République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université M'hamed Bougara Boumerdès

Département Mathématiques

Mémoire
En vue de l'obtention du Diplôme de MASTER
Spécialité : Mathématiques Financières

Thème

Modélisation du coût des sinistres extrêmes en Assurance automobile

Présenté par : Kouici Manel et Saidoun Manel

Encadré par : M.ZITOUNI Mahieddine

Soutenu le …Septembre 2022 devant le jury composé de :

Président : M. khaldi khaled

Examinateur : Mm.Meddahi samira

Les Listes

Liste des Abréviations

Abréviation	Intitulé
BDG	Brille De Glace.
DC	Dommage Collision.
DASC	Dommage Avec et Sans Contrepartie.
RC	Risque Obligatoire
TR	Tous Risques.
VI	Vol et Incendie.
Min	Minimum
Max	Maximum
1Q	1erQARTILE
3Q	**3 ^{éme}** QARTILE

.

Liste des Figures

Liste des Graphes

Liste des Tableaux

Sommaire

Sommaire

Introduction générale

Introduction générale

Afin de concrétiser les connaissances théoriques acquises pendant le cycle de formation, LA FACULTE DES SCIENCE, DEPARTEMENT MATH propose à ses étudiants la préparation du projet de fin d'études pour l'obtention du diplôme de master en mathématique financière. Les secteurs des assurances en général est parmi les différents secteurs des finances qui ont besoin de l'analyses et modélisations mathématiques et en particulier de l'assurance automobiles (AUTO PARTICULIER) pour que l'on modélise toutes les données liées aux sinistres et que l'on calcule les meilleures propositions afin de satisfaire le client et la société.

Au cours de ce travail, l'étudiant est appelé à appliquer les connaissances théoriques acquises, et il se met face aux problèmes mathématiques existants concernant l'étude et l'analyse des donnée des assurés. Le travail qu'on élaborera consiste à : la modalisation de cout des sinistres extrêmes en assurance automobile.

Nous entamons notre étude dans le premier chapitre "**Généralités sur les assurances**" par Toutes les généralités, les notions de base nécessaires sur l'assurance et l'assurance automobile en particulier assurance auto particulier et finalement une présentation de l'organisme d'accueil la compagnie d'assurance SAA.

Le deuxième chapitre "**Les Outils mathématiques utilisés pour modéliser le coût des sinistres** ", sera dédié essentiellement aux notions de la théorie des valeurs extrêmes.

Le troisième chapitre représente **l'application des outils mathématiques dans l'assurance automobile** (auto particulier). Pour cela nous travaillerons sur trois modèles suivants : le modèle linéaire généralisé, la régression quantile et la théorie des valeurs extrêmes.

La distribution des coûts des sinistres à une queue lourde et de ce fait elle est délicate à étudier. Nous axons notre étude sur la prise en compte des sinistres extrêmes en assurance automobile (AUTO PARTICULIER) en particulier sur la direction régionale de TIZI OUZOU. Les coûts sont modélisés garantie par garantie.

Les assureurs sont soumis à un certain nombre de règles qui sont communes à tous comme nous l'avons évoqué (conventions d'assurance, code des assurances) et qui ont des conséquences directes sur l'indemnisation des sinistres. Mais chaque assureur possède une tarification qui lui est propre et qui est construite en fonction de son portefeuille. Une partie de cette tarification repose sur la modélisation des coûts des sinistres.

Nous allons tester différents modèles prenant en compte la sinistralité extrême.

Dans notre mémoire on s'intéresse à l'assurance AUTO PARTICULIER durant quatre exercices 2017, 2018, 2019 et 2020 (Direction régionale de TIZI OUZOU).

Chapitre 1

Généralités sur les assurances

Introduction

Dans sa vie, l'homme peut être atteint par certains sinistres, il recherche tout naturellement le moyen de supporter la charge du dommage subi ou de la responsabilité encourue.
Le besoin de sécurité est ressenti, au moins par tout individu, exposé aux conséquences de l'adversité il ne peut assumer seul le fardeau, aussi se tourne-t-il, en toute circonstance, vers la collectivité qui prend en charge le dommage résultant pour lui d'une éventualité qu'il redoute. La notion d'assurance est née de cette nécessité et est considérée pour cette raison comme application spéciale de l'instinct d'association.
Pour mieux comprendre le système des assurances il est utile pour ce premier chapitre, de projeter l'attention sur son histoire, ses concepts de base ainsi que ses techniques.

1.1 L'Assurance

1.1.1. Définition de l'assurance

Nous allons tout d'abord définir l'assurance de façon générale, technique, puis juridique.

a. Définition générale :
D'une manière générale, l'assurance peut être définie comme : une réunion de personnes qui, craignant l'arrivée d'un événement dommageable pour elles, se cotisent pour permettre à ceux qui seront frappés par cet événement, de faire face à ses conséquences.[1]

b. Définition juridique
Selon la formulation proposée par le professeur Hérmad :
« 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. »[2]

c. Définition technique
« L'assurance est l'opération par laquelle un assureur, organisant en mutualité une multitude d'assurés exposés à la réalisation de certains risques, indemnise ceux d'entre eux qui Subissent un sinistre grâce à la masse commune des primes collectées. »[3]

Les trois définitions de l'assurance ont l'avantage de faire ressortir les éléments qui caractérisent l'opération d'assurance.

1.1.2 Les acteurs d'une opération d'assurance

Éléments qui découlent d'une opération d'assurance :

a. L'assureur

C'est la société d'assurance ou la personne auprès de laquelle le contrat d'assurance est souscrit, et qui s'engage à fournir les prestations prévues en cas de réalisation du risque.[4]
Techniquement, le rôle de l'assureur est d'évaluer le prix de ce risque à partir des garanties offertes et des paramètres statistiques. Il tiendra compte à la fois du montant moyen des sinistres, de ses valeurs extrêmes, de sa répartition, de sa probabilité de survenance . . .

b. Le souscripteur

Le souscripteur est la personne qui souscrit un contrat d'assurance, c'est à dire qui signe les différents documents du contrat d'assurance (devis ou proposition d'assurance, questionnaire, conditions particulières) et qui s'engage à payer les primes dues à l'assureur.
Le souscripteur n'est pas obligatoirement l'assuré : il peut souscrire un contrat d'assurance pour son propre compte, ou pour celui d'autres personnes indiquées **aux** conditions particulières.[5]

c. L'assuré

Personne physique ou morale sur la tête ou sur les intérêts de qui pèse le risque assuré. L'assuré est la personne à laquelle s'appliquent les garanties du contrat d'assurance.[6]

d. Le contrat d'assurance

Le contrat d'assurance est une convention passée entre une entreprise d'assurance et un souscripteur (individu ou collectivité), fixant à l'avance, pour une période déterminée, des charges financières en fonction d'un ensemble bien défini d'évènements aléatoires.[7]

- ✓ Les conditions générales (risques garantis, exclusions, franchises, démarches pour déclarer un sinistre, paiement des cotisations...)
- ✓ Les conditions particulières (identité de l'assuré et de l'assureur, description du risque, montant de la garantie et de la première cotisation...)

