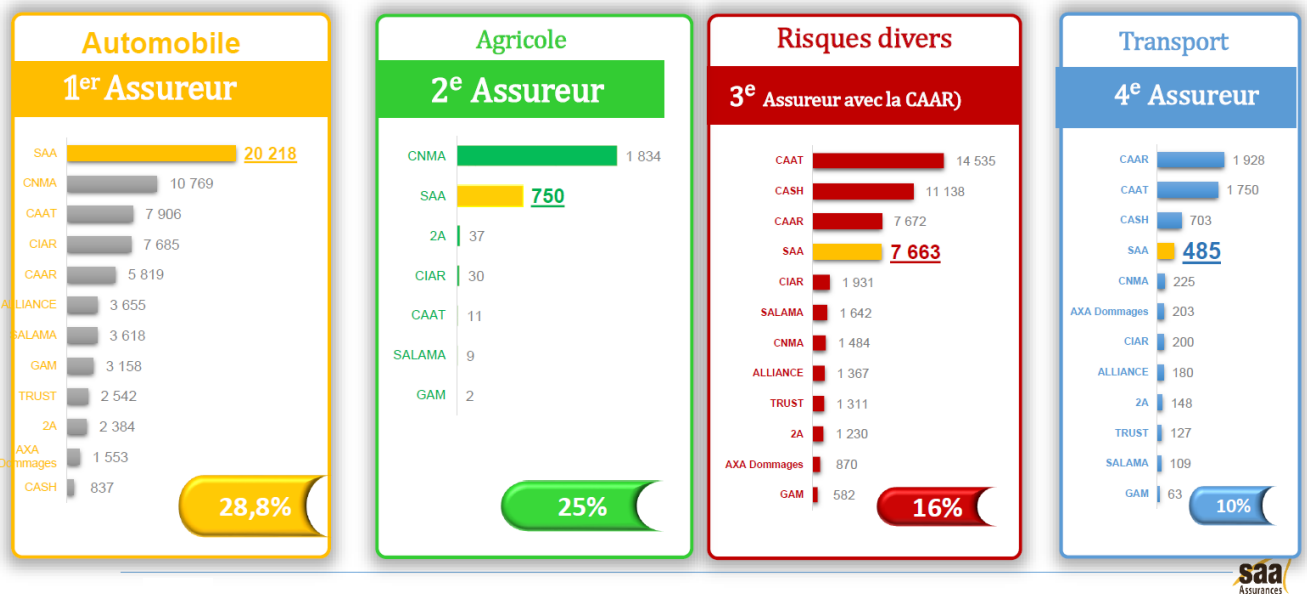



FIGURE 4.5 – Positionnement de la SAA sur le marché d'assurance en 2018. Montants en millions de DA.[34]

Comme on peut le voir sur la figure 4.5, la SAA est le premier assureur automobile en Algérie avec une assez grande marge d'avance sur le second. De plus, grâce à ses stratégies de diversification, la SAA a réussi à s'imposer également sur les autres secteurs.

### 4.10.3 Organigramme de la SAA

J'ai eu le plaisir et l'honneur de faire mon stage dans la branche production de la division automobile au sein de la SAA. Le direction de la production automobile se charge de mettre en place de nouveaux produits d'assurance auto et de porter un suivi sur ceux-ci pour évaluer leur rentabilité. De plus, la tarification et les conventions sont élaborées dans ce service.

## 4.11 Contexte actuel économique de l'assurance automobile en Algérie

### 4.11.1 Situation des assurances pour l'année 2020

Si un secteur d'activité n'est pas aussi rentable dans un pays comparé au reste du monde, il ne relève pas toujours entièrement de la responsabilité des acteurs de ce secteur d'activité. A l'instar des chiffres précédemment cités, l'assurance en Algérie ne représente qu'une faible proportion du PIB algérien mais cela est d'abord dû à la culture du pays, à la façon dont est menée cette activité mais elle est également étroitement liée à l'économie du pays. En effet, les hydrocarbures représentent la principale rentrée en devises du pays et il s'en suit que l'économie algérienne dépend entièrement de ses ventes en ressources fossiles. Ainsi, lorsque les cours de ces dites ressources subissent d'importantes baisses, tous les secteurs d'activités en Algérie en ressentent les effets sur leurs finances. De la macroéconomie à la microéconomie, tout est imbriqué.[13]
Le marché des assurances ne déroge pas à cette règle et aurait pu lui aussi subir une forte régression. Fort heureusement, les assureurs ont eu l'initiative de diversifier leurs produits pour maintenir une certaine rentabilité de l'activité. Cependant, la situation exceptionnelle née du mouvement social et politique de février 2019 et de la crise sanitaire en 2020 dont les retombées n'ont pas fini de faire des dégâts sur l'économie algérienne ne fera qu'empirer l'état fragile de l'assurance en Algérie.
Les statistiques des notes de conjoncture de la CNA [9] sont là pour confirmer ces faits.

On peut clairement voir que la majeur partie des rentrées de liquidités en automobile proviennent des différentes garanties facultatives (hors RC) proposées par les assureurs – 76,1 % d'entre elles. Et plus particulièrement, 51,7 % proviennent des garanties Tous Risques ou Dommage Collision.
En Algérie, la tarification de la RC ne dépend pas des assureurs. La réglementation oblige les assureurs à tous pratiquer le même système tarifaire pour la responsabilité civile. Il serait intéressant d'étudier les chiffres de la RC mais pas dans le but de rentabiliser cette garantie.
Cependant, les assureurs peuvent pratiquer des prix différents pour les garanties alternatives et c'est sur cela qu'il est possible d'améliorer la situation économique d'un assureur et surtout rester compétitif dans un marché clairement déloyal.

**Remarque** – Il est à noter que depuis la nationalisation de l'activité d'assurance, le gouvernement a tenté de garder l'esprit social en tarifant la garantie obligatoire RC à partir du salaire minimum algérien. Ce tarif a été instauré il y a plus d'une quarantaine d'années et aujourd'hui, la prime moyenne est de 1500 da et sachant que l'Algérie enregistre un très fort taux de sinistralité, les charges sinistres en RC à payer dépassent de loin les primes récoltées : **23,9 % de production contre 56,1 % des sinistres à payer en 2019 [10]**

Pour compenser la garantie RC, les assureurs utilisent donc les garanties plus rentables et c'est là où la diversification des garanties instaurées par les compagnies d'assurance algériennes a induit des profits et résultats intéressants.

## 4.12 Problématique et motivations de ce mémoire

**Point Important 1** – L'étude menée dans ce mémoire concerne la garantie Tous Risques mais il est clairement possible d'appliquer la même méthodologie pour estimer les sinistres de chaque garantie et déterminer un tarif adéquat pour chacune d'elle. Pour la majorité des individus voulant souscrire un contrat d'assurance, un des facteurs les plus importants de décision pour le choix de contracter ou non une garantie dommage est le tarif de celle-ci. Le système actuel ne

tient compte que de la valeur assurée du véhicule pour déterminer la prime d'un contrat mais il serait logique de penser que cela pourrait énormément affecter l'intention d'un assuré à prendre une garantie Tous Risques. Le tableau 4.6 simule et compare quelques primes de garanties Tous Risques en fonction du véhicule et de sa valeur assurée.

| Véhicule | Valeur du véhicule | Prime Tous Risques | **Prime totale demandée** |
|---|---|---|---|
| Dacia Sandero | 2 950 000 DA | 147 500 DA | **221 965 DA** |
| Seat Leon | 7 000 000 DA | 350 000 DA | **579 931 DA** |
| Renault Clio | 4 200 000 DA | 210 000 DA | **313 046 DA** |
| Peugeot Expert | 4 800 000 DA | 240 000 DA | **356 606 DA** |
| Kia Rio | 3 500 000 DA | 175 000 DA | **261 895 DA** |

TABLE 4.5 – Tableau des primes qu'auraient payé des assurés particuliers. Les véhicules choisis sont des modèles fréquents en Algérie. [2]

En 2021, la valeur d'un véhicule neuf est d'au moins 2,5 millions de dinars, et la prime Tous Risques correspondante à cette VA est de 125 000, sans compter les différentes taxes dont la TVA. Avec un salaire moyen en Algérie de 40 000 da en 2020 [28], et un niveau de vie qui ne cesse de diminuer à cause de la situation économique algérienne, ce genre de tarification même pour un véhicule de moyenne gamme pourrait en repousser plus d'un, en plus de devoir être payé en une seule fois lors de la souscription d'un contrat et en liquide (facteur ayant tendance à repousser lorsqu'il s'agit de payer un service non matériel – on s'imagine mal se rendre dans une agence d'assurance avec une importante somme pour payer une assurance – Donc solutions : diminuer **stratégiquement** les primes aux individus à bas risques **et** proposer le paiement en ligne ou en plusieurs fois.).
Certes, pour remédier à la cherté de cette garantie, les assureurs ont eu recours à l'attribution de réductions mais cette stratégie ne se base sur aucune étude statistique pour l'affectation des différents taux de réduction et il est fort probable que beaucoup d'assurés qui ne bénéficient d'aucune réduction paient une prime complète alors qu'elle n'est pas proportionnelle au risque auquel ces assurés sont exposés. Inversement, il y a une forte probabilité qu'il y ait attribution de réductions à des groupes d'assurés qui sont très sinistrés et donc ne devraient pas bénéficier de forts taux de réduction. La concurrence déloyal dans le marché algérien joue sur ce taux de réduction pour attirer de nouveaux assurés, même si aucune étude n'est menée sur les risques encourus et les pertes potentielles dues à l'attribution de fortes réductions.

**Point Important 2** – En plus du caractère couteux des garanties dommages, il est important de considérer l'évolution du marché automobile algérien car les garanties facultatives en dépendent fortement.
En effet, d'après les données mises à ma disposition, 87 % des véhicules assurés en Tous Risques sont des véhicules de 5 ans ou moins [3].

On en déduit que la rentabilité de cette garantie dépend fortement du marché du véhicule neuf en Algérie. Et comme soulevé précédemment à maintes reprises, la situation économique en Algérie se détériore et le marché des véhicules neufs algériens a subit un énorme déclin. L'importation des véhicules neufs est à l'arrêt depuis 2015. Le soit disant montage des véhicules a débuté peu après et le marché a gardé sa forme pendant quelques années. Depuis février 2019, toutes les usines de montages sont à l'arrêt et il devient très difficile pour les algériens de se procurer un véhicule neuf et cela a eu une répercussion direct sur l'économie des assurances.

2. Les données des valeurs du véhicules ont été tirées de Ouedkniss et il semble légitime de croire que c'est une référence convaincante concernant la valeur des véhicules.
3. Dans la catégorie des véhicules particuliers

Seuls les véhicules importés par les licences Moudjahidines sont aujourd'hui en vente mais ceux-ci sont rares et très couteux pour un algérien moyen (plus de 5 millions de DA en général). De plus, le portefeuille des assureurs auto et surtout la rentabilité de ce portefeuille dépend fortement des véhicules de petites et moyennes gammes (80 % des véhicules assurés entre 2014 et 2018 ont des valeurs assurées de moins de 2 millions de DA [4]).

Les données disponibles dans la Base de Données Centralisée des Statistiques (BDCS) s'arrêtent à l'année 2017 pour les informations concernant l'importation des véhicules, mais selon un article de Algérie-eco, environ 8000 véhicules ont été importés par les particuliers durant le premier trimestre 2021, et si on ramène ce chiffre à l'année, cela donne environ 32 000 véhicules par an. Si on utopise le fait que tous ces véhicules souscrivent une assurance Tous Risques, cela reste difficile de concevoir que cela suffise à garder la rentabilité des portefeuilles automobiles des assureurs. Pour aller plus loin, il serait intéressant de recueillir les données sur les nombres de sinistres (qui sont très élevés en Algérie) et les recouper avec le faible nombre de véhicules importés afin de suivre l'évolution du patrimoine automobile assurable.

**Conclusion** – Si la situation ne s'améliore pas et que le marché des véhicules neufs ne redécolle pas, les assureurs vont voir les rentabilités de leurs garanties dommages fortement baisser à l'horizon 2023. Etant donné que les actions que peuvent mener les assureurs sont limitées pour remédier à la crise du marché automobile, il ne leur reste qu'à trouver des stratégies afin de fidéliser leurs assurés et essayer de les garder dans leur portefeuille Tous Risques ou Dommages Collision plus longtemps. A notre sens, la révision de la tarification aura un impact majeur sur la stratégie de fidélisation de l'assuré à la garantie facultative. D'où l'objet de ce mémoire.

## 4.13 Contexte actuel économique de l'assurance automobile en Algérie

### 4.13.1 Situation des assurances pour l'année 2020

Si un secteur d'activité n'est pas aussi rentable dans un pays comparé au reste du monde, il ne relève pas toujours entièrement de la responsabilité des acteurs de ce secteur d'activité. A l'instar des chiffres précédemment cités, l'assurance en Algérie ne représente qu'une faible proportion du PIB algérien mais cela est d'abord dû à la culture du pays, à la façon dont est menée cette activité mais elle est également étroitement liée à l'économie du pays. En effet, les hydrocarbures représentent la principale rentrée en devises du pays et il s'en suit que l'économie algérienne dépend entièrement de ses ventes en ressources fossiles. Ainsi, lorsque les cours de ces dites ressources subissent d'importantes baisses, tous les secteurs d'activités en Algérie en ressentent les effets sur leurs finances. De la macroéconomie à la microéconomie, tout est imbriqué.[13]

Le marché des assurances ne déroge pas à cette règle et aurait pu lui aussi subir une forte régression. Fort heureusement, les assureurs ont eu l'initiative de diversifier leurs produits pour maintenir une certaine rentabilité de l'activité. Cependant, la situation exceptionnelle née du mouvement social et politique de février 2019 et de la crise sanitaire en 2020 dont les retombées n'ont pas fini de faire des dégâts sur l'économie algérienne ne fera qu'empirer l'état fragile de l'assurance en Algérie.