1.1.3 Les éléments clés d'une opération d'assurance

On distingue quatre éléments essentiels d'une opération d'assurance :

a. Le risque

En matière d'assurance le mot « risque » s'emploie pour désigner l'objet de la garantie. Il en est l'élément constitutif, c'est pourquoi il doit être défini avec la plus grande précision possible.[8]

Le risque est l'éventualité de la survenue d'un fait dommageable tel que le vol, la perte, l'incendie, l'accident … sur le bien assuré.

Le risque à un caractère aléatoire puisque, il dépend d'un événement hasardeux provoquant le sinistre.

b. Sinistre

Le sinistre est la réalisation d'un accident entrant dans l'objet du contrat d'assurance. Le sinistre fait naître l'obligation pour une entreprise d'assurance d'exécuter la garantie prévue dans un contrat d'assurance.
La déclaration du sinistre doit être faite par écrit en principe par une lettre recommandée adressée à la société ou à son représentant.
La déclaration comporte le nom, prénom, adresse, numéro du contrat, nom et adresse du courtier, nature, date, heure et lieu de sinistre, circonstances, victime, dommage, témoigne.[9]
Il y a deux sortes de sinistre le sinistre matériel et le sinistre corporel :

- ✓ **Le sinistre matériel** : accident entraînant seulement des dégâts aux victimes adversaire ou bien important à des tiers.
- ✓ **Le sinistre corporel** : accident entraîne des lésions corporelles et des terrasses personne.
- ✓ **Le sinistre mixte** : accident à la Foire des dégâts matériels et corporels dans la réalité.

1.2 L'assurance Automobile

1.2.1 Définition L'assurance Automobile

Le contrat assurance automobile est une assurance obligatoire qui a pour but de garantir le conducteur d'un véhicule automobile contre les conséquences des dommages matériels ou corporels causés par son véhicule à des tiers. En fonction du type de contrat souscrit, l'assurance automobile peut également couvrir les dommages matériels pour le véhicule assuré et les dommages corporels.[10]

1.2.2 Les types d'un contrat d'assurance Automobile

Il existe deux types de contrat d'assurance :

a. Le contrat particulier

Le contrat pour particulier est destiné pour couvrir un seul véhicule, la distinction de particulier est un peu différent du terme employé en général car dans ce contexte le terme particulier ne veut pas dire uniquement l'usage privé mais aussi que le contrat prend en charge un seul véhicule car il peut exister des contrats flotte pour particuliers (un particulier qui possède plusieurs véhicules). Contrairement aux contrats flotte, les contrats pour particuliers n'ont pas d'avantage de réduction, d'absence de franchise ou d'absence de bonus-malus.[11]

b. Contrat flotte

[9] Luc grynbaum, « assurances », éditions L'ARCUS de l'assurance 2011, p211.
[10] <https : https://www.mataf.net/fr/edu/glossaire/assurance-automobile> .Consulté le 20 août 2020.
[11] Djilali B, Cours de droit des assurances Université de, Khemis Miliana.

La flotte est l'ensemble de véhicules à moteur couverts au sein d'une même police automobile, elle est divisée en deux catégories (c.jean.p, 2016) [12]

- ✓ **Les flottes naturelles** : Sont constituées d'un ensemble de véhicules appartenant ou exploités par un même propriétaire ou entité juridique ayant souscrit un contrat collectif pour la couverture globale de son parc. Ainsi tous les véhicules sont soumis aux mêmes règles tarifaires, les primes de chacun d'eux sont recouvrées en une seule fois et les conditions du contrat sont applicables indistinctement à tous.
- ✓ **Les flottes artificielles** : Correspondent au regroupement « mutualisé » de contrats automobiles couvrant des clients distincts d'un prescripteur ayant les mêmes besoins en termes d'assurance, chacun acquittant la prime relative à son véhicule.

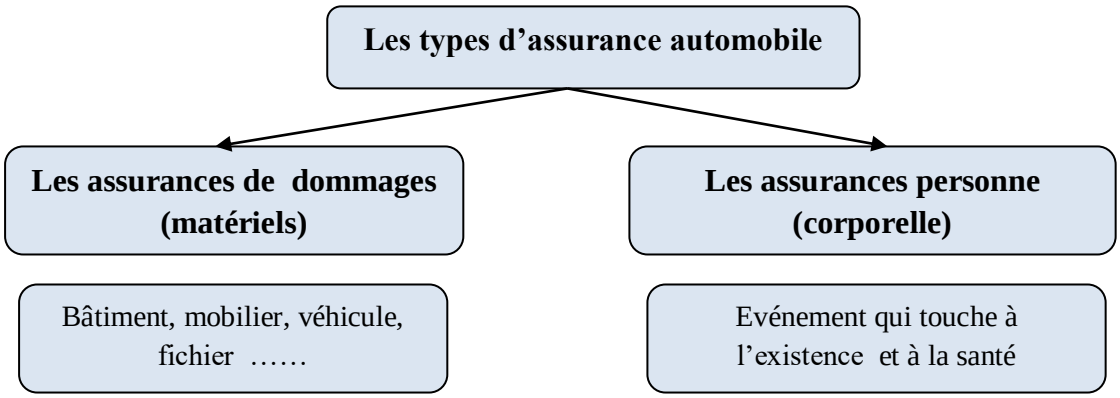


FIGURE 01.1 – Types d'assurance automobile

La prime nette de l'assurance automobile :

Est établie à partir de plusieurs facteurs :
1-Le véhicule lui-même (marque, modèle, année, valeur, âge, coût de réparation, etc.)
2-L'usage que vous en faites (fonctionnaire, auto-école, taxi ;..)
3-Votre lieu de résidence et le lieu d'utilisation du véhicule
4-Le profil du ou des conducteurs (âge, sexe, etc.)
5-Les protections (ou avenants) choisies.
6- La valeur du véhicule (voiture sportive, de luxe ...etc.)

1.3 L'Assurance Automobile en Algérie

1.3.1 L'importance de l'Assurance Automobile dans le marché d'assurance En Algérie [13]

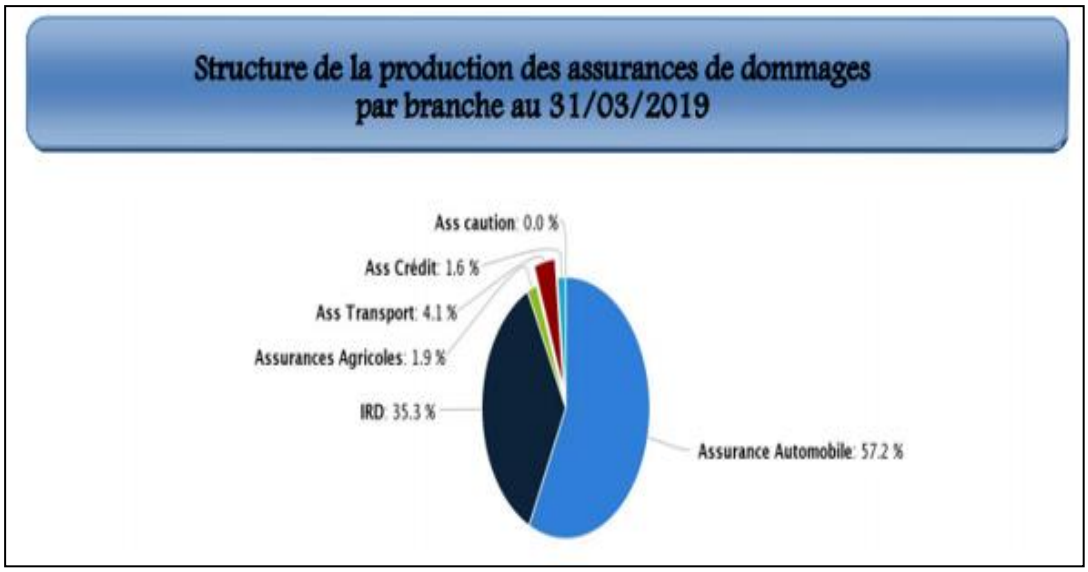


FIGURE 1.2 : La production des assurances de dommages par branche en 2019

Le diagramme circulaire représentant la structure de la production d'assurance des dommages par branche du premier trimestre de 2019 on remarque que le plus grand pourcentage est celui de l'assurance automobile avec 57.2% suivis par IRD avec 35.3%, puis l'assurance transport avec 4.1%, et finalement l'assurance-crédit avec 1.6%.

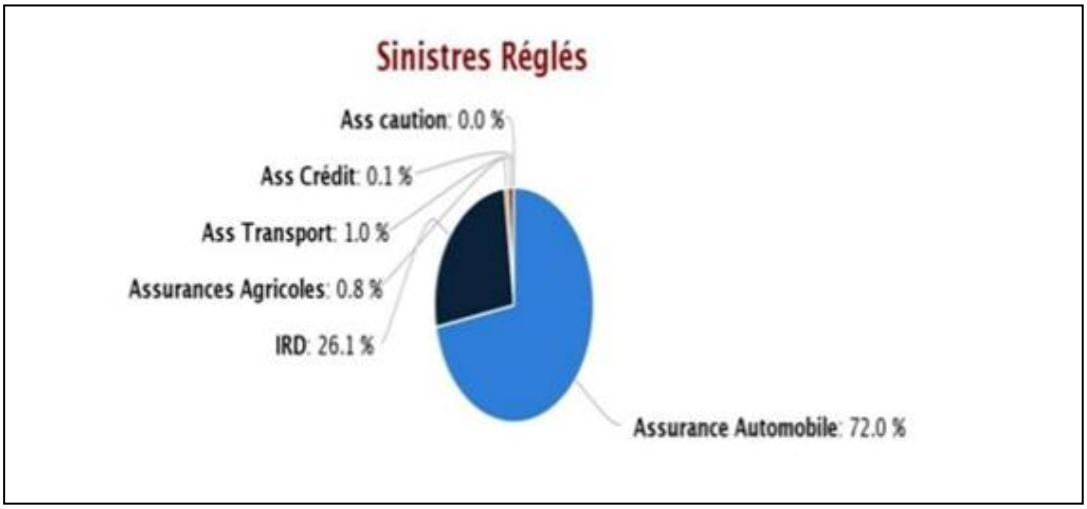


FIGURE 1.3– La sinistralité des assurances par branche en 2019

Le diagramme circulaire représentant la structure des sinistres des assurances dommages majoré par l'assurance automobile avec 72%, les sinistres réglés sont suivis par IRD avec 26.1%, puis l'assurance transport avec 1%, en 4eme place l'assurance agricole de 0.8% et finalement l'assurance- crédit avec 0.1 %.

L'assurance automobile constitue pour le marché algérien la branche principale avec plus 57% de parts du marché. C'est une branche importante de l'assurance étant donné le nombre D'assurés et de sinistres concernés par cette branche.

1.4. Composante du marché d'assurance : La SAA

1.4.1 Présentation de la SAA :

La SAA est une Entreprise Publique Économique par Action relevant du Trésor Public dont le seul actionnaire est l'État, elle dispose d'une riche expérience, de 59 ans dans le domaine des assurances, elle a été Créée le 12 décembre 1963, c'est l'une des plus anciennes compagnies d'assurances à capitaux publics, agréée pour pratiquer toutes branches d'assurances de dommages et de réassurance, travaille avec la banque partenaire la BADR. La SAA dispose également d'une filiale à 100% d'expertise, et actionnaire dans l'assurances de personnes AMANA.

Ainsi Avec un effectif de 3 319 collaborateurs, la SAA propose aux particuliers et aux entreprises, quel que soit leur domaine d'activité, des solutions avantageuses et adaptées à des tarifs étudiés. Par contre au regard de la structure du portefeuille de la SAA, elle est toujours dominée par la branche automobile
.[14]

a. Les branche d'assurance de la SAA

- **AUTO** : Automobile
- **DRPP** : Risque simple
- **DRI** : Risque industriel
- **DAT** : Transport
- **DAA** : Agricole
-

b. Les garanties de la branche AUTO

✓ **Risque obligatoire (RO)**

- **RC :** Responsabilité Civil

✓ **Risque Non Obligatoire (RNO)**

- **TR** : la Tous Risques.
- **DASC** : Dommage Avec ou Sans Contre partie
- **DC** : Dommage Collision
- **VI** : Vole et Incendie
- **BDG** : Bri De Glace

c. Réseau commercial

Le réseau commercial de la SAA est composé de 16 directions régionales (**DR**), qui gèrent 971 points de vente.