Les statistiques des notes de conjoncture de la CNA [9] sont là pour confirmer ces faits.

La branche «automobile» enregistre, à fin 2020, un chiffre d'affaires de 62,8 milliards de DA. Comparativement aux 69,2 milliards de DA de 2019, il y a eu donc une régression de 9,2 %.

4. Selon les données mises à ma disposition

### 4.13.2 Rentabilité de l'assurance automobile

Le diagramme suivant [10] montre la répartition des revenues de la branche automobile par garantie.

On peut clairement voir que la majeur partie des rentrées de liquidités en automobile proviennent des différentes garanties facultatives (hors RC) proposées par les assureurs – 76,1 % d'entre elles. Et plus particulièrement, 51,7 % proviennent des garanties Tous Risques ou Dommage Collision.
En Algérie, la tarification de la RC ne dépend pas des assureurs. La réglementation oblige les assureurs à tous pratiquer le même système tarifaire pour la responsabilité civile. Il serait intéressant d'étudier les chiffres de la RC mais pas dans le but de rentabiliser cette garantie.
Cependant, les assureurs peuvent pratiquer des prix différents pour les garanties alternatives et c'est sur cela qu'il est possible d'améliorer la situation économique d'un assureur et surtout rester compétitif dans un marché clairement déloyal.

**Remarque** – Il est à noter que depuis la nationalisation de l'activité d'assurance, le gouvernement a tenté de garder l'esprit social en tarifant la garantie obligatoire RC à partir du salaire minimum algérien. Ce tarif a été instauré il y a plus d'une quarantaine d'années et aujourd'hui, la prime moyenne est de 1500 da et sachant que l'Algérie enregistre un très fort taux de sinistralité, les charges sinistres en RC à payer dépassent de loin les primes récoltées : **23,9 % de production contre 56,1 % des sinistres à payer en 2019 [10]**

Pour compenser la garantie RC, les assureurs utilisent donc les garanties plus rentables et c'est là où la diversification des garanties instaurées par les compagnies d'assurance algériennes a induit des profits et résultats intéressants.

## 4.14 Problématique et motivations de ce mémoire

**Point Important 1** – L'étude menée dans ce mémoire concerne la garantie Tous Risques mais il est clairement possible d'appliquer la même méthodologie pour estimer les sinistres de chaque garantie et déterminer un tarif adéquat pour chacune d'elle. Pour la majorité des individus voulant souscrire un contrat d'assurance, un des facteurs les plus importants de décision pour le choix de contracter ou non une garantie dommage est le tarif de celle-ci. Le système actuel ne tient compte que de la valeur assurée du véhicule pour déterminer la prime d'un contrat mais il serait logique de penser que cela pourrait énormément affecter l'intention d'un assuré à prendre une garantie Tous Risques. Le tableau 4.6 simule et compare quelques primes de garanties Tous Risques en fonction du véhicule et de sa valeur assurée.

| Véhicule | Valeur du véhicule | Prime Tous Risques | **Prime totale demandée** |
|---|---|---|---|
| Dacia Sandero | 2 950 000 DA | 147 500 DA | **221 965 DA** |
| Seat Leon | 7 000 000 DA | 350 000 DA | **579 931 DA** |
| Renault Clio | 4 200 000 DA | 210 000 DA | **313 046 DA** |
| Peugeot Expert | 4 800 000 DA | 240 000 DA | **356 606 DA** |
| Kia Rio | 3 500 000 DA | 175 000 DA | **261 895 DA** |

TABLE 4.6 – Tableau des primes qu'auraient payé des assurés particuliers. Les véhicules choisis sont des modèles fréquents en Algérie. [5]

5. Les données des valeurs du véhicules ont été tirées de Ouedkniss et il semble légitime de croire que c'est une référence convaincante concernant la valeur des véhicules.

En 2021, la valeur d'un véhicule neuf est d'au moins 2,5 millions de dinars, et la prime Tous Risques correspondante à cette VA est de 125 000, sans compter les différentes taxes dont la TVA. Avec un salaire moyen en Algérie de 40 000 da en 2020 [28], et un niveau de vie qui ne cesse de diminuer à cause de la situation économique algérienne, ce genre de tarification même pour un véhicule de moyenne gamme pourrait en repousser plus d'un, en plus de devoir être payé en une seule fois lors de la souscription d'un contrat et en liquide (facteur ayant tendance à repousser lorsqu'il s'agit de payer un service non matériel – on s'imagine mal se rendre dans une agence d'assurance avec une importante somme pour payer une assurance – Donc solutions : diminuer **stratégiquement** les primes aux individus à bas risques **et** proposer le paiement en ligne ou en plusieurs fois.).
Certes, pour remédier à la cherté de cette garantie, les assureurs ont eu recours à l'attribution de réductions mais cette stratégie ne se base sur aucune étude statistique pour l'affectation des différents taux de réduction et il est fort probable que beaucoup d'assurés qui ne bénéficient d'aucune réduction paient une prime complète alors qu'elle n'est pas proportionnelle au risque auquel ces assurés sont exposés. Inversement, il y a une forte probabilité qu'il y ait attribution de réductions à des groupes d'assurés qui sont très sinistrés et donc ne devraient pas bénéficier de forts taux de réduction. La concurrence déloyal dans le marché algérien joue sur ce taux de réduction pour attirer de nouveaux assurés, même si aucune étude n'est menée sur les risques encourus et les pertes potentielles dues à l'attribution de fortes réductions.

**Point Important 2** – En plus du caractère couteux des garanties dommages, il est important de considérer l'évolution du marché automobile algérien car les garanties facultatives en dépendent fortement.
En effet, d'après les données mises à ma disposition, 87 % des véhicules assurés en Tous Risques sont des véhicules de 5 ans ou moins [6].

On en déduit que la rentabilité de cette garantie dépend fortement du marché du véhicule neuf en Algérie. Et comme soulevé précédemment à maintes reprises, la situation économique en Algérie se détériore et le marché des véhicules neufs algériens a subit un énorme déclin. L'importation des véhicules neufs est à l'arrêt depuis 2015. Le soit disant montage des véhicules a débuté peu après et le marché a gardé sa forme pendant quelques années. Depuis février 2019, toutes les usines de montages sont à l'arrêt et il devient très difficile pour les algériens de se procurer un véhicule neuf et cela a eu une répercussion direct sur l'économie des assurances. Seuls les véhicules importés par les licences Moudjahidines sont aujourd'hui en vente mais ceux-ci sont rares et très couteux pour un algérien moyen (plus de 5 millions de DA en général). De plus, le portefeuille des assureurs auto et surtout la rentabilité de ce portefeuille dépend fortement des véhicules de petites et moyennes gammes (80 % des véhicules assurés entre 2014 et 2018 ont des valeurs assurées de moins de 2 millions de DA [7]).

### 4.14.1 Modèle individuel

Le modèle individuel se place au niveau de chaque police [5]. Les coûts totaux des sinistres causés par les $n$ polices du portefeuille sont notés $S_1, S_2, \ldots, S_n$. Ces variables sont supposées indépendantes, mais pas identiquement distribuées. Ceci permet de tenir compte de l'hétérogénéité du portefeuille et d'une éventuelle segmentation a priori effectuée par l'assureur. Si on note $F_i$ la fonction de répartition de $S_i$, c'est-à-dire :

$$F_i(x) = \mathbb{P}(S_i \leq x), \quad x \in \mathbb{R},$$

6. Dans la catégorie des véhicules particuliers
7. Selon les données mises à ma disposition

et que la charge de sinistre $S_i$ causée par la police numéro $i$ est représentée par

$$S_i = \begin{cases} 0, & \text{si la police } i \text{ ne cause aucun sinistre,} \\ Y_i & \text{sinon} \end{cases},$$

avec $Y_i$ représentant le coût total des sinistres relatifs à la police $i$ lorsque cette police a donné lieu à au moins un sinistre. La charge totale de sinistre du portefeuille, notée $S^{ind}$ vaut alors

$$S^{ind} = \sum_{i=1}^{n} S_i.$$

Considérons à présent $q_i$ la probabilité que la police $i$ produise au moins un sinistre sur la période d'exposition et $p_i$ celle qu'elle n'en produise aucun. On a alors dans la réalité $F_i(0) = p_i < 1$.
Si on note $G_i(x) = \mathbb{P}(S_i \leq x | S_i > 0)$ la fonction de répartition de la charge de sinistre de la police $i$ sachant que celle ci a produit au moins un sinistre, on a alors :

$$F_i(x) = p_i \mathbb{1}_{x \geq 0} + q_i G_i(x), \quad x > 0,$$

ce qui donne

$$G_i(x) = \frac{F_i(x) - F_i(0)}{1 - F_i(0)}.$$

L'espérance et la variance de la charge totale du portefeuille dans le modèle individuel s'expriment aisément en fonction des deux premiers moments des montants de sinistres par police :

$$\begin{cases} \mathbb{E}(S^{ind}) = \mathbb{E}(\sum_{i=1}^{n} S_i) = \sum_{i=1}^{n} \mathbb{E}(S_i) \\ Var(S^{ind}) = \sum_{i=1}^{n} Var(S_i) \end{cases}$$

Le montant de sinistre $S_i$ engendré par la police $i$ peut s'exprimer sous la forme

$$S_i = \mathbb{1}_i Y_i$$

avec $\mathbb{1}$, vaut 1 si le contrat $i$ a été touché par au moins un sinistre et 0 sinon.
Les variables aléatoires $\mathbb{1}, , \ldots, Y_1, Y_2, \ldots, Y_n$ sont supposées mutuellement indépendantes. Alors sous cette hypothèse, le montant de la prime pure dans le modèle individuel est donné par :

$$\mathbb{E}(S_i) = \mathbb{E}(\mathbb{1})\mathbb{E}(Y_i).$$

Sous cette même hypothèse et en notant $H^{ind}$ la fonction de répartition de $S^{ind}$, on a alors :

$$H^{ind}(x) = F_1 \star F_2 \star \cdots \star F_n.$$

Malheureusement, le nombre $n$ de polices est en général très grand, ce qui rend impossible le calcul direct de $H^{ind}$ (chacun des produits de convolution nécessitant une intégration numérique). Afin de contourner le problème lié au calcul de $H^{ind}$,les actuaires ont suggéré d'approximer le modèle individuel par un équivalent collectif, dans lequel les calculs sont (sous certaines conditions) plus aisés à effectuer.

### 4.14.2 Modèle collectif

Le modèle collectif de théorie du risque ne distingue plus les polices composant le portefeuille mais voit ce dernier comme un ensemble soumis à une série de chocs causées par l'occurrence des sinistres. Le modèle individuel se place au niveau de la police et distingue la charge

des sinistres $S_i$ générée par la police $i$ dans la charge totale $S^{ind} = \sum_{i=1}^{n} S_i$ relative au portefeuille. Contrairement au modèle individuel, le modèle collectif ne distingue plus les polices composant le portefeuille mais voit ce dernier comme un tout, comme un collectif de risques. Les coûts des sinistres touchant le collectif de risques sont modélisés par des variables positives, indépendantes et de même loi. L'identique distribution des coûts dans le modèle collectif s'explique par le fait que l'actuaire renonce à savoir quelle police a causé le sinistre, et gomme donc les différences de sinistralité existant entre les assurés du portefeuille.
Dans la vision collective, $N$ désigne le nombre des sinistres survenus durant une certaine période et $X_i, \quad i = 1, 2, \ldots,$ les montants de ceux-ci. La charge totale des sinistres $S^{coll}$ pour la compagnie s'écrit alors

$$S^{coll} = \sum_{i=1}^{N} X_i,$$

avec la convention que $S^{coll} = 0$ lorsque $N = 0$. Les variables aléatoires $X_i, \quad i = 1, 2, \ldots$ sont supposées indépendantes et identiquement distribuées, et $N$ est supposée indépendante des $X_i$. La loi de $S^{coll}$ est donc composée. Très souvent en assurance, $N$ sera supposé de loi de Poisson, de sorte que $S^{coll} \sim CPoi(\lambda, F)$ avec $F$ représentant la fonction de répartition commune des $X_i$. L'espérance et la variance de la charge totale du portefeuille dans le modèle collectif s'expriment comme ceci :

$$\begin{cases} \mathbb{E}(S^{coll}) = \mathbb{E}(\sum_{i=1}^{N} X_i) = \mathbb{E}(N)\mathbb{E}(X_i) = \mathbb{E}(N)\mathbb{E}(X) \\ Var(S^{coll}) = \mathbb{E}(N)Var(X) + \mathbb{E}^2(X)Var(N) \end{cases}$$

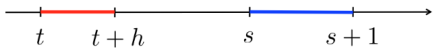
## 4.15 Régression sur variable de comptage

### 4.15.1 Régression de Poisson

On dit qu'une variable aléatoire $Y$ suit une loi de Poisson de paramètre $\lambda$ et on écrit $Y \sim \mathcal{P}(\lambda)$ si sa densité s'écrit :

$$P(Y = y) = e^{-\lambda} \frac{\lambda^y}{y!}.$$

L'espérance et la variance de $Y$ sont alors $\mathbb{E}(N) = Var(N) = \lambda \in \mathbb{R}^+$.
On dit que $(N_t)_{t\geq 0}$ est un processus de Poisson homogène [5, 24, 2] (d'intensité $\lambda$) s'il est à accroissements indépendants, et le nombre de sauts observés pendant la période $[t, t+h]$ suit une loi $\mathcal{P}(\lambda \cdot h)$.