- 303 Agences directes
- 228 AGA
- 42 Antennes AD
- 17 Antennes AGA
- 201 Agences bancaires

d. Capacité financière

La SAA possède 30 milliards DA de Capital social, 35 milliards DA Fonds propres et 34,8 Milliard DA Marge de solvabilité.

1.4.2 Les définitions des garanties de la branche AUTO

Les garanties présentes dans un contrat d'assurance auto conditionnent directement l'étendue de la couverture proposée, en complément la garantie obligatoire de responsabilité civile, L'assureur automobile propose d'autres garanties facultatives relatives aux dommages subis par le véhicule ainsi que celles relatives aux personnes transportées à bord.[15]

a. Garantie Responsabilité Civile (RC) : Risque Obligatoire (RO)

La garantie responsabilité civile, l'assurance automobile obligatoire Tout propriétaire d'un Véhicule à moteur doit obligatoirement l'assurer avec, au minimum, une garantie Responsabilité civile (également appelée « assurance au tiers ». Cette garantie couvre Exclusivement les dommages que pourrait causer la voiture assurée à un tiers (passager, autre Conducteur, piéton. . .) ou à un autre véhicule.

b. Les Risques Non Obligatoires (RNO) :

Elle se compose de :

✓ **Garantie Tous Risques ou Dommages avec ou sans contrepartie (DASC):**

- Collision avec un autre véhicule.
- Choc contre un corps fixe ou mobile.
- Renversement sans collision préalable du véhicule assuré.
- Dommages causés par : Hautes eaux, Inondations, Éboulements de rochers, Chutes
- de pierres, Glissement de terrain, Grêle.

✓ **Garantie Dommages-Collision (DC) :**

Cette garantie rembourse les dégâts causés au véhicule en cas de collision avec une autre voiture, un piéton ou un animal. Dans chaque cas de figure, le propriétaire du véhicule concerné, le piéton ou l'animal doit être identifié. Si ce n'est pas le cas, c'est-à-dire que la voiture prend la fuite ou qu'il s'agit d'un animal sauvage, aucun Remboursement n'est alors possible.

✓ **Garantie Vol/Incendie (VI) :**

Au titre de cette garantie, sont couverts les dommages subis par les véhicules assurés Leurs accessoires et les pièces de rechange dont le catalogue du constructeur prévoit la livraison en même temps que le véhicule, en cas :

.

- De vol ou de tentative de vol du véhicule assuré.
- D'incendie, combustion spontanée, chute de la foudre et explosion.

✓ **Garantie Bris de Glace (BDG) :**

Garantit l'assuré contre les dommages causés au pare-brise, lunette arrière et aux glaces Latérales du véhicule assuré, par projection de cailloux, de gravillons ou autre corps.

c. Quelque Autre garanties d'assurance automobile :

✓ **Garantie Personnes Transportées :**

Garantit dans les limites des sommes fixées aux Conditions Particulières, le paiement d'indemnités, en cas d'accident corporel subi par les personnes transportées, dans le véhicule assuré.

✓ **Garantie Défense et Recours :**

- la prise en charge de la défense des intérêts de l'assuré devant les juridictions concernées, chaque fois qu'il est mis en cause du fait de l'utilisation du véhicule assuré.
- l'exercice du recours contre le tiers responsable ou son assureur, pour récupérer le remboursement des dommages subis par le véhicule assuré.

✓ **Garantie Assistance :**

Introduite depuis peu (avril 2007), l'assistance automobile est de plus en plus proposée par les sociétés d'assurance en partenariat avec des sociétés d'assistance. Cette garantie permet, en cas d'accident ou de panne et sur simple appel téléphonique, la mise à la disposition de l'assuré et autres bénéficiaires d'une aide matérielle immédiate :

✓ **Dépannage ou remorquage du véhicule :**

Séjour et déplacement des bénéficiaires à cause de l'immobilisation du véhicule assuré.

Chapitre 2

Outils mathématiques utilisés pour modéliser le coût des sinistres

2.1 Analyse descriptive

Dans le but de comprendre les données mises à notre disposition nous avons décidé d'effectuer des statistiques descriptives sur notre base de données**.**

Voici les définitions suivantes :

Le maximum est la plus grande valeur du caractère effectivement obtenue.
Le minimum est la plus petite valeur du caractère effectivement obtenue.
La médiane : la médiane est un nombre qui divise en 2 parties la population telle que chaque partie contient le même nombre de valeurs. Dans la même logique, il y a les quartiles, déciles et centiles, qui divisent respectivement en 4, 10 et 100 la population.
La moyenne : La moyenne arithmétique est la somme des valeurs de la variable divisée par le nombre d'individus.
La variance : La variance est la moyenne des carrés des écarts à la moyenne.
L'écart-type fi : c'est la racine carrée de la variance.
Quartiles Les quantiles habituellement calculés sont les quartiles :

Q1 : 25% des valeurs sont inférieures au premier quartile
Q3 : 75% des valeurs sont inférieures au troisième quartile

La boite à moustaches permet de visualiser ces différentes valeurs :
c'est un rectangle dont un coté est un trait allant de Q1 à Q3 sur le quelle un trait vertical indique la valeur de la médiane et d'où deux traits horizontaux (les moustaches) débordent : l'un va de la valeur minimum à Q1 et l'autre de Q3 à la valeur maximum. Sur ces deux moustaches, on trouve quelquefois deux traits verticaux indiquant la valeur du premier et du neuvième décile.

Q-Q plot (diagramme Quantile-Quantile) un outil graphique permettant d'évaluer la pertinence de l'ajustement d'une distribution donnée à un modèle théorique.
Le terme de quantile-quantile provient du fait que l'on compare la position de certains quantiles dans la population observée avec leur position dans la population théorique.[1]

QQ-plot des données Un graphique QQ-plot est un outil convenable pour examiner si la distribution d'une variable dans un échantillon provient d'une distribution théorique spécifique. Il donne les quantiles de la distribution empirique en fonction des quantiles de la distribution théorique envisagée. Si l'échantillon provient bien de cette distribution théorique, alors la forme du graphique QQ-plot sera linéaire.

2.2. Le test de Kolmogorov-Smirnov

Le test de Kolmogorov-Smirnov(K-S) porte le nom du mathématicien Nikoláyevich Kolmogorov, est un test d'hypothèse qui compare la distribution observée d'un échantillon statistique à une distribution théorique.

Principe :
Fondé sur les fonctions de répartition sa statistique :

$$D_n = sup_x |F_n(x) - F(x)|$$

Pour un échantillon ordonné $x_1; x_2; \ldots; x_n$, tester l'adéquation $F_n(x)$à une loi $F(x)$revient à mesurer l'écart maximum donné par :

$$D_{KS} = max_{i=1,,,,n}\left\{\left|F(x_i) - \frac{i}{n}\right|, \left|F(x_i) - \frac{i-1}{n}\right|\right\}$$

Hypothèse testée : $\mathcal{H}_0$: La loi de X, a pour fonction de répartition F(x) avec un risque à
La valeur critique du test Kolmogorov est donnée par :

$$ks(\alpha) = \sqrt{\frac{\frac{1}{2}\ln(\frac{\alpha}{2})}{n}}$$

Si D_{KS}- KS (a) alors on accepte $\mathcal{H}_0$, sinon, on la rejette.

2.2 Le modèle linéaire généralisé

Le modèle linéaire généralisé est une extension du modèle linéaire. Il a été introduit par John Nelderet Robert Wedderburn en 1972 et permet d'étudier le lien entre une variable qualitative à expliquer est un ensemble de variables explicatives qualitatives ou quantitatives.
Le modèle linéaire classique est défini par :

$$\begin{cases} Y \approx N(\mu, \sigma) \\ \mu = X\beta \end{cases}$$

Le modèle linéaire généralisé permet notamment de s'éloigner d'hypothèses fortes de ce modèle telles que la normalité de *Y* et la linéarité entre l'espérance et la variable explicative.

2.2.1 Définition du modèle

Le modèle linéaire généralisé est composé de trois éléments :

- **Un ensemble de variables explicatives**

Nous considérons p variables explicatives : $X = \left(\mathrm{X}_1 \cdots \mathrm{X}_p\right)$

ou $X_i \in R^n \forall i \in \{1 \ldots n\}X$ est donc une matrice *n* * *p*

La i-ème ligne de *X* correspond aux valeurs des variables explicatives pour l'observation i.

La j-ème colonne de *X* correspond aux valeurs de la variable explicative j.

- **Une variable à expliquer**

Nous la notons dans la suite de l'étude : $Y \in R^n, Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$ s'agit d'une composante aléatoire qui conditionnellement à la variable X, suit une loi de probabilité appartenant à la famille exponentielle.

Définition 1 : (Famille exponentielle)

La loi de probabilité P appartient à une famille de loi de type exponentielle générale $\{P_\theta\}$ S'il existe une mesure dominante n telle que les lois P_θ ont pour densité par rapport à n :

$$f_\theta(y) = c(\theta)h(y)\exp\left\{\sum_{j=1}^{p} \alpha_j(\theta)T_j(y)\right\} y \in Y$$

Avec $T_1(.) \dots T_p(.) \dots \alpha_1(.), \dots \alpha_p(.)$ fonctions mesurables et Y l'ensemble de définition de la Densité f_θ.

Dans le cadre des modèles linéaires généralisés et par simplification, nous supposerons que la variable à expliquer suit une loi de probabilité de densité par rapport à la mesure dominante n : W Ì R^2 q:est appelé paramètre d'échelle et j paramètre de dispersion.
Nous obtenons donc les résultats suivants (les démonstrations sont en annexe):

$$f_{\theta,\phi}(y) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}$$

Où a(.), b(.) et c(.) sont des fonctions connues et dérivables, b(.) est trois fois dérivable et sa dérivée première est inversible. Les paramètres (q, j) appartiennent à

$$E(Y) = b'(\theta)$$
$$V(Y) = b''(\theta)a(\phi)$$

Exemple : **Cas d'une distribution Gamma**

Soit : $Y \approx \Gamma(\alpha\ ,\ \beta\ )$

Y admet donc pour densité : $f_{\alpha,\beta}(y) = \frac{1}{\alpha^{\beta} * \Gamma(\beta)} y^{\beta-1} e^{-y/\alpha}$

que nous pouvons également écrire sous la forme :

$$f_{\alpha,\beta}(y) = [\exp(\beta - 1)\, y - \frac{y}{\alpha} - \beta \ln \varepsilon - \ln(\Gamma(\beta))]$$

Nous retrouvons alors les éléments de la famille exponentielle :

$$\phi = \frac{1}{\beta}$$
$$\theta = -\frac{1}{\alpha\beta}$$
$$b(\theta) = -\ln(-\theta)$$
$$c(y, \phi) = -\ln(\Gamma(\beta)) + \beta \ln(y\beta) - \ln y$$

- ·**Une fonction de lien**

Contrairement au modèle linéaire traditionnel, nous ne modélisons non pas l'espérance de Y directement mais une transformation de cette valeur. Ainsi, la fonction de lien notée g est la liaison entre la composante aléatoire Y et une combinaison linéaire des variables explicatives.
G est une fonction bijective et différentiable.

$$\eta = g(E(Y)) = \sum_{f=1}^{p} X_j \beta_j$$

La fonction de lien canonique définie par $g(.) = (b')^{-1}$, simplifie les calculs théoriques puisque dans ce cas : $g(E(Y)) = ((b')^{-1}\ (b'(q))) = q$.

Voici les fonctions de lien canoniques des lois exponentielles les plus courantes

Loi	Nom du lien	Fonction de lien
Poisson	Log	$g(\mu) = \log(\mu)$
Normale	Identité	$g\ (\mu\ ) = \mu$
Gamma	Réciproque	$g(\mu) = -\frac{1}{\mu}$

Tableau 2.1 : Récapitulatif des fonctions de lien canoniques des modèles usuels

Loi Binomiale négative :

Une variable pouvant prendre seulement des valeurs entières et positives (0,1,2,3, 4…) et dont la variance est particulièrement forte.

Loi de poisson :

Une variable pouvant prendre seulement des valeurs entières et positives (0,1,2,3, 4…)

- La distribution résiduelle d'un modèle le nombre de fées moyen dépendrait de l'enchantement **n'est pas gaussienne**… Les fées semblent en outre avoir tendance à se rassembler au sein d'un même arbre (soit par tendance grégaire, soit parce que les arbres en question répondent à des critères de choix qui, peut-être, nous échappent du fait de notre condition de simples mortels).

  Et une transformation, type transformation log, n'y changerait rien ! En effet, on a ici affaire à un type de distribution typique des données de comptage.

  Classiquement, le modèle de distribution pour de telles données de comptage peut être soit

- ✓ Une **distribution de Poisson** (du nom d'un mathématicien français) si la variance est à peu près égale à la moyenne,
- ✓ Soit une **distribution binomiale négative** quand la variance est plus forte que la moyenne.

En pratique, les calculs sont menés par des logiciels, simplifier les calculs n'est donc pas une priorité et d'autres fonction de lien peuvent être privilégiées en fonction des données.
De plus, h peut prendre n'importe quelle valeur de *R* mais dans le cas de certaines distributions exponentielles les valeurs de m peuvent être restreintes. Le choix de la fonction de lien canonique est alors discutable.