Les données disponibles dans la Base de Données Centralisée des Statistiques (BDCS) s'arrêtent à l'année 2017 pour les informations concernant l'importation des véhicules, mais selon un article de Algérie-eco, environ 8000 véhicules ont été importés par les particuliers durant le premier trimestre 2021, et si on ramène ce chiffre à l'année, cela donne environ 32 000 véhicules par an. Si on utopise le fait que tous ces véhicules souscrivent une assurance Tous Risques, cela reste difficile de concevoir que cela suffise à garder la rentabilité des portefeuilles automobiles des assureurs. Pour aller plus loin, il serait intéressant de recueillir les données sur les nombres de sinistres (qui sont très élevés en Algérie) et les recouper avec le faible nombre de véhicules importés afin de suivre l'évolution du patrimoine automobile assurable.

**Conclusion** – Si la situation ne s'améliore pas et que le marché des véhicules neufs ne redécolle pas, les assureurs vont voir les rentabilités de leurs garanties dommages fortement baisser à l'horizon 2023. Etant donné que les actions que peuvent mener les assureurs sont limitées pour remédier à la crise du marché automobile, il ne leur reste qu'à trouver des stratégies afin de fidéliser leurs assurés et essayer de les garder dans leur portefeuille Tous Risques ou Dommages Collision plus longtemps. A notre sens, la révision de la tarification aura un impact majeur sur la stratégie de fidélisation de l'assuré à la garantie facultative. D'où l'objet de ce mémoire.

**Introduction** – Dans ce chapitre, nous allons présenter le contenu de la base de données et passer en revue toutes les manipulations qui ont été faites sur celle-ci afin de préparer au mieux les données pour une bonne modélisation de la sinistralité automobile du portefeuille de la SAA. Nous allons également proposer quelques variables (informations) supplémentaires qui pourraient s'avérer utiles pour mieux décrire la sinistralité en Algérie.

## 4.16 Objectif de l'étude

Le but de cette étude est d'essayer de modéliser au mieux la sinistralité automobile algérienne. Cette étude permettra une ébauche d'un système de tarification de la garantie Tous Risques mais pourra également servir de base pour les autres types de garanties dommages car le but sera de déterminer la prime pure [36] d'un contrat d'assurance. La prime pure correspond au montant moyen d'un sinistre que devra payer la compagnie d'assurance si le risque survenait. Le calcul de la prime pure a pour but d'évaluer, pour chaque assuré, le montant attendu des sinistres pour la période d'assurance étudiée. Cette évaluation se fait le plus souvent par des méthodes statistiques. La sinistralité est divisée en plusieurs composantes, chacune étant évaluée indépendamment :

— La probabilité d'un sinistre normal.

— Le coût d'un sinistre normal.

— La probabilité d'un sinistre grave.

— Le coût d'un sinistre grave.

S'il est possible de modéliser avec précision le potentiel futur sinistre d'un individu, il est également possible d'utiliser cette information pour le provisionnement d'un dossier sinistre lors de son ouverture. Cette information pourrait être utilisée plus généralement dans les PSAP (Provisions Sinistres A Payer).

# Chapitre 5

# Modélisation linéaire

Nous allons commencer par donner brièvement quelques notions sur la modélisation linéaire avant de parler de modèles généralisés.
La régression linéaire permet le traitement de la fluctuation d'une variable par celui d'une ou plusieurs autres variables. Le modèle linéaire classique, ou modèle linéaire simple, constitue la base des Modèles Linéaires Généralisés (ou GLM : Generalized Linear Models en anglais) et une compréhension approfondie est essentielle pour maîtriser les GLM. Beaucoup de concepts de régression trouvés dans les GLM ont leur genèse dans le modèle linéaire simple. Les distributions des variables réponses rencontrées dans le monde des assurances sont généralement non normales et c'est pour cette raison que les GLM existent. Mais la compréhension de ceux-ci nécessitent des outils que nous verrons dans ce chapitre mais qui ne sont pas utilisables directement en assurance.

## 5.1 Définition

On appelle **modèle linéaire** un modèle statistique qui peut s'écrire sous la forme [8] :

$$Y = \sum_{j=1}^{k} \beta_j X^j + \epsilon.$$

On définit les quantités qui interviennent dans ce modèle :

- $Y$ est une variable aléatoire réelle (v.a.r) que l'on observe et que l'on souhaite expliquer et/ou prédire ; on l'appelle **variable à expliquer** ou **variable réponse** ; On suppose que la variance de $Y$ est constante : c'est ce qu'on appelle l'hypothèse d'homoscédasticité.
- les $k$ variables $X^1, \ldots, X^k$ sont des variables réelles ou discrètes, non aléatoires et également observées ; l'écriture de ce modèle suppose que l'ensemble des $X^j$ est sensé expliquer $Y$ par une relation de cause à effet ; les variables $X^j$ sont appelées **variables explicatives, prédicteurs** ou **régresseurs**.
- les $\beta_j, j = 1, \ldots, k$ sont les paramètres du modèle, non observés, et donc à estimer par des techniques statistiques appropriées.
- $\epsilon$ est le terme d'erreur dans le modèle ; c'est une v.a.r non observée pour laquelle on pose les hypothèses suivantes :
$$\mathbb{E}(\epsilon) = 0, \quad Var(\epsilon) = \sigma^2 > 0.$$
où $\sigma^2$ est un paramètre inconnu à estimer. Une hypothèse sera faite plus tard sur ce point.
- Les hypothèses posées sur $\epsilon$ impliquent les caractéristiques suivantes sur $Y$ :

$$\mathbb{E}(Y) = \sum_{j=1}^{k} \beta_j X^j, \quad Var(Y) = \sigma^2.$$

En moyenne, $Y$ s'écrit donc comme une combinaison linéaire des $X^j$ : la liaison entre les $X^k$ et $Y$ est linéaire. C'est la raison pour laquelle ce modèle est appelé **modèle linéaire.**

## 5.2 Critère de qualité

En réalité, lorsqu'on fait une régression linéaire, on cherche une certaine fonction $f$ telle que :

$$Y \approx f(X).$$

Pour faire une régression, il faudrait définir un critère quantifiant la qualité de l'ajustement de $Y$ par la fonction $f$ sur les données. On suppose que $f$ est une fonction inconnue que l'on cherche et qui est dans une classe de fonctions $\mathcal{G}$. Le problème mathématique s'écrit alors sous cette forme :

$$\underset{f \in \mathcal{G}}{\operatorname{argmin}} \sum_{i=1}^{n} l(y_i - f(x_i)),$$

avec :

— $n$ représente le nombre de données

— $l$ est la **fonction de coût** ou la **fonction de perte**.

La fonction de coût la plus utilisée est la fonction **coût quadratique** : $l(u) = u^2$.

## 5.3 Modèle de régression multiple

### 5.3.1 Généralités

Un modèle de régression multiple sera dorénavant noté sous forme vectorielle [14] :

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1}.$$

avec :

— $Y$ est un vecteur aléatoire de dimension $n$

— $n$ nombre d'observation de la variable aléatoire $Y$

— $X$ est une matrice de taille $n \times p$ connue, appelée matrice du **plan d'expérience** ou matrice **design**. $X$ est la concaténation des $p$ variables $X_j$ : $X = (X_1, X_2 \ldots, X_p)$. Nous noterons la $i^e$ ligne de la matrice $X$ par le vecteur ligne $x_i^{'} = (x_{i1}, \ldots, x_{ip})$

— $\beta$ est le vecteur de dimension $p$ des paramètres inconnus du modèle

— $\epsilon$ est le vecteur centré, de dimension $n$, des erreurs.

Pour pouvoir modéliser une variable aléatoire $Y$, il va falloir faire des suppositions sur la matrice $X$. La première consistera à supposer que la matrice $X$ est de plein rang. On notera cette hypothèse $\mathcal{H}_1$. Dans la plupart des situations et spécialement dans la notre, le nombre d'observations $n$ est supérieur au nombre des paramètres à estimer $p$, alors le rang de la matrice $X$ sera égal à $p$.

Il arrive que parmi les variables explicatives $X_j$, certaines peuvent interagir entre elles. Il est par exemple naturel de penser que la variable `densité d'habitation` soit liée à la variable `localisation`. Pour représenter mathématiquement cette interaction, nous écrivons le produit entre les variables explicatives qui interagissent.

Ce type d'écriture reste un cas de régression linéaire. En effet, nous pouvons considérer que n'importe quelle transformation connue et fixée des variables explicatives (log, exp, produit,

...) rentre dans le modèle de régression linéaire. Ceci est important car pour le traitement des variables continues ou numériques plus généralement, il est souvent question de transformer les variables pour obtenir de meilleures approximations de $\mathbb{E}(Y)$.

### 5.3.2 Estimation des paramètres - Moindres Carrés Ordinaires MCO

On appelle estimateur des (**MCO**) $\widehat{\beta}$ de $\beta$ la valeur suivante :

$$\widehat{\beta} = \underset{\beta_1,\ldots,\beta_p}{\operatorname{argmin}} \sum_{i=1}^{n} \left( y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \|Y - X\beta\|^2.$$

**Théorème 4** (Estimateur des MCO [14]). *Si l'hypothèse $\mathcal{H}_1$ est vérifiée, alors l'estimateur des moindres carrés ordinaires $\widehat{\beta}$ de $\beta$ vaut :*

$$\widehat{\beta} = (X'X)^{-1}X'Y.$$

### 5.3.3 Approche géométrique

Il est intéressant de considérer les vecteurs des observations dans l'espace des variables $\mathbb{R}^n$. Géométriquement, le vecteur $Y$ définit dans $\mathbb{R}^n$ un vecteur $\overrightarrow{OY}$ d'origine $O$ et d'extrémité $Y$ avec pour coordonnées $(y_1, \ldots, y_n)$. La matrice $X$ du plan d'expérience est formée de $p$ vecteurs colonnes. Chaque vecteur $X_j$ définit dans $\mathbb{R}^n$ un vecteur $\overrightarrow{OX_j}$ d'origine $O$ et d'extrémité $X_j$. Ce vecteur a pour coordonnées $(x_{1j}, \ldots, x_{nj})$. Ces $p$ vecteurs linéairement indépendants (hypothèse $\mathcal{H}_1$) engendrent un sous-espace vectoriel de $\mathbb{R}^n$ que l'on notera $\mathfrak{F}(X)$, de dimension $p$.

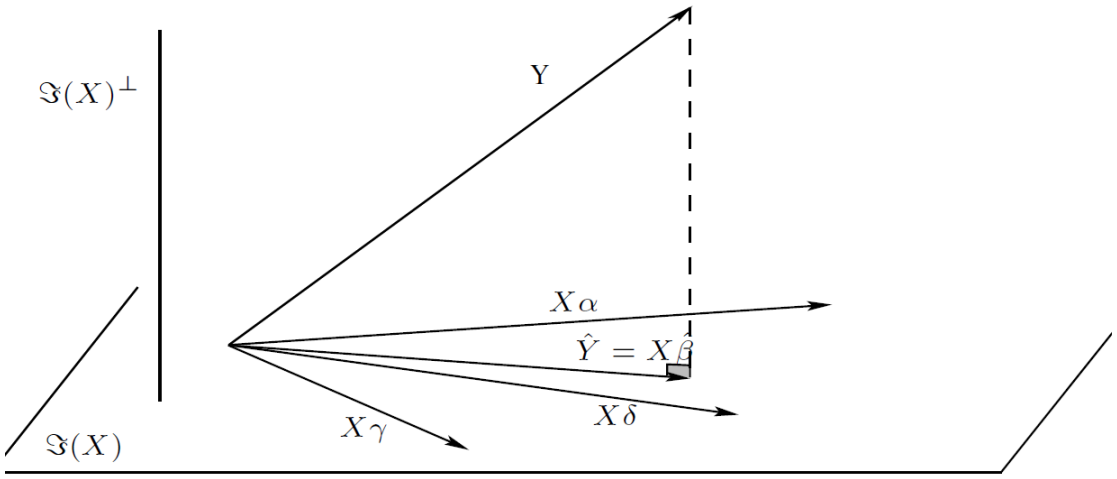



FIGURE 5.1 – Représentation graphique dans l'espace des variables.[14]

L'espace $\mathfrak{F}(X)$ est engendré par les colonnes de $X$. L'espace orthogonal à $\mathfrak{F}(X)$ est noté $\mathfrak{F}(X)^{\perp}$ est quant à lui appelé espace des résidus.

Minimiser $S(\beta) = \sum_{i=1}^{n} \left( y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$ revient à chercher un élément de $\mathfrak{F}(X)$ qui soit le plus proche de $Y$, au sens de la norme euclidienne classique. Par définition, cet unique élément noté $\widehat{Y}$ est appelé projection orthogonale de $Y$ sur $\mathfrak{F}(X)$ noté :

$$\widehat{Y} = P_X Y = X\widehat{\beta}.$$

La matrice $P_X$ est la matrice de projection orthogonale sur $\mathfrak{F}(X)$ ou **hat matrix** et $\widehat{\beta}$ est l'estimateur des moindres carrés de $\beta$. Le vecteur $\widehat{Y}$ contient les valeurs ajustées de $Y$ par le modèle.

**Proposition 7.** *[14] L'estimateur des moindres carrés $\widehat{\beta}$ est un estimateur sans biais de $\beta$ et sa variance vaut $Var(\widehat{\beta}) = \sigma^2(X'X)^{-1}$*

**Théorème 5** (Gauss-Markov [14])**.** *L'estimateur des MCO est optimal parmi les estimateurs linéaires sans biais de $\beta$.*

### 5.3.4 Résidus

Les résidus sont définis par la relation suivante :

$$\widehat{\epsilon} = Y - \widehat{Y} = (I - P_X)Y = P_{X^\perp}Y.$$

Les résidus appartiennent donc à $\mathfrak{F}(X)^\perp$. On peut réécrire les résidus sous la forme suivante :

$$\widehat{\epsilon} = P_{X^\perp}Y = P_{X^\perp}(X\beta + \epsilon) = P_{X^\perp}\epsilon.$$

Ici $\epsilon$ est appelé le terme d'erreur ou résidus. La terminologie **erreur** est trompeuse : il n'y a aucune incidence que la déviation de Y du modèle théorique est de quelque manière que ce soit erronée. En économétrie, le terme **perturbation** est utilisé.