2.2.2. Estimations des coefficients

Le paramètre de dispersion j est supposé connu, si ce n'est pas le cas le paramètre est estimé au préalable et est ensuite supposé connu. Nous avons à notre disposition n observations des variables X et Y.
La fonction de lien du modèle et la densité de la variable aléatoire Y sont connues puisqu'elles sont choisies par l'utilisateur. Il reste donc à estimer q. Pour cela nous utilisons la méthode du maximum de vraisemblance.

- **Calcul de la log-vraisemblance**

Pour une observation i, la log-vraisemblance est :

$$l(y_i, \beta, \phi) = \log\left(f_{\theta,\phi}(y_i)\right) = \frac{y_i\theta - b(\theta)}{a(\phi)} + c(y_i, \phi) = \frac{y_i\theta - b(\theta)}{\phi} + c(y_i, \phi)$$

Ici, nous supposons que le poids de l'observation i est constant et donc que la fonction *a* est égale à la fonction identité.

La méthode du maximum de vraisemblance nous permet d'identifier le paramètre $\hat{\theta}$ qui maximise lalog-vraisemblance.

Etant donné le lien entre b et q : m = g^{-1} (h) = g^{-1} (*x*'b) et m = *b*' (q) , une condition nécessaire d'optimum est d'annuler la dérivé partielle de la log-vraisemblance par rapport au paramètre b .

Notations :

$$\mu_i = E(Y \mid X = X_i)$$
$$\eta_i = X_i'\beta$$

La valeur du paramètre q au point $X = X_i$

Le lien entre la log-vraisemblance et le paramètre b n'étant pas direct, il est plus pratique de considérer la décomposition suivante :

$$\frac{\partial l}{\partial \beta_j} = \frac{\partial l}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j}$$

1. Nous savons que $\frac{\partial l}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{\phi}$ d'après l'expression de *l*.

2. D'après l'expression de la variance :

$$V(Y \mid X = X_i) = b''(\theta)\phi \text{ or } E(Y) = \mu = b'(\theta)$$

Nous avons donc la relation :$V(Y \mid X = X_i) = \frac{\partial \mu_i}{\partial \theta_i}\phi$

Et donc : $\frac{\partial \theta_i}{\partial \mu_i} = \frac{\phi}{V(Y|X=X_i)}$

3.$\frac{\partial \mu_i}{\partial \eta_i} = \frac{\partial g^{-1}(\eta_i)}{\partial \eta_i} = (g^{-1})'(\eta_i) = h'(\eta_i)$avec h l'inverse de la fonction de lien.

4. $\frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial\left(\sum_{k=1}^{p} X_{ik}\beta_k\right)}{\partial \beta_j} = X_{ij}$

-Nous obtenons donc :

$$\frac{\partial l}{\partial \beta_j} = \frac{y_i - b'(\theta_i)}{\phi} * \frac{\phi}{V(Y \mid X = X_i)} * h'(\eta_i) * X_{ij}$$

Finalement :

$$\frac{\partial l}{\partial \beta} = \left(y_i - b'(\theta_i)\right) * \frac{h'(\eta_i)}{V(Y \mid X = X_i)} * X_i = (y_i - \mu_i) * \frac{h'(\eta_i)}{V(Y \mid X = X_i)} * X_i$$

· **Problème**

Nous ne connaissons pas μ_i qui dépend de β et nous ne pouvons pas trouver d'expression simple de β en annulant cette dérivée.

Il est donc nécessaire d'utiliser un algorithme pour maximiser la vraisemblance.

· **Processus itératif**

Principe de l'algorithme :

1. Maximiser la log-vraisemblance *L* revient à minimiser *–L*

2. Nous initialisons l'algorithme par b_0 En pratique, le logiciel utilisé (SAS) ne nous demande pas de choisir de valeur initiale, il le fait automatiquement.

3. Construction de b_{k+1} à partir de b *k* tel que $-L(\beta_{k+1}) \leq -L(\beta_k)$

Pour cela nous utilisons la construction suivante : $\beta_{k+1} = \beta_k + A_k \nabla L_k$

Où A_k est la matrice de pas de l'algorithme et et∇L_k le gradient calculé au point b_k.

4. Arrêt de l'algorithme lorsque $\beta_{k+1} \approx \beta_k$ ou $L(\beta_{k+1}) \approx L(\beta_k)$

La procédure GENMOD de SAS utilise l'algorithme de Newton-Raphson pour calculer b D'après cette méthode : $A_k = -\left(E(\nabla^2 L_k)\right)^{-1}$

Nous allons calculer le hessien $\nabla^2 L_k = \begin{pmatrix} \vdots & \\ \cdots & \frac{\partial^2 L}{\partial \beta_j \beta_k} \end{pmatrix}$:

$L = \sum_{i=1}^{n} l(y_i, \beta, \phi)$

$\frac{\partial l}{\partial \beta_j} = \frac{(y_i - \mu_i)}{\phi} * \frac{w_i}{h'(\eta_i)} * X_{ij}$ Avec la notation $W_i = \frac{\left(h'(\eta_i)\right)^2}{b''(\theta_i)}$ qui est le poids associé à

L'observation i.

$$\frac{\partial l}{\partial \beta_j \beta_j} = \frac{\partial}{\partial \beta_k}\left(\frac{(y_i - \mu_i)}{\phi} * \frac{W_i}{h^{'}(\eta_i)} * X_{ij}\right)$$
$$= \left(\frac{\partial}{\partial \beta_k}\left(\frac{y_i - \mu_i}{\phi}\right)\right) * \frac{W_i}{h^{'}(\eta_i)} * X_{ij} + \frac{y_i - \mu_i}{\phi}\frac{\partial}{\partial \beta_k}\left(\frac{W_i}{h^{'}(\eta_i)}\right) * X_{ij}$$

Le terme de droite ayant une espérance nulle car $E(y_i - \mu_i) = 0$nous le notons K.

Nous utilisons la décomposition suivante pour calculer le terme de gauche :

$$\frac{\partial \mu_i}{\partial \beta_k} = \frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_k} = h^{'}(\eta_i)^{*} X_{ik}$$

-Nous obtenons donc :

$$\frac{\partial^2 l}{\partial \beta_j \beta_k} = \frac{1}{\phi}\left(-h^{'}(\eta_i) X_{ik} W_i \frac{1}{h^{'}(\eta_i)} X_{ij}\right) + K$$
$$= -\frac{1}{\phi} X_{ik} W_i X_{ij} + K$$

Nous passons à l'espérance, ce qui revient à calculer la matrice d'information de **Fisher:**

$$-E\left(\frac{\partial^2 l}{\partial \beta_j \beta_k}\right) = \frac{1}{\phi} X_{ik} W_i X_{ij}$$

Nous avons donc :

$$-E(\nabla^2 L) = -E\left(\sum_{i=1}^{n} \frac{\partial^2 l}{\partial \beta^2}\right) = \sum_{i=1}^{n}\left(\cdots \frac{1}{\phi} X_{ik} W_i X_{ij}\right) = \frac{1}{\phi} X^{'} W X$$

Et donc : $-\left(E(\nabla^2 L)\right)^{-1} = \phi (X^{'} W^{k} X)^{-1}$

Remarque : Il ne faut pas oublier que *W* dépend de β_k

Enfin : $\beta_{k+1} = \beta_k + \phi (X^{'} W X)^{-1} \nabla L_k$

Or comme calculé précédemment : $\frac{\partial l}{\partial \beta_j} = \frac{(y_i - \mu_i)}{\phi} * \frac{W_i}{h'(\eta_i)} * X_{ij}$ donc

$$\nabla L_k = \frac{1}{\phi}(Y - \mu)X' \mathrm{W}\ \mathrm{diag}\left(\frac{1}{h'(\eta_i)}\right)$$

Finalement :

$$\beta_{k+1} = \beta_k + \left(X'WX\right)^{-1}(Y - jj)X'W^k\ \mathrm{diag}\left(\frac{1}{h'(\eta_i)}\right)$$

2.2.3. Intervalles de confiance

► Propriété 1 : Efficacité de l'estimateur du maximum de vraisemblance

Dans un modèle régulier, l'estimateur du maximum de vraisemblance est asymptotiquement efficace.

Soit $\hat{\alpha}$ l'estimateur du maximum de vraisemblance du paramètre a.

$$\sqrt{n}(\hat{\alpha} - \alpha) \xrightarrow{loi} N(0, \Sigma)$$

où Σ est l'inverse de la matrice d'information de Fisher calculée au point a.

Remarque :

Les hypothèses d'un modèle régulier sont rappelées en annexe.

Le modèle exponentiel est un modèle régulier, cette propriété s'applique donc à $\hat{\beta}$

$$\sqrt{n}\left(\hat{\beta} - \beta\right) \xrightarrow{loi} N(0, \phi(X'WX)^{-1})$$

W est la matrice de poids calculée au point b qui est inconnu, en pratique nous la calculons donc au Point$\hat{\beta}$.

Nous pouvons donc construire un intervalle de confiance de niveau a du paramètre $\hat{\beta}_j$

$$IC(\beta_j) = \left[\hat{\beta}_j - u_{1-\frac{\alpha}{2}}\frac{1}{\sqrt{n}}\hat{\sigma}_{\hat{\beta}_j}; \hat{\beta}_j + u_{1-\frac{\alpha}{2}}\frac{1}{\sqrt{n}}\hat{\sigma}_{\hat{\beta}_j}\right]$$

Considérons une suite de vecteurs aléatoires $(Z_n) \in R^k, z \in R^k$ et (a_n) une suite de réels strictement positifs.

Supposons : $(Z_n) \in R^k, z \in R^k$et (a_n)

Soit g une fonction définie sur un voisinage V de z et différentiable en z . Alors :

$$a_n\big(g(Z_n) - g(z)\big) \to g'(z)Z$$

$u_{1-\frac{\alpha}{2}}$ désigne le quantile de niveau $1 - \frac{\alpha}{2}$de la loi normale centrée.

$$\hat{\sigma}_{\hat{\beta}_j} = \frac{1}{n}[I(\hat{\beta})]^{-1}_{jj}$$

La prévision par le modèle linéaire généralisé au point $X = x_i$ est la moyenne $\hat{\mu}_i$.

En utilisant **la delta-méthode**, nous pouvons obtenir un intervalle de confiance de cette estimation à partir de celui construit pour le paramètre b*j*.

➢ Propriété 2 : Delta-méthode

Pour obtenir l'intervalle de confiance des estimations des moyennes conditionnelles $\hat{\mu}_i = g^{-1}(x_i'\hat{\beta})$ il faut appliquer deux fois la delta-méthode.

En l'appliquant une première fois, nous obtenons un intervalle de confiance de niveau a pour $x_i'\widehat{\beta}$:

$$IC(x_i'\beta) = \left[x_i'\hat{\beta} - u_{1-\frac{\alpha}{2}}\frac{1}{\sqrt{n}}\hat{\sigma}_{\hat{\eta}_i}; x_i'\hat{\beta} + u_{1-\frac{\alpha}{2}}\frac{1}{\sqrt{n}}\hat{\sigma}_{\eta_i}\right]$$

Avec $\hat{\sigma}_{\hat{\eta}_i} = \sqrt{\phi x_i'(X'WX)^{-1}x_i}$

Nous appliquons une seconde fois la delta-méthode pour $g^{-1}(x_i'\hat{\beta})$ et nous obtenons :

$$IC(\mu_i) = \left[x_i'\hat{\beta} - u_{1-\frac{\alpha}{2}}\frac{1}{\sqrt{n}}(g^{-1})(x_i'\hat{\beta})\hat{\sigma}_{\hat{\eta}_t}; x_i'\hat{\beta} + u_{1-\frac{\alpha}{2}}\frac{1}{\sqrt{n}}(g^{-1})(x_i'\hat{\beta})\hat{\sigma}_{\hat{\eta}_i}\right]$$

2.2.4. Comparaison des modèles

La difficulté, dans l'utilisation des modèles linéaires généralisés est de choisir le modèle le plus adéquat. Pour cela, il existe des critères permettant de comparer les modèles entre eux.

- **La déviance :**

La méthode du **Fisher scoring** nous a permis de calculer $\hat{\beta}$

Notons $\mu_i = E(Y \mid X = X_i)$

Nous avons alors : $\widehat{\boldsymbol{\mu}}_{\boldsymbol{i}} = \boldsymbol{g}^{-1}(\boldsymbol{x}_{\boldsymbol{i}}'\widehat{\boldsymbol{\beta}})$ Si le modèle construit était parfait, $\hat{\mu}_i$serait égal à la moyennedes observations de Y lorsque $X = X_i$. Ce modèle considéré comme parfait est appelé modèle saturé.

L'algorithme de **Newton-Raphson** nous permet de déterminer $\hat{\beta}$, ce qui permet également dedéterminer $\hat{\theta}$à partir de la relation $\theta = (b')^{-1}\left(g^{-1}(x_i'\beta)\right)$ en remplaçant b par $\hat{\beta}$.