D'autres hypothèses sur la modélisation sont imposées pour pouvoir faire une régression linéaire. On admettra donc que :

- **Homoscédasticité** – Hypothèse $\mathcal{H}_2$ : la variance de $\epsilon$ est finie et ne varie pas avec les variables $x$, $Var(\epsilon) = \sigma^2$
- **Normalité et indépendance** – Hypothèse $\mathcal{H}_3$ : La distribution de $\epsilon$ est normale. De plus, les observations $y_i$ sont indépendantes. On écrira $\epsilon \sim \mathcal{N}(0, \sigma^2)$

**Proposition 8.** *[14] Sous les hypothèses $\mathcal{H}_1$ et $\mathcal{H}_2$, nous avons :*

$$\begin{array}{ll} \mathbb{E}(\widehat{\epsilon}) = P_{X^\perp}(\epsilon) = 0, & Var(\widehat{\epsilon}) = \sigma^2 P_{X^\perp}, \\ \mathbb{E}(\widehat{Y}) = X\beta, & Var(\widehat{Y}) = \sigma^2 P_X, \\ Cov(\widehat{\epsilon}, \widehat{Y}) = 0. & \end{array}$$

Il est également possible de compléter la figure 5.1 pour visualiser les résidus géométriquement.

Le théorème de Pythagore donne directement l'égalité suivante :

$$\|Y - \overline{y}\,\mathbb{1}\|^2 = \|\widehat{Y}\|^2 + \|\widehat{\epsilon}\|^2.$$

On définit le coefficient de détermination multiple $R^2$ comme suit :

$$R^2 = \frac{\|\widehat{Y} - \overline{y}\,\mathbb{1}\|^2}{\|Y - \overline{y}\,\mathbb{1}\|^2} = \cos^2\theta = \frac{\text{Variation expliquée par le modèle}}{\text{Variation totale}}.$$

Le coefficient de détermination est une mesure permettant de quantifier la qualité d'un modèle. Plus elle se rapproche de 1, mieux le modèle modélise la variable à expliquer $Y$. Ce coefficient ne tient pas compte de la dimension de $\mathfrak{F}(X)$, ce qui est problématique car l'ajout de variables augmente le $R^2$ mais augmente également la complexité du modèle. (*cf Coefficient de détermination ajusté*). Nous verrons plus tard d'autres mesures de la qualité d'un modèle.

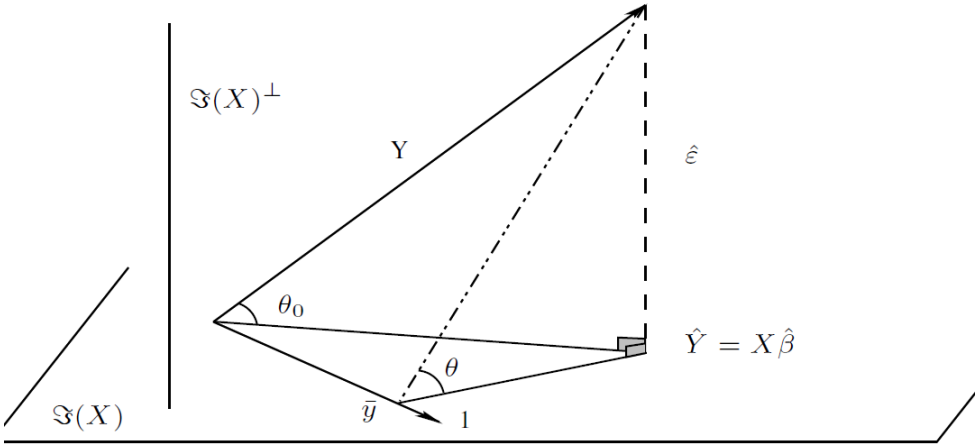



FIGURE 5.2 – Représentation graphique des résidus dans l'espace des variables.[14]

**Proposition 9.** *[14] La statistique que l'on appelle* ***somme des carrés résiduels*** *ou* ***SCR*** *définit un estimateur sans biais de* $\sigma^2$

$$\widehat{\sigma}^2 = \frac{\|\widehat{\epsilon}\|^2}{n-p} = \frac{SCR}{n-p}.$$

Ce dernier estimateur nous permet alors de créer un estimateur de la variance de $\widehat{\beta}$

$$\widehat{\sigma}^2_{\widehat{\beta}} = \widehat{\sigma}^2 (X'X)^{-1} \Rightarrow \widehat{\sigma}_{\widehat{\beta}_j} = \sqrt{\widehat{\sigma}^2 [(X'X)^{-1}]_{jj}}.$$

### 5.3.5 Prévision

Un des plus importants buts de la régression est de pouvoir proposer des prévisions pour la variable à expliquer $y$ lorsque nous avons de nouvelles valeurs de $x$. Soit une nouvelle valeur que l'on notera $x'_{n+1} = (x'_{n+1,1}, x'_{n+1,2}, \ldots, x'_{n+1,p})$. Nous souhaitons alors prédire la valeur de $y$. On a alors alors :

$$y_{n+1} = x'_{n+1}\beta + \epsilon_{n+1},$$

avec $\mathbb{E}(\epsilon_{n+1}) = 0, \quad V(\epsilon_{n+1}) = \sigma^2$ et $Cov(\epsilon_{n+1}, \epsilon_i) = 0$ pour $i = 1, \ldots, n$.

Il est possible de prédire la valeur correspondante grâce au modèle ajusté

$$\widehat{y} = x'_{n+1}\widehat{\beta},$$

Cependant, la variance de la prévision n'est pas égale à celle des $\widehat{y}$. En effet, deux erreurs viennent perturber l'exactitude de la prévision, la première due à l'incertitude sur $\epsilon_{n+1}$ et l'autre à l'incertitude due à l'estimation. Calculons la variance de l'erreur de prévision :

$$\mathrm{V}\left(y_{n+1} - \hat{y}^p_{n+1}\right) = \mathrm{V}\left(x'_{n+1}\beta + \varepsilon_{n+1} - x'_{n+1}\hat{\beta}\right) = \sigma^2 + x'_{n+1} V(\hat{\beta}) x_{n+1} = \sigma^2 \left(1 + x'_{n+1}\left(X'X\right)^{-1} x_{n+1}\right).$$

Ce résultat permet clairement de voir l'erreur due à $\sigma^2$ à laquelle vient s'ajouter l'incertitude d'estimation.

### 5.3.6 Inférence statistique : tests

Grâce à l'hypothèse gaussienne $\mathcal{H}_3$ , il est possible d'établir les résultats suivants [14].

**Proposition 10.** *[14] $(\widehat{\beta},\widehat{\sigma}^2)$ est une statistique complète et $(\widehat{\beta},\widehat{\sigma}^2)$ est de variance minimale dans la classe des estimateurs sans biais.*

**Proposition 11** (variance connue [14])**.** *Sous l'hypothèse $\mathcal{H}_1$ et $\mathcal{H}_3$, nous avons :*

- *$\widehat{\beta}$ est un vecteur gaussien de moyenne $\beta$ et de variance $\sigma^2(X'X)^{-1}$,*
- *$(n-p)\widehat{\sigma}^2/\sigma^2$ suit un $\chi^2_{n-p}$ à $n-p$ ddl,*
- *$\widehat{\beta}$ et $\widehat{\sigma}^2$ sont indépendants.*

Or dans la réalité, la variance est rarement connue, on utilise alors son estimateur sans biais $\widehat{\sigma^2}$. On a alors la proposition suivante.

**Proposition 12** (variance inconnue [14])**.** *Sous l'hypothèse $\mathcal{H}_1$ et $\mathcal{H}_3$, nous avons :*

- *pour $j = 1,\dots,p$, on a : $T_j = \dfrac{\widehat{\beta}_j - \beta_j}{\sqrt{\widehat{\sigma}^2\left[(X'X)^{-1}\right]_{jj}}} \sim \mathcal{T}(n-p)$,*
- *Soit $R$ une matrice de taille $q \times p$ de rang $q, q \leq p$, alors*

$$\frac{1}{q\widehat{\sigma}^2}(R(\widehat{\beta}-\beta))'\left[R(X'X)^{-1}R'\right]^{-1}R(\widehat{\beta}-\beta) \sim \mathcal{F}_{q,n-p},$$

*$\mathcal{T}$ étant la loi de Student et $\mathcal{F}$ celle de Fischer.*

**Utilisation du test de Student** – En pratique, dans la modélisation [16], il est souvent question de faire les tests d'hypothèses suivants : $H_0 : \beta_j = \beta_j^0$ constituant l'hypothèse nulle contre $H_1 : \beta_j \neq \beta_j^0$ avec $\beta_j^0$ est une valeur supposé de la valeur de $\beta_j$. Alors la statistique

$$T_j = \frac{\widehat{\beta}_j - \beta_j}{\sqrt{\widehat{\sigma}^2\left[(X'X)^{-1}\right]_{jj}}}.$$

est utilisée pour faire ce test d'hypothèse.
Chaque test de ce genre sur les $j = 1,\dots,p$ permet de déterminer si réellement la valeur de $\beta_j$ est statistiquement et de façon significative différente de $\beta_j^0$.

Supposons que nous voulons savoir si la variable explicative $x_j$ est significative et est bien corrélée à la variable réponse $y$, alors il suffit de poser $\beta_j^0 = 0$, calculer la statistique de Student et conclure sur la significativité du coefficient $\beta_j$.

**Utilisation du test de Fischer** – Supposons que nous souhaitons tester la nullité simultanée des $q$ derniers coefficients du modèle avec $q \leq p$, le problème s'écrit alors de la façon suivante :

$$H_0 : \beta_{q-p+1} = \cdots = \beta_p = 0 \qquad \text{contre} \qquad H_1 : \exists j \in \{p-q+1,\dots,p\} : \beta_j \neq 0.$$

Ce test est en réalité un moyen de déterminer s'il n'existe pas un modèle emboité avec $q$ variables explicatives et qui possède le même pouvoir prédictif et le même ajustement que le modèle complet à $p$ régresseurs. On a le théorème suivant pour faire ce test utilisant la statistique de Fischer.

**Théorème 6.** *[14] Soit un modèle de régression à $p$ variables $Y = X\beta + \epsilon$ satisfaisant $\mathcal{H}_1$ et $\mathcal{H}_3$. Nous souhaitons tester la validité d'un sous-modèle (ou modèle emboîté) où un ou plusieurs coefficients sont nuls. Le plan d'expérience privé de ces variables sera noté $X_0$, les $p_0$ colonnes de $X_0$ engendreront un sous-espace noté $\mathfrak{F}_0$ et le sous-modèle sera $Y = X_0\beta_0 + \epsilon$. Notons l'hypothèse nulle (modèle restreint) $H_0 : \mathbb{E}(Y) \in \mathfrak{F}_0$ et l'hypothèse alternative (modèle complet) $H_1 : \mathbb{E}(Y) \in \mathfrak{F}(X)$. Pour tester ces deux hypothèses, nous utilisons la statistique de test F ci-dessous qui possède comme loi sous $H_0$ :*

$$F = \frac{\|\widehat{Y}_0 - \widehat{Y}\|^2/(p - p_0)}{\|Y - \widehat{Y}\|^2/(n - p)} \sim \mathcal{F}_{p-p_0,n-p}.$$

*Il est également possible d'écrire de façon équivalente cette statistique :*

$$F = \frac{n - p}{p - p_0}\frac{SCR_0 - SCR}{SCR} \sim \mathcal{F}_{p-p_0,n-p}.$$

*L'hypothèse $H_0$ sera repoussée en faveur de $H_1$ si l'observation de la statistique F est supérieure à $f_{p-p_0,n-p}(1 - \alpha)$. La valeur $\alpha$ est le niveau du test et $f$ la répartition de la loi de Fischer.*

## 5.4 Choix des variables

Cette étape de la régression est très importante pour tout utilisateur ayant à manipuler plusieurs variables explicatives. En assurance automobile par exemple, la base de données des assureurs est pleine d'informations qui sont peut être inutiles et il est primordial pour tout utilisateur de régression de pouvoir identifier les seules variables qui rentrent dans l'explicabilité de la réponse $y$ afin de réduire la complexité des modèles et peut être diminuer drastiquement les temps de calculs de ceux-ci. Dans notre cas malheureusement, c'est le manque de variables qui affectera la qualité du modèle. En effet, manquer de variables explicatives peut s'avérer plus couteux en termes d'erreurs que d'en prendre certaines qui n'apportent rien au modèle. C'est pour cela que nous avons consacré la section 5.5.2 à proposer de nouvelles informations à demander aux assurés lors de la souscription des contrats.

L'objectif de la sélection des variables est de déterminer au mieux l'ensemble des variables explicatives pertinentes $X_j$ tel que leurs coefficients $j$ soient non nuls dans le modèle.

### 5.4.1 Choix incorrect de variables

Il faut savoir que faire un mauvais choix de variables explicatives peut avoir des conséquences sur la qualité des modèles. Par « mauvais choix », il faut comprendre soit en prendre trop peu, soit en prendre le bon nombre mais pas les bonnes, soit en prendre trop. Admettons que nous ayons trois variables explicatives potentielles $X_1$, $X_2$ et $X_3$ et que le vrai modèle soit :

$$Y = \beta_1 X_1 + \beta_2 X_2 + \epsilon = X_{12}\beta_{12} + \epsilon.$$

La variable $X_3$ n'intervient pas dans le modèle pour prédire $Y$, mais cette information n'est pas disponible. Avec trois variables, nous pouvons donc produire $(2^3 - 1)$ modèles différents, trois modèles à une variable, trois modèles à deux variables et un modèle à trois variables.