Nous pouvons donc calculer la log-vraisemblance maximisée :

$$l(y_i, \hat{\beta}, \phi) = \frac{y_i\hat{\theta} - b(\hat{\theta})}{\phi} + c(y_i, \phi)$$
$$= \frac{y_i^*(b')^{-1}(\hat{\mu}_i) - b_0(b')^{-1}(\hat{\mu}_i)}{\phi} + c(y_i, \phi)$$

Pour le modèle saturé :

$$l_{\text{ saturé }}(y_i, \hat{\beta}, \phi) = \frac{y_i^*(b')^{-1}(y_i) - b_o(b')^{-1}(y_i)}{\phi} + c(y_i, \phi)$$

La déviance est définie par :

$$D = 2\phi\left[\sum_{i=1}^{n}\left(l_{\text{ saturé }}(y_i) - l(y_i, \hat{\beta}, \phi)\right)\right]$$

Et la déviance standardisée par :

$$D_{\text{scaled}} = 2\left[\sum_{i=1}^{n}\left(l_{\text{saturé}}(y_i) - l(y_i, \hat{\beta}, \phi)\right)\right]$$

Plus la déviance ou la déviance standardisée sont grandes, moins le modèle considéré est proche du modèle saturé et donc moins l'estimation est de bonne qualité.

Exemple : Cas d'une distribution Gamma

Soit $Y \approx \Gamma(\alpha, \beta)$

Nous avons alors : $D = \frac{1}{\phi}\sum_{i=1}^{n}\left(-\ln\left(\frac{y_i}{\hat{\mu}_i}\right) + \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i}\right)$

- **L'AIC: Akaike Informative Criterion**

AIC = -2L + 2 p

L :log vraisemblance du modèle

P: nombre de paramètres

- **Le BIC:Bayesian Informative Criterion**

BIC = -2L + p log(n)

n : nombre d'observations

L'utilisation de ces critères est relativement simple puisque le modèle sélectionné est le modèle minimisant le critère considéré.

2.2.5. Utilisation

Le modèle linéaire généralisé est un outil courant en assurance dommage et de nombreux logiciels ont développé des outils facilitant son utilisation R

En pratique, cinq étapes sont essentielles à sa mise en place :

_ le choix de la loi de la distribution

_ le choix de la fonction de lien

_ Le choix des variables explicatives

_ L'estimation de b

_ Le calcul des valeurs estimées

Dans certaines situations, les choix de la loi de la variable d'intérêt ou des variables explicatives peuvent s'avérer délicat. Nous verrons comment mettre en pratique cette théorie dans le chapitre 3.

2.3 La régression quantile :

La méthode de régression quantile a été introduite par Koenker et Basset en 1978. Elle permet d'estimer les quantiles de la variable d'intérêt conditionnellement à un ensemble de variables explicatives.
La régression quantile est une extension de la méthode des moindres carrés qui permet d'estimer la fonction moyenne de la distribution conditionnelle de la variable à expliquer. L'avantage de la régression quantile est qu'elle caractérise des points particuliers de la distribution de la variable, elle peut donc apporter plus particulièrement des informations sur les queues de distribution. La régression quantile est un outil adapté à l'étude des valeurs extrêmes puisque contrairement à la méthode des moindres carrées, cette méthode est robuste : elle est donc moins sensible aux variations sur les données.
Cette méthode sera utilisée dans nos modèles pour déterminer des seuils à partir desquels les sinistres seront considérés comme extrêmes. D'après les résultats de la régression quantile, nous construirons deux échantillons : l'un contenant les sinistres bas et l'autre les sinistres considérés comme extrêmes.

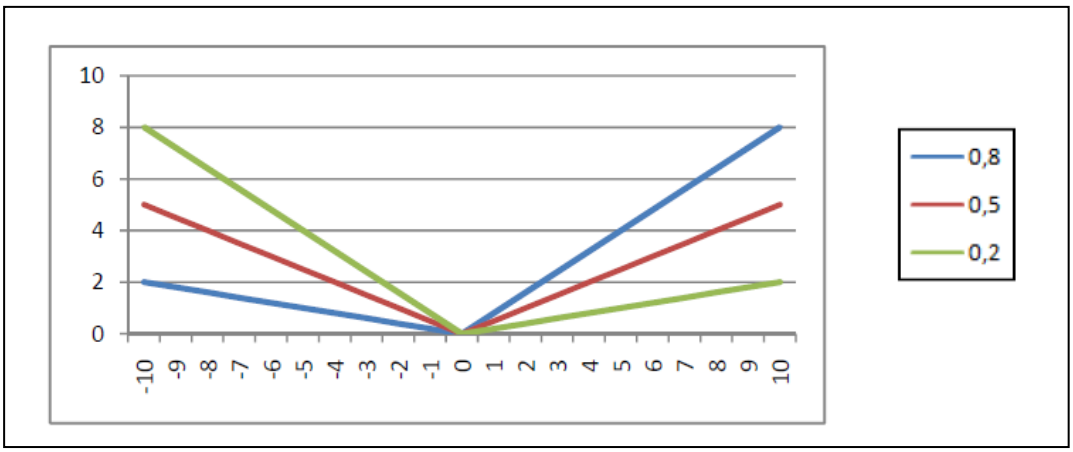
2.3.1 Définition du modèle

➢ Définition 1 : (θ -quantile)

Soit Y une variable aléatoire réelle de fonction de répartition Fy et θ un réel compris entre 0 et 1.
Le θ -quantile de F_Y , noté $q_Y(\theta)$ est la solution de l'équation : $F_Y(q) = \theta$ c'est-à-dire,

$$\mathbf{q_Y(\theta) = \inf\{y : F_Y(y) \geq \theta\}}$$

Remarque : La fonction quantile est l'inverse généralisé de la fonction de répartition.
Nous introduisons une nouvelle fonction : ρ_θ est une fonction de pondération appelée « check function » en anglais et définie par : $\rho_\theta(u) = u(\theta - 1_{u<0})$
Voici une représentation de cette fonction pour différentes valeurs de θ:



Graphe 2.1– Check fonction pour différentes valeurs de q.

Le θ -quantile peut aussi être vu, notamment lors de l'utilisation de la méthode des moindres carrés ordinaires, comme la quantité minimisant par rapport à q la fonction objective suivante :

$$E(\rho_\theta(y-q))= \theta\int_{y>q} |y-q|\, dF_Y(y) + (1-\theta)\int_{y<q} |y-q|\, dF_Y$$

$$=\int_{y>q}(y-q)\, dF_Y(y) - (