### 5.4.2 Biais des estimateurs

L'analyse des sept modèles produit les résultats sur les biais des estimateurs suivants :

| modèle | estimations | propriétés |
|---|---|---|
| $Y_1 = X_1\beta_1 + \varepsilon$ | $\hat{Y}_1 = X_1\hat{\beta}_1$ | $B(\hat{Y}_1) = -P_{X_1^\perp}X_2\beta_2$ |
| | $\hat{\sigma}_1^2 = \frac{\|P_{X_1^\perp}Y\|^2}{n-1}$ | $B(\hat{\sigma}_1^2) = \frac{1}{n-1}\beta_2^2\|P_{X_1^\perp}X_2\|^2$ |
| $Y = X_2\beta_2 + \varepsilon$ | $\hat{Y}_2 = X_2\hat{\beta}_2$ | $B(\hat{Y}_2) = -P_{X_2^\perp}X_1\beta_1$ |
| | $\hat{\sigma}_2^2 = \frac{\|P_{X_2^\perp}Y\|^2}{n-1}$ | $B(\hat{\sigma}_2^2) = \frac{1}{n-1}\beta_1^2\|P_{X_2^\perp}X_1\|^2$ |
| $Y = X_3\beta_3 + \varepsilon$ | $\hat{Y}_3 = X_3\hat{\beta}_3$ | $B(\hat{Y}_3) = -P_{X_3^\perp}X_{12}\beta_{12}$ |
| | $\hat{\sigma}_3^2 = \frac{\|P_{X_3^\perp}Y\|^2}{n-1}$ | $B(\hat{\sigma}_3^2) = \frac{1}{n-1}\beta_{12}'X_{12}'P_{X_{12}^\perp}X_{12}\beta_{12}$ |
| $Y = X_{12}\beta_{12} + \varepsilon$ | $\hat{Y}_{12} = X_{12}\beta_{12}$ | $B(\hat{Y}_{12}) = 0$ |
| | $\hat{\sigma}_{12}^2 = \frac{\|P_{X_{12}^\perp}Y\|^2}{n-2}$ | $B(\hat{\sigma}_{12}^2) = 0$ |
| $Y = X_{13}\beta_{13} + \varepsilon$ | $\hat{Y}_{13} = X_{13}\hat{\beta}_{13}$ | $B(\hat{Y}_{13}) = -P_{X_{13}^\perp}X_{12}\beta_{12}$ |
| | $\hat{\sigma}_{13}^2 = \frac{\|P_{X_{13}^\perp}Y\|^2}{n-2}$ | $B(\hat{\sigma}_{13}^2) = \frac{1}{n-2}\beta_{12}'X_{12}'P_{X_{13}^\perp}X_{12}\beta_{12}$ |
| $Y = X_{23}\beta_{23} + \varepsilon$ | $\hat{Y}_{23} = X_{23}\hat{\beta}_{23}$ | $B(\hat{Y}_{23}) = -P_{X_{23}^\perp}X_{12}\beta_{12}$ |
| | $\hat{\sigma}_{23}^2 = \frac{\|P_{X_{23}^\perp}Y\|^2}{n-2}$ | $B(\hat{\sigma}_{23}^2) = \frac{1}{n-2}\beta_{12}'X_{12}'P_{X_{23}^\perp}X_{12}\beta_{12}$ |
| $Y = X_{123}\beta_{123} + \varepsilon$ | $\hat{Y}_{123} = X_{123}\hat{\beta}_{123}$ | $B(\hat{Y}_{123}) = 0$ |
| | $\hat{\sigma}_{123}^2 = \frac{\|P_{X_{123}^\perp}Y\|^2}{n-3}$ | $B(\hat{\sigma}_{123}^2) = 0$ |

FIGURE 5.3 – Biais des différents estimateurs. [14]

Nous constatons alors que dans les modèles «trop petits» à une variable, c'est-à-dire admettant moins de variables que le modèle « correct » inconnu du statisticien, les estimateurs obtenus sont biaisés. A l'inverse, lorsque les modèles sont « trop grands » (ici à 3 variables), les estimateurs ne sont pas biaisés. Il semblerait donc qu'il vaille mieux travailler avec des modèles « trop grands » afin d'éviter de créer du biais en enlevant des variables explicatives qu'on estimerait inutiles.

### 5.4.3 Variance des estimateurs

Cependant, ce dont nous venons de parler ne concerne que les biais des estimateurs. L'autre aspect de la qualité d'une régression est de mesurer la variance des estimateurs. Le calcul des variances selon les différents modèles donne :

| Modèle | Variance |
|---|---|
| $Y = X_1\beta_1 + \epsilon$ | $V(\widehat{Y}_1) = P_{X_1}\sigma^2$ |
| $Y = X_{12}\beta_{12} + \epsilon$ | $V(\widehat{Y}_{12}) = P_{X_{12}}\sigma^2 = P_{X_1}\sigma^2 + P_{X_2\cap X_1^\perp}\sigma^2$ |
| $Y = X_{123}\beta_{13} + \epsilon$ | $V(\widehat{Y}_{123}) = P_{X_{123}}\sigma^2 = P_{X_1}\sigma^2 + P_{X_{23}\cap X_1^\perp}$ |

**Conclusion** – La variance des données ajustées dans le modèle le plus petit est plus faible que celle des données ajustées dans le modèle le plus grand. C'est pour cette raison que l'utilisateur

de la régression doit "rechercher" le meilleur modèle en évitant de passer à coté de variables explicatives pour éviter de créer du biais, et aussi l'insertion de variables inutiles pour éviter d'augmenter la variance des données ajustées par le modèle.

### 5.4.4 Régression linéaire pondérée

La régression pondérée est une méthode pouvant être utilisée lorsque l'hypothèse de variance constante dans les valeurs résiduelles pour les moindres carrés est contredite (hétéroscédasticité). Avec une pondération adaptée, cette procédure minimise la somme des carrés des valeurs résiduelles pondérés, de manière à générer des valeurs résiduelles présentant une variance constante (on parle aussi d'homoscédasticité).
Ainsi dans la suite $\omega_i$ sera la pondération de chaque observation et $\mathbf{\Omega} = \text{diag}(\omega)$.

Au lieu de $\sum_{i=1}^{n} \left(y_i - \mathbf{x}_i^\top \beta\right)^2$, on considère $\sum_{i=1}^{n} \omega_i \left(y_i - \mathbf{x}_i^\top \beta\right)^2$,

$$W(\boldsymbol{\beta}) = (\mathbf{Y} - \mathbf{X}\beta)^\top \mathbf{\Omega}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}), \quad \mathbf{\Omega} = \text{diag}(\omega),$$

$$\frac{\partial W(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2\mathbf{X}^\top \mathbf{\Omega}\mathbf{Y} + 2\mathbf{X}^\top \mathbf{\Omega}\mathbf{X}\beta,$$

et

$$\frac{\partial^2 W(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} = 2\mathbf{X}^\top \mathbf{\Omega}\mathbf{X}.$$

Aussi

$$\frac{\partial W(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2\mathbf{X}^\top \mathbf{\Omega}\mathbf{Y} + 2\mathbf{X}^\top \Omega\mathbf{X}\boldsymbol{\beta} = \mathbf{0} \text{ si } \widehat{\beta} = \left(\mathbf{X}^\top \Omega\mathbf{X}\right)^{-1} \mathbf{X}^\top \Omega\mathbf{Y},$$

De plus, si on pose maintenant :

$$\Omega^{1/2} = \text{diag}(\sqrt{\omega}),$$
$$\tilde{\mathbf{X}} = \Omega^{1/2}\mathbf{X} \quad \text{et} \quad \tilde{\mathbf{Y}} = \Omega^{1/2}\mathbf{Y},$$

nous avons alors :

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^\top \Omega\mathbf{X}\right)^{-1} \mathbf{X}^\top \Omega\mathbf{Y} = \left(\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}}\right)^{-1} \tilde{\mathbf{X}}^\top \tilde{\mathbf{Y}}.$$

## 5.5 Base de données et préparation des données

Les données qui ont été mises à notre disposition ont d'abord subi une consolidation pour n'en faire qu'une seule table, puis ont subi un nettoyage. Malheureusement, un fait inhérent aux systèmes d'informations algériens est la mauvaise construction des bases de données et des informations qu'elles contiennent. Il est clair et admis de tout le monde que la qualité de saisie des données n'est malheureusement pas optimale dans l'administration algérienne, et cela affectera grandement l'étude car les fausses données créent du biais et peuvent mener à de fausses conclusions. Il est donc primordial avant toute étude de mener une exploration des données pour chercher et traiter ces mauvaises données. Pour cette étape, il n'y a malheureusement pas de méthode universelle ayant une certaine efficacité dans le traitement des données manquantes, mal saisies ou aberrantes. Cette partie de l'étude est en général la plus coûteuse en temps car il n'est pas possible de savoir vraiment ce que l'on doit chercher comme mauvaises informations. Ensuite, il s'agira de décider quoi faire de ces mauvaises données. La solution la plus facile serait bien sûr de les éliminer, mais il serait préférable d'essayer de trouver un moyen d'expliquer ou de corriger ces données car elles pourraient appartenir à une sous population qui ne possèdent

pas beaucoup d'observations. Une autre option serait d'attribuer une valeur moyenne aux valeurs manquantes ou une valeur aléatoire de la variable en question. De plus la modification d'une donnée revient à supposer la survenue d'un évènement qui n'a pas forcément eu lieu et la suppression revient à supposer que l'évènement n'a pas eu lieu. Si ce travail de récupération est mené sur beaucoup de données, il serait nécessaire de procéder à un traitement automatique des données à réviser.

Dans notre étude, par défaut de temps, nous avons dû éliminer toute observation paraissant aberrante.

### 5.5.1 Contenu de la base de données et ses variables

La base de données contient les informations de toutes les polices d'assurances souscrites auprès de la SAA dans le cadre de la garantie Tous Risques durant les années 2014 à 2018. Avant nettoyage, il y avait au total quelques 850 000 observations. Le fichier qui nous a été remis en comporte 740 000 car toutes les observations où des informations importantes ont été omises ou mal saisies ont dues être retirées. La table contient la liste des polices d'assurances de deux types de catégories de véhicules : les véhicules particuliers ou les véhicules appartenant à des flottes. Il n'est bien entendu pas possible d'étudier ces deux catégories de la même façon car selon une analyse descriptive, ces deux types de catégories présentent une sinistralité très différente, en plus de ne pas contenir les mêmes variables explicatives ; les véhicules des flottes ne sont pas supposés appartenir à des individus à qui l'on pourrait demander un âge, un sexe, un âge de permis, . . .

La base de données comporte 40 variables mais ne sont pas toutes significatives. Il sera question de faire une étude sur la fréquence des sinistres puis une étude sur la gravité des sinistres. Ces deux études portent donc sur la variable `sinistres` qui est le nombre de sinistres survenus durant la période d'exposition d'une police d'assurance, et la seconde variable `ChargeSin` qui est le montant que l'assureur a dû payer en totalité à son assuré pour les sinistres survenus durant la période d'exposition.

La liste des variables utiles sont les suivantes :

1. `id` : chaîne de caractères servant de clé d'identification de chaque police d'assurance. Elle est la concaténation du numéro d'agence, du numéro de police, du code d'appartenance à l'une des catégories citées plus haut.
2. `dr` : variable catégorielle. La direction régionale auquel appartient l'agence dans laquelle le contrat d'assurance a été souscrit. Cette variable est très importante car elle nous servira de variable spatiale, permettant une segmentation géographique du portefeuille. Cette variable contient 15 modalités mais on ne retiendra que 14. La modalité `annaba` a été retirée car elle ne contient que 700 observations, ce qui ne représente pas la quantité d'observations réelles qui se situe à 50 000. L'extraction des données de la direction régionale de Annaba est incorrecte. Ceci est problématique car beaucoup des régions du Nord-Est du pays dépendent de cette direction régionale.
3. `agence` : variable catégorielle. Code de l'agence dans laquelle la police d'assurance a été souscrite. Pour l'instant, nous n'avons pas utilisé cette variable car c'est une variable catégorielle et elle possède 592 modalités (592 agences SAA en Algérie en 2018) et cela représenterait un énorme coût en temps de calcul. Il serait judicieux d'utiliser cette variable dans une étude future car on pourrait supposer que la plupart des individus souscrivent une assurance dans une agence près de leur domicile et on pourrait donc peut être associer la zone de circulation du véhicule à la localisation de l'agence (segmentation par Daïras-Communes). Une autre idée serait d'utiliser les codes d'agences des contrats d'assurances

pour déduire les densités d'habitations de certaines régions car cette variable peut intuitivement et conceptuellement aider à expliquer la sinistralité.

4. `categorie` : variable catégorielle. Cette variable catégorielle possède deux modalités déjà citées précédemment. Elle différencie les contrats des flottes de ceux des véhicules particuliers. Cette variable permettra de créer deux bases de données sur lesquelles nous travaillerons : `basePartic` et `baseFlotte`
5. `annee` : valeur entière. Représente l'année de souscription du contrat. Etant donné l'instabilité de l'économie algérienne, il sera important d'utiliser cette variable pour actualiser les valeurs des montants des sinistres et des valeurs assurées des véhicules. Plus de détails sur ce point par la suite.
6. `duree` : variable catégorielle à deux niveaux. Période d'exposition au risque d'un contrat : 6 mois ou une année.
7. `police` : chaîne de caractères. Numéro de police du contrat.
8. `avenant` : valeur entière. Représente le numéro d'avenant du contrat. Par définition, un avenant est un : « acte par lequel on modifie les termes d'un contrat ». En assurance, plus précisément un avenant est une révision du contrat pour corriger, apporter des modifications, ajouter ou retirer des garanties ou bien renouveler celui-ci. La variable suivante a été créée à partir de `avenant`
9. `cutAvenant` : variable catégorielle. Variable déduite de la précédente possédant deux modalités : `new` et `old` permettant de différencier les nouvelles polices d'assurances des anciennes. Elle permettra de séparer les nouveaux assurés de l'agence de ceux qui sont à la SAA depuis au moins 1 an, en supposant que la proportion des avenants de modification liés à des rectifications des nouveaux contrats est faible. L'idée derrière la création de cette variable est de savoir s'il existe bien une tendance des nouveaux contrats à être plus sinistrés que ceux des assurés fidèles ou de savoir s'il y a un phénomène de souscription à la garantie Tous Risques dans un seul but : réparer son véhicule. La création de cette variable est intéressante mais mérite une étude à elle seule car la séparation en 2 modalités exclue le cas des contrats RC renouvelés pendant seulement une année pour une Tous Risques.
10. `code` : valeur entière. Code risque. Un contrat d'assurance peut porter sur plusieurs véhicules et le code risque liste ceux-ci, notamment pour les contrats des flottes.
11. `sex` : variable catégorielle. Sexe du souscripteur de la police d'assurance. Variable pas très intéressante comme on le verra plus tard dans l'étude. L'interprétation sera donnée plus tard. Variable indisponible dans `BaseFlotte`.
12. `age` : variable continue. Age du souscripteur du contrat. Variable indisponible dans `BaseFlotte`.
13. `permis` : variable continue. Ancienneté du permis en années. Variable indisponible dans `BaseFlotte`.
14. `nvPermis` : variable catégorielle. Variable permettant de séparer les assurés avec un nouveau permis ou non. Variable indisponible dans `BaseFlotte`. Cetta variable peut être déduite de la précédente en séparant les anciennetés de permis `0` des autres.
15. `marque` : variable catégorielle avec 113 modalités. Elle représente la marque du véhicule.
16. `brand` : variable catégorielle à 56 modalités représentant les marques mais regroupées selon une certaine logique. Par exemple les marques `RENAULT` et `DACIA` sont regroupées en un groupe `RENAULT-DACIA`.
17. `ageVeh` : variable continue. Représente l'âge du véhicule en années.
18. `genre` : variable catégorielle avec 12 modalités. Elle représente le genre de véhicule ; exemples : Transport Publique de Marchandises, Véhicules de plus de 3, 5 tonnes, Tracteurs Routiers, ...

19. `usage` : variable catégorielle avec 10 niveaux. L'usage auquel le véhicule est supposé être assuré ; exemples : affaire, fonctionnaire, location, taxi, . . .
20. `Fiscal` : variable continue. Puissance fiscale du véhicule.
21. `zone` : variable catégorielle à deux modalités `nord` et `sud`. Variable permettant de séparer les véhicules sensés être utilisés dans le nord ou dans le sud du pays.
22. `inflam` : variable catégorielle à deux modalités permettant de séparer les véhicules transportant des marchandises inflammables ou non.
23. `tauxRed` : variable numérique. Représente un taux de réduction appliqué sur la prime demandée pendant la souscription.
24. `PleinTarifAct` : variable continue. Représente la prime Tous Risques avant application du taux de réduction. Ce montant est actualisé selon l'année de souscription du contrat.
25. `prime_trAct` : variable continue représentant le montant de la prime exact payé par l'assuré pour la garantie Tous Risques. Réduction appliquée et montant actualisé.
26. `sinistres` : valeur entière. Variable sur laquelle portera une des études dans ce mémoire. Représente le nombre de sinistres reportés durant la période d'exposition d'un contrat.
27. `VA` : variable continue. Représente la valeur assurée du véhicule. Ce montant est demandé par l'assuré pendant la souscription du contrat. Ce montant a été actualisé.
28. `ChargeSin` : variable continue. Deuxième variable sur laquelle portera l'étude. Représente le montant total des déboursements qu'a dû faire l'assureur sur un contrat concernant uniquement la garantie Tous Risques. Valeur actualisée.

```
         dr              agence           categorie       annee          duree          avenant           code
tizi ouzou:123561   1501    : 12969   Flotte     :122357   2014:163711   Min.   :0.5833   Min.   :  0.000   Min.   :   1.00
alger1     : 77964   2056    : 11621   Particulier:612779   2015:161517   1st Qu.:1.0000   1st Qu.:  0.000   1st Qu.:   1.00
setif      : 65976   2001    :  7916                        2016:149949   Median :1.0000   Median :  2.000   Median :   1.00
alger2     : 64149   1252    :  7248                        2017:134758   Mean   :0.9002   Mean   :  5.111   Mean   :  34.45
alger3     : 63467   2010    :  6865                        2018:125201   3rd Qu.:1.0000   3rd Qu.:  5.000   3rd Qu.:   1.00
mouzaia    : 59114   2552    :  6772                                      Max.   :1.0000   Max.   :339.000   Max.   :3926.00
(Other)    :280905   (Other):681745
sex             age             permis          nvPermis                  brand             marque          ageVeh
 :   704   Min.   :18.00    Min.   : 0.0    flotte:122357   RENAULT-DACIA  :168231   RENAULT   :111234   Min.   : 0.000
F: 43260   1st Qu.[redacted]   1st Qu.:[redacted]    1      :  6787   PEUGEOT        : 88153   PEUGEOT   : 88153   1st Qu.:[redacted]
M:691172   Median           Median :        0      :605992   MARQUE CHINOISE: 67570   HYUNDAI   : 59801   Median :
           Mean             Mean   :                         HYUNDAI        : 59801   DACIA     : 56997   Mean   :
           3rd Qu.          3rd Qu.:                         TOYOTA         : 51684   TOYOTA    : 51684   3rd Qu.:
           Max.   :85.00    Max.   :67.0                     VOLKSWAGEN     : 41080   VOLKSWAGEN: 41080   Max.   :14.000
           NA's   :122357   NA's   :122357                   (Other)        :258617   (Other)   :326187
 genreComplet     usageComplet        Fiscal        zone        inflam     PleinTarifAct        tauxRed          sinistres
VP     :610979   aff    :413466   Min.   : 0.000   N:692200   N:733102   Min.   :   1250   Min.   :0.0000   Min.   :0.0000
TPMm2  : 32056   com    :123523   1st Qu.: 5.000   S: 42936   O:  2034   1st Qu.[redacted]   1st Qu.:[redacted]   1st Qu.:[redacted]
V3.5   : 31543   tax    : 48199   Median : 6.000                         Median           Median :         Median :
TPV    : 22733   TPM    : 44863   Mean   : 7.588                         Mean             Mean   :         Mean   :
TR     : 15200   comB   : 34623   3rd Qu.: 8.000                         3rd Qu.          3rd Qu.:         3rd Qu.:1.0000
TPMp2  :  9427   TPV    : 22733   Max.   :99.000                         Max.   :9828225   Max.   :0.9500   Max.   :7.0000
(Other): 13198   (Other): 47729
      VA              chargeSin
Min.   :    25000   Min.   :       0
1st Qu.[redacted]   1st Qu.:[redacted]
Median           Median
Mean             Mean
3rd Qu.          3rd Qu.
Max.   :196564500   Max.   :18258658
```

[redacted] Certaines données sont masquées pour des raisons de confidentialité

FIGURE 5.4 – Statistiques de la base de données.

### 5.5.2 Variables explicatives manquantes

La liste des variables contenues dans cette base de données est conséquente mais il manque malheureusement de précieuses informations qui à notre sens pourraient aider grandement à

mieux prédire la fréquence et la gravite des sinistres. Une liste non exhaustive des variables qui pourraient apporter une amélioration à la modélisation de la sinistralité :

— `localisation` : Déduire la ville dans lequel est supposé circuler le véhicule en fonction de la localité de l'agence dans lequel a été souscrit le contrat. Pour améliorer la qualité de cette donnée, il faudrait faire correspondre également l'usage du véhicule. Car un véhicule assuré à un endroit peut être utilisé pour aller à un travail situé dans une autre ville. Cas fréquent dans le centre-nord du pays. Enfin l'ajout de cette donnée pourra déterminer la densité de population de la région dans laquelle le véhicule devra circuler.

— `profession` : cette variable existe dans les bases de données mais n'a pas été retenue car jugée de mauvaise qualité. Une idée serait d'utiliser la carte grise des véhicules pour déterminer correctement cette donnée.

— `dateSinistre` : la date pourrait renseigner sur une certaine saisonnalité des sinistres notamment sur la fréquence. On pourrait partitionner cette variable en mois ou en trimestre pour déterminer les périodes de l'année les plus sinistrées. Par exemple le mois de ramadan ou la période estivale.

— `energie` : le type de carburant ou énergie utilisée par les véhicules pour fonctionner. Cette donnée existe dans les bases de données de la SAA mais a été jugée inutile. Nous ne partageons pas ce point de vue car elle pourrait donner un indice sur l'utilisation du véhicule. En effet, en général, un véhicule acheté dans le but de rouler beaucoup est de type Diesel ou GPL. D'autres interprétations pourraient être faites en fonction des marques et genres de véhicules.

— `kilométrage` : le kilométrage du véhicule pourrait renseigner sur la manière dont est utilisé le véhicule et son ancienneté. De plus, si cette valeur est renseignée sur au moins deux années consécutives, il serait possible de déterminer le nombre de km que le véhicule parcourt sur une période d'exposition.

— `enfantPlusDe18` : variable intéressante permettant de savoir si un assuré a au moins un enfant de plus de 18 ans. Il est très rare qu'un jeune conducteur avec un nouveau permis souscrive un contrat d'assurance en son nom. Généralement, les jeunes conducteurs utilisent le véhicule des parents et aucune information sur la base actuelle ne permet de les repérer.

— `situationMaritale` : On pourrait penser qu'en Algérie, les véhicules sont assurés au nom du mari au sein d'un couple même si celui ci est utilisé par les deux membres du couple, d'où la non consistance de la variable sexe. De plus la variable sexe est également biaisée par le fait que le gouvernement ait accordé des crédits ANSEJ à des femmes qui n'utilisent pas le véhicule assuré en leur nom. C'est pour cette raison que la base de données contient anormalement des usages ou genres de véhicules liés à un assuré de sexe féminin. Il serait logique et non sexiste de supposer qu'en Algérie, il est rare de voir une femme au volant d'un semi remorque de plusieurs tonnes.

— `typeSinistre` : une variable qui pourrait être très importante selon moi pour bien modéliser la gravité des sinistres. En effet, dans les bases de données des assureurs étrangers, la séparation des sinistres en véhicules à réparer ou véhicules réformés (le montant du versement à l'assuré est la valeur assurée du véhicule moins quelques frais) est faite pour connaitre la proportion des véhicules à réparer ou totalement détruits. En Algérie, seule une description du sinistre est faite mais chaque dossier en possède une différente. Cette variable pourrait servir à déterminer la proportion des petits sinistres, moyens sinistres et gros sinistres. La définition de la gravité de ces sinistres pourrait faire l'objet d'une étude à elle seule.

— `retraitPermis` : variable difficile à obtenir en Algérie mais pourrait aider à repérer les conducteurs à risque.

— `valeurVenale` : Selon moi, la variable `VA` correspondant à la valeur assurée du véhicule devrait être remplacée par une valeur vénale déterminée selon le marché. En effet, après une exploration des données, la base contient beaucoup d'observations telles que des véhicules identiques possèdent des valeurs assurées très différentes, ce qui est anormal. De plus, comme la prime est calculée en fonction de cette valeur assurée, un assuré pourrait de plein gré demander une police d'assurance avec une très faible valeur assurée afin de payer une petite prime. Il y a bien sûr des conséquences à cela mais en ce qui nous concerne pour l'étude de la gravité des sinistres, la qualité de cette donnée est importante et ne doit pas être biaisée par de tels phénomènes.

### 5.5.3 Préparation des données

Comme dit précédemment, le fichier qui nous a été remis a déjà subi un gros travail de consolidation et de nettoyage. Cependant il reste encore des observations aberrantes ou manquantes [1] qu'il faudrait traiter. Nous allons passer en revue le travail supplémentaire mené sur la base de données.

**Durée d'exposition** – la SAA permet aux assurés de souscrire des contrats Tous Risques pour des durées de 3, 6 ou 12 mois. Seuls les contrats de 6 et 12 mois ont été conservés car les contrats de 3 mois sont peu nombreux, mal saisis et très sinistrés, ce qui pourrait grandement affecter la qualité des modèles. De plus, si on suppose que la fréquence des sinistres de chaque individu est un processus de poisson d'intensité $\lambda$, alors un conducteur exposé au risque pendant 12 mois devrait avoir une sinistralité deux fois plus importante qu'un conducteur exposé au risque pendant 6 mois. Par exemple, si un assuré a subit 2 sinistres durant les 12 derniers mois, on pourrait dire qu'il subi un sinistre tous les 6 mois. Cependant, les données de notre base ne reflètent pas cette logique et nous voyons une fréquence de sinistralité plus élevée sur les contrats de 6 mois si on ramène ces contrats à 12 mois. En effet, les contrats de 6 mois sont en moyenne 9% plus sinistrés que ceux d'un an, ce qui correspond en réalité à une durée d'exposition au risque de 7 mois plutôt que 6. L'interprétation de cette remarque reste à faire mais nous avons pris l'initiative de transformer les contrats de 6 mois en 7 mois et il en sort une meilleure modélisation.

**Variable** `age` – l'étendu de la variable `age` est $[18; 114]$. Il est bien évidemment très rare si ce n'est impossible de rencontrer un individu de plus de 110 conduire un véhicule. Il serait donc préférable d'éliminer tout contrat dont l'âge de l'assuré est supérieur à 85 ans, car 98% des contrats sont souscrits par des assurés de moins de 85 ans.

**Variable** `sinistres` – correspondant au nombre de sinistres reportés pendant la période d'exposition. C'est une valeur entière allant de 0 à 18. Il y a 600 polices d'assurances avec plus de 7 sinistres. il est préférable de les éliminer.

**Variable** `genre` – pour faire une bonne segmentation, il faudrait s'appuyer sur un nombre conséquent d'observations pour que chaque sous population ait assez d'observations pour pouvoir établir une estimation des sinistres avec un bon niveau de certitude. Ainsi, si par exemple un genre de véhicule (tel que les camions d'ordures) possède moins de 100 observations, il serait illogique de conclure sur la sinistralité de ce genre de véhicule. Ainsi tous les genres ayant moins de 100 observations ont été retirés : 5 genres ont été retirés.

**Prise en considération des inflations de tous les montants** – la base de données contient toutes les polices d'assurances établies par la SAA de 2014 à 2018. Durant ces 5 années, l'économie algérienne a fortement évolué et l'inflation a eu un fort impact sur le prix des véhicules et les prix des réparations. L'inflation a fluctué entre 3% et 8%. Pour éviter de complexifier cette tâche d'actualisation, il a été retenu un taux d'inflation annuel moyen de $4,5\%$. Par conséquent, les variables `VA, PleinTarifAct, prime_trAct` et `chargeSin` ont été actualisées avec ce taux et bien sûr en prenant en compte l'année de souscription du contrat.

1. le package `R–mice` a été très utile pour repérer les valeurs manquantes

**La valeur assurée des véhicules** `VA` – cette variable est très liée au montant de remboursement d'un sinistre. Premièrement, le montant de l'indemnisation ne pourrait dépasser la valeur assurée inscrite dans le contrat. Ensuite, cette valeur dépend du véhicule et représente supposément sa valeur vénale : si deux véhicules de marques différentes par exemple sont endommagés de la même manière, les montants des réparations ont de fortes chances d'être différents. Etant donnée l'étendue des valeurs assurées allant de 25 000 DA (absurde bien sûr) à plus de 18 millions de DA en valeur non actualisée, il serait préférable de séparer les grosses valeurs du reste des valeurs assurées sachant que ces dernières représentent 90% de tous les contrats si on fixe un seuil à 3 150 000 DA.

**Variable** `marque` – dans le même esprit que le traitement de la variable `genre`, il faudrait éliminer les marques qui ne possèdent pas beaucoup d'observations. Pour cette variable, les 60 premières marques en termes de nombres ont été retenues pour un minimum de 1000 observations par marque.

**Variables** `inflam` et `Fiscal` – 8995 observations dans toute la base ont le champ `Fiscal` égal à 0, ce qui correspond à une puissance fiscale nulle, ce qui est aberrant et non admissible. Ces observations ont été retirées. 2042 observations sont déclarées comme des véhicules transportant des marchandises inflammables, ce qui est insuffisant pour conclure sur ce type de véhicule car ces 2000 observations font partie de plusieurs genres, marques, région, ...
Pour effectuer l'étude, il reste 735 136 observations à notre disposition, ce qui est suffisant pour avoir des résultats convaincants.

### 5.5.4 Contenu de la base de données et ses variables

La base de données contient les informations de toutes les polices d'assurances souscrites auprès de la SAA dans le cadre de la garantie Tous Risques durant les années 2014 à 2018. Avant nettoyage, il y avait au total quelques 850 000 observations. Le fichier qui nous a été remis en comporte 740 000 car toutes les observations où des informations importantes ont été omises ou mal saisies ont dues être retirées. La table contient la liste des polices d'assurances de deux types de catégories de véhicules : les véhicules particuliers ou les véhicules appartenant à des flottes. Il n'est bien entendu pas possible d'étudier ces deux catégories de la même façon car selon une analyse descriptive, ces deux types de catégories présentent une sinistralité très différente, en plus de ne pas contenir les mêmes variables explicatives ; les véhicules des flottes ne sont pas supposés appartenir à des individus à qui l'on pourrait demander un âge, un sexe, un âge de permis, ...

La base de données comporte 40 variables mais ne sont pas toutes significatives. Il sera question de faire une étude sur la fréquence des sinistres puis une étude sur la gravité des sinistres. Ces deux études portent donc sur la variable `sinistres` qui est le nombre de sinistres survenus durant la période d'exposition d'une police d'assurance, et la seconde variable `ChargeSin` qui est le montant que l'assureur a dû payer en totalité à son assuré pour les sinistres survenus durant la période d'exposition.

### 5.5.5 Modèle individuel

Le modèle individuel se place au niveau de chaque police [5]. Les coûts totaux des sinistres causés par les $n$ polices du portefeuille sont notés $S_1, S_2, \ldots, S_n$. Ces variables sont supposées indépendantes, mais pas identiquement distribuées. Ceci permet de tenir compte de l'hétérogénéité du portefeuille et d'une éventuelle segmentation a priori effectuée par l'assureur. Si on note

$F_i$ la fonction de répartition de $S_i$, c'est-à-dire :

$$F_i(x) = \mathbb{P}(S_i \leq x), \quad x \in \mathbb{R},$$

et que la charge de sinistre $S_i$ causée par la police numéro $i$ est représentée par

$$S_i = \begin{cases} 0, & \text{si la police } i \text{ ne cause aucun sinistre,} \\ Y_i & \text{sinon} \end{cases},$$

avec $Y_i$ représentant le coût total des sinistres relatifs à la police $i$ lorsque cette police a donné lieu à au moins un sinistre. La charge totale de sinistre du portefeuille, notée $S^{ind}$ vaut alors

$$S^{ind} = \sum_{i=1}^{n} S_i.$$

Considérons à présent $q_i$ la probabilité que la police $i$ produise au moins un sinistre sur la période d'exposition et $p_i$ celle qu'elle n'en produise aucun. On a alors dans la réalité $F_i(0) = p_i < 1$.
Si on note $G_i(x) = \mathbb{P}(S_i \leq x | S_i > 0)$ la fonction de répartition de la charge de sinistre de la police $i$ sachant que celle ci a produit au moins un sinistre, on a alors :

$$F_i(x) = p_i \mathbb{1}_{x \geq 0} + q_i G_i(x), \quad x > 0,$$

ce qui donne

$$G_i(x) = \frac{F_i(x) - F_i(0)}{1 - F_i(0)}.$$

L'espérance et la variance de la charge totale du portefeuille dans le modèle individuel s'expriment aisément en fonction des deux premiers moments des montants de sinistres par police :

$$\begin{cases} \mathbb{E}(S^{ind}) = \mathbb{E}(\sum_{i=1}^{n} S_i) = \sum_{i=1}^{n} \mathbb{E}(S_i) \\ Var(S^{ind}) = \sum_{i=1}^{n} Var(S_i) \end{cases}$$

Le montant de sinistre $S_i$ engendré par la police $i$ peut s'exprimer sous la forme

$$S_i = \mathbb{1}_i Y_i$$

avec $\mathbb{1}$, vaut 1 si le contrat $i$ a été touché par au moins un sinistre et 0 sinon.
Les variables aléatoires $\mathbb{1}, , \ldots, Y_1, Y_2, \ldots, Y_n$ sont supposées mutuellement indépendantes. Alors sous cette hypothèse, le montant de la prime pure dans le modèle individuel est donné par :

$$\mathbb{E}(S_i) = \mathbb{E}(\mathbb{1})\mathbb{E}(Y_i).$$

Sous cette même hypothèse et en notant $H^{ind}$ la fonction de répartition de $S^{ind}$, on a alors :

$$H^{ind}(x) = F_1 \star F_2 \star \cdots \star F_n.$$

Malheureusement, le nombre $n$ de polices est en général très grand, ce qui rend impossible le calcul direct de $H^{ind}$ (chacun des produits de convolution nécessitant une intégration numérique). Afin de contourner le problème lié au calcul de $H^{ind}$,les actuaires ont suggéré d'approximer le modèle individuel par un équivalent collectif, dans lequel les calculs sont (sous certaines conditions) plus aisés à effectuer.

### 5.5.6 Modèle collectif

Le modèle collectif de théorie du risque ne distingue plus les polices composant le portefeuille mais voit ce dernier comme un ensemble soumis à une série de chocs causées par l'occurrence des sinistres. Le modèle individuel se place au niveau de la police et distingue la charge des sinistres $S_i$ générée par la police $i$ dans la charge totale $S^{ind} = \sum_{i=1}^{n} S_i$ relative au portefeuille. Contrairement au modèle individuel, le modèle collectif ne distingue plus les polices composant le portefeuille mais voit ce dernier comme un tout, comme un collectif de risques. Les coûts des sinistres touchant le collectif de risques sont modélisés par des variables positives, indépendantes et de même loi. L'identique distribution des coûts dans le modèle collectif s'explique par le fait que l'actuaire renonce à savoir quelle police a causé le sinistre, et gomme donc les différences de sinistralité existant entre les assurés du portefeuille.
Dans la vision collective, $N$ désigne le nombre des sinistres survenus durant une certaine période et $X_i, \quad i = 1, 2, \ldots,$ les montants de ceux-ci. La charge totale des sinistres $S^{coll}$ pour la compagnie s'écrit alors

$$S^{coll} = \sum_{i=1}^{N} X_i,$$

avec la convention que $S^{coll} = 0$ lorsque $N = 0$. Les variables aléatoires $X_i, \quad i = 1, 2, \ldots$ sont supposées indépendantes et identiquement distribuées, et $N$ est supposée indépendante des $X_i$. La loi de $S^{coll}$ est donc composée. Très souvent en assurance, $N$ sera supposé de loi de Poisson, de sorte que $S^{coll} \sim CPoi(\lambda, F)$ avec $F$ représentant la fonction de répartition commune des $X_i$. L'espérance et la variance de la charge totale du portefeuille dans le modèle collectif s'expriment comme ceci :

$$\begin{cases} \mathbb{E}(S^{coll}) = \mathbb{E}(\sum_{i=1}^{N} X_i) = \mathbb{E}(N)\mathbb{E}(X_i) = \mathbb{E}(N)\mathbb{E}(X) \\ Var(S^{coll}) = \mathbb{E}(N)Var(X) + \mathbb{E}^2(X)Var(N) \end{cases}$$

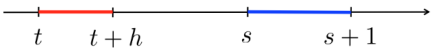
## 5.6 Régression sur variable de comptage

### 5.6.1 Régression de Poisson

On dit qu'une variable aléatoire $Y$ suit une loi de Poisson de paramètre $\lambda$ et on écrit $Y \sim \mathcal{P}(\lambda)$ si sa densité s'écrit :

$$P(Y = y) = e^{-\lambda}\frac{\lambda^y}{y!}.$$

L'espérance et la variance de $Y$ sont alors $\mathbb{E}(N) = Var(N) = \lambda \in \mathbb{R}^+$.
On dit que $(N_t)_{t\geq 0}$ est un processus de Poisson homogène [5, 24, 2] (d'intensité $\lambda$) s'il est à accroissements indépendants, et le nombre de sauts observés pendant la période $[t, t+h]$ suit une loi $\mathcal{P}(\lambda \cdot h)$.




La liste des variables utiles sont les suivantes :

1. `id` : chaîne de caractères servant de clé d'identification de chaque police d'assurance. Elle est la concaténation du numéro d'agence, du numéro de police, du code d'appartenance à l'une des catégories citées plus haut.
2. `dr` : variable catégorielle. La direction régionale auquel appartient l'agence dans laquelle le contrat d'assurance a été souscrit. Cette variable est très importante car elle nous servira de variable spatiale, permettant une segmentation géographique du portefeuille. Cette variable contient 15 modalités mais on ne retiendra que 14. La modalité `annaba` a été retirée car elle ne contient que 700 observations, ce qui ne représente pas la quantité d'observations réelles qui se situe à 50 000. L'extraction des données de la direction régionale de Annaba est incorrecte. Ceci est problématique car beaucoup des régions du Nord-Est du pays dépendent de cette direction régionale.
3. `agence` : variable catégorielle. Code de l'agence dans laquelle la police d'assurance a été souscrite. Pour l'instant, nous n'avons pas utilisé cette variable car c'est une variable catégorielle et elle possède 592 modalités (592 agences SAA en Algérie en 2018) et cela représenterait un énorme coût en temps de calcul. Il serait judicieux d'utiliser cette variable dans une étude future car on pourrait supposer que la plupart des individus souscrivent une assurance dans une agence près de leur domicile et on pourrait donc peut être associer la zone de circulation du véhicule à la localisation de l'agence (segmentation par Daïras-Communes). Une autre idée serait d'utiliser les codes d'agences des contrats d'assurances pour déduire les densités d'habitations de certaines régions car cette variable peut intuitivement et conceptuellement aider à expliquer la sinistralité.
4. `categorie` : variable catégorielle. Cette variable catégorielle possède deux modalités déjà citées précédemment. Elle différencie les contrats des flottes de ceux des véhicules particuliers. Cette variable permettra de créer deux bases de données sur lesquelles nous travaillerons : `basePartic` et `baseFlotte`
5. `annee` : valeur entière. Représente l'année de souscription du contrat. Etant donné l'instabilité de l'économie algérienne, il sera important d'utiliser cette variable pour actualiser les valeurs des montants des sinistres et des valeurs assurées des véhicules. Plus de détails sur ce point par la suite.
6. `duree` : variable catégorielle à deux niveaux. Période d'exposition au risque d'un contrat : 6 mois ou une année.
7. `police` : chaîne de caractères. Numéro de police du contrat.
8. `avenant` : valeur entière. Représente le numéro d'avenant du contrat. Par définition, un avenant est un : « acte par lequel on modifie les termes d'un contrat ». En assurance, plus

précisément un avenant est une révision du contrat pour corriger, apporter des modifications, ajouter ou retirer des garanties ou bien renouveler celui-ci. La variable suivante a été créée à partir de `avenant`

9. `cutAvenant` : variable catégorielle. Variable déduite de la précédente possédant deux modalités : `new` et `old` permettant de différencier les nouvelles polices d'assurances des anciennes. Elle permettra de séparer les nouveaux assurés de l'agence de ceux qui sont à la SAA depuis au moins 1 an, en supposant que la proportion des avenants de modification liés à des rectifications des nouveaux contrats est faible. L'idée derrière la création de cette variable est de savoir s'il existe bien une tendance des nouveaux contrats à être plus sinistrés que ceux des assurés fidèles ou de savoir s'il y a un phénomène de souscription à la garantie Tous Risques dans un seul but : réparer son véhicule. La création de cette variable est intéressante mais mérite une étude à elle seule car la séparation en 2 modalités exclue le cas des contrats RC renouvelés pendant seulement une année pour une Tous Risques.
10. `code` : valeur entière. Code risque. Un contrat d'assurance peut porter sur plusieurs véhicules et le code risque liste ceux-ci, notamment pour les contrats des flottes.
11. `sex` : variable catégorielle. Sexe du souscripteur de la police d'assurance. Variable pas très intéressante comme on le verra plus tard dans l'étude. L'interprétation sera donnée plus tard. Variable indisponible dans `BaseFlotte`.
12. `age` : variable continue. Age du souscripteur du contrat. Variable indisponible dans `BaseFlotte`.
13. `permis` : variable continue. Ancienneté du permis en années. Variable indisponible dans `BaseFlotte`.
14. `nvPermis` : variable catégorielle. Variable permettant de séparer les assurés avec un nouveau permis ou non. Variable indisponible dans `BaseFlotte`. Cetta variable peut être déduite de la précédente en séparant les anciennetés de permis `0` des autres.
15. `marque` : variable catégorielle avec 113 modalités. Elle représente la marque du véhicule.
16. `brand` : variable catégorielle à 56 modalités représentant les marques mais regroupées selon une certaine logique. Par exemple les marques `RENAULT` et `DACIA` sont regroupées en un groupe `RENAULT-DACIA`.
17. `ageVeh` : variable continue. Représente l'âge du véhicule en années.
18. `genre` : variable catégorielle avec 12 modalités. Elle représente le genre de véhicule ; exemples : Transport Publique de Marchandises, Véhicules de plus de 3, 5 tonnes, Tracteurs Routiers, ...
19. `usage` : variable catégorielle avec 10 niveaux. L'usage auquel le véhicule est supposé être assuré ; exemples : affaire, fonctionnaire, location, taxi, ...
20. `Fiscal` : variable continue. Puissance fiscale du véhicule.
21. `zone` : variable catégorielle à deux modalités `nord` et `sud`. Variable permettant de séparer les véhicules sensés être utilisés dans le nord ou dans le sud du pays.
22. `inflam` : variable catégorielle à deux modalités permettant de séparer les véhicules transportant des marchandises inflammables ou non.
23. `tauxRed` : variable numérique. Représente un taux de réduction appliqué sur la prime demandée pendant la souscription.
24. `PleinTarifAct` : variable continue. Représente la prime Tous Risques avant application du taux de réduction. Ce montant est actualisé selon l'année de souscription du contrat.
25. `prime_trAct` : variable continue représentant le montant de la prime exact payé par l'assuré pour la garantie Tous Risques. Réduction appliquée et montant actualisé.

26. `sinistres` : valeur entière. Variable sur laquelle portera une des études dans ce mémoire. Représente le nombre de sinistres reportés durant la période d'exposition d'un contrat.
27. `VA` : variable continue. Représente la valeur assurée du véhicule. Ce montant est demandé par l'assuré pendant la souscription du contrat. Ce montant a été actualisé.
28. `ChargeSin` : variable continue. Deuxième variable sur laquelle portera l'étude. Représente le montant total des déboursements qu'a dû faire l'assureur sur un contrat concernant uniquement la garantie Tous Risques. Valeur actualisée.

**Conclusion** – Après avoir passé en revue toutes les informations contenues dans la base de données, les modifications qui y ont été apportées ainsi que les potentielles variables qui pourraient aider à mieux expliquer la sinistralité, il serait temps maintenant de commencer à traiter le problème lui même en commençant par la présentation de quelques outils théoriques.

# Conclusion générale

L'objectif de ce mémoire était de pouvoir proposer un nouveau système de tarification pour la branche automobile des assureurs algériens. Le système actuel ne prend pas du tout en compte une segmentation du portefeuille et tous les assurés sont donc considérés comme exposés au même risque. D'un point de vue éthique et concurrentiel, ce système n'est pas très adapté et il semble donc judicieux de procéder à une révision de ce système tarifaire.
Dans ce mémoire, nous avons commencé par illustrer l'état actuel de l'activité d'assurance en Algérie et présenté quelques arguments pour justifier le besoin de réforme du système tarifaire actuel. Puis nous avons analysé la base de données de la SAA et donné quelques idées d'informations supplémentaires que l'on pourrait demander aux assurés qui pourraient s'avérer très utiles dans la modélisation de la sinistralité en assurance auto. Puis les mathématiques de la modélisation linéaire généralisée ont été données et suivies de l'étude elle même.
Ainsi, après avoir utilisé plusieurs modèles GLM pour estimer la fréquence des sinistres puis la gravité de ceux-ci, nous avons maintenant une ébauche d'un système de tarification plus juste que celui qui est appliqué aujourd'hui. Les performances des GLM pour faire ce genre d'étude est bien entendu discutable car de nouvelles techniques notamment d'apprentissage automatique ont été développées pour apporter plus de précision dans la prédiction des sinistres. De plus, les résultats de nos modèles sont légèrement inférieurs à ceux retrouvés dans d'autres études ou dans la littérature actuarielle mais ceci est justifiable par la qualité des données et l'hétérogénéité extrême de certaines catégories du portefeuille. Nous n'avons pas choisi d'éliminer les valeurs aberrantes car nous estimons qu'elles font partie de la réalité et les ignorer pour seulement gagner en précision n'est pas correct.
Néanmoins, le travail que nous avons effectué est un bon début pour réviser la tarification actuelle car elle montre qu'il y a d'importantes hétérogénéités qui ne sont pas prises en compte et qu'il y a un manque à gagner à ne pas réviser les systèmes d'aujourd'hui.

# Bibliographie


[1] BOUFIDJELINE, H. (2021). *Bases techniques de l'assurance*. Document de formation interne.

[2] CAMERON, A-C. & TRIVEDI, P. (1998). *Regression Analysis of Counts Data*. Cambridge University Press, New York.

[3] CHARPENTIER, A. *Modèles Linéaires Appliqués*. Cours de modélisation linéaire à l'université du Quebec à Montreal.

[4] CHARPENTIER, A. (2015). *Computational Actuarial Science with R*. CRC Press.

[5] CHARPENTIER, A. & DENUIT, M. (2004). *Mathématiques de l'Assurance Non Vie Tome I et II*. Economica.

[6] CHEIKH, B. *L'histoire de l'assurance en Algérie*. Assurances et gestion des risques. URL : https://www.revueassurances.ca/wp-content/uploads/2016/01/2013_81_no3_4_p285_290.pdf.

[7] CHHIKARA, R-S. & FOLKS, J-L. (1989). *The Inverse Gaussian Distribution : Theory, Methodology and Applications*. New York, NY, USA : Marcel Dekker.

[8] CHOUQUET, C. *Modèles Linéaires*. Cours de modèles linéaires à Université de Toulouse. URL : https://www.math.univ-toulouse.fr/~barthe/M1modlin/poly.pdf.

[9] CNA, Conseil National d'Assurances. *Note conjoncture Trimestre 4 2020*. URL : https://cna.dz/content/download/56874/387374/version/1/file/NC_2020_T4.pdf.

[10] CNA, Conseil National d'Assurances. *Note statistique 2018/2019*. URL : https://cna.dz/content/download/55747/380760/version/1/file/Note+statistique+-+Le+march%C3%A9+de+l%27assurance+automobile+%282018-2019%29.pdf.

[11] CNA, Conseil National d'Assurances. *Patrimoine assurable algérien*. URL : https://www.bdcs.dz/index.php?p=45&g=45.

[12] CNA, Conseil National d'Assurances. *Site web*. URL : https://www.bdcs.dz/index.php.

[13] CNA, Conseil National d'Assurances. (2020). *Assurance retraite*. Revue mensuelle.

[14] CORNILLON, P-A. & HENGARTNER, N. & MATZNER-LOBER, E. & ROUVIÈRE, L. (2019). *Régression avec R*. edp sciences.

[15] CRAMÉR, H. (1946). *Mathematical Methods of Statistics*. Princeton : Princeton University Press.

[16] DUNN, P-K. & SMYTH, G-K. (2018). *Generalized Linear Models with examples in R*. Springer.

[17] FARAWAY, J. (2016). *Extending the Linear Model with R, Generalized Linear, Mixed effects Regression Models*. CRC Press.

[18] FREES, E. (2009). *Regression Modeling with Actuarial and Financial Applications*. Cambridge University Press, New York.

[19] HASTIE, T-J. & TIBSHIRANI, R-J. (1990). *Generalized Additive Models*. Chapman et Hall.

[20] HILBE, J. (2011). *Negative Binomial Regression*. Wiley, New York.

[21] JONG, P. & HELLER, G. (2008). *Generalized Linear Models for Insurance Data*. International Series on Actuarial Science.

[22] KAAS, R. & GOOVAERTS, M. & DHAENE, J. & DENUIT, M. (2008). *Modern Actuarial Risk Theory Using R*. Springer.

[23] LUSSAC, M.L. (2018). *Comparaison de modèles prédictifs pour l'évaluation des coûts matériels automobiles*.

[24] MCCULLAGH, P. & NELDER, J-A. (1989). *Generalized Linear Models (Second edition)*. Chapman et Hall, London.

[25] MCCULLOCH, C-E. & SEARLE, S-R. (2001). *Generalized, Linear, and Mixed Models*. Wiley Series in Probability et Statistics.

[26] MCELREATH, R. (2016). *Statistical Rethinking : A Bayesian Course with Examples in R and Stan*. CRC Press.

[27] NIST/SEMATECH. (2018). *e-Handbook of Statistical Methods*. Engineering Statistics Handbook. URL : https://www.itl.nist.gov/div898/handbook/.

[28] ONS, Office National Statistiques. *Chapitre 3 - Salaires*. URL : https://www.ons.dz/IMG/pdf/CH3-SALAIRES.pdf.

[29] PETRICOUL, J-L. (2017). *Guide pratique de l'assurance*.

[30] RE, Swiss. *Document SIGMA 4/2019*. URL : https://www.swissre.com/dam/jcr:05ba8605-48d3-40b6-bb79-b891cbd11c36/sigma4_2020_en.pdf.

[31] ROSSI, J-R. (2018). *Mathematical Statistics : An Introduction to Likelihood Based Inference*. New York : John Wiley et Sons.

[32] SAA, Société Nationale d'assurance. *Site web*. URL : https://la.saa.dz/fr/about.

[33] SAA, Société Nationale d'assurance. (2020). *Rapport annuel 2019*.

[34] SAA, Société Nationale d'assurance. (2020). *Rapport de gestion 2019*.

[35] SHANNON, C-E. & WEAVER, W. (1963). *The Mathematical Theory of Communication*. University of Illinois Press.

[36] WIKIPÉDIA. *Prime d'assurance*. URL : https://fr.wikipedia.org/wiki/Prime_d%27assurance.

# Résumé

Ce mémoire à pour objet l'étude de la tarification de la garantie Tous Risques en assurance automobile dans le cas du- marché algérien. Le système tarifaire actuel de cette garantie n'est pas équitable car chaque assuré paie une prime proportionnelle à la valeur assurée du véhicule (valeur demandée par l'assuré) sans aucune autre distinction. Ceci est bien évidemment problématique car les assurés ne sont pas tous exposés au même risque. Le travail de ce mémoire a pour objectif de proposer une esquisse pour un système de tarification basé sur une individualisation de chaque prime d'assurance qui devrait être proportionnelle au risque de chaque assuré. Le principal outil mathématique ayant servi à l'élaboration de ces systèmes est le Modèle Linéaire Généralisé (MLG). En statistique, le GLM (Generalized Linear Model en anglais) est une généralisation souple de la régression linéaire. Le GLM généralise la régression linéaire en permettant au modèle linéaire d'être relié à la variable réponse via une fonction lien et en autorisant l'amplitude de la variance de chaque mesure d'être une fonction de sa valeur prévue. Dans ce document, il sera donné plusieurs modèles GLM basés sur des distributions différentes et une comparaison sera effectuée pour déterminer les modèles qui sont les plus performants dans la prédiction de la sinistralité en Algérie.

# Abstract

The purpose of this thesis is to study the pricing of the All Risks guarantee in automobile insurance in the case of the Algerian market. The current pricing system for this guarantee is not fair because each policyholder pays a premium proportional to the insured value of the vehicle (value requested by the policyholder) without any other distinction. This is obviously problematic because policyholders are not all exposed to the same risk. The purpose of this thesis is to provide a sketch for a pricing system based on an individualization of each insurance premium which should be proportional to the risk of each policyholder. The main mathematical tool used in the development of these systems is the Generalized Linear Model (MLG). In statistics, the GLM (Generalized Linear Model) is a flexible generalization of linear regression. GLM generalizes linear regression by allowing the linear model to be related to the response variable via a link function and allowing the magnitude of the variance of each measure to be a function of its predicted value. In this document, several GLM models based on different distributions will be given and a comparison will be made to determine which models are the most efficient in predicting claims in Algeria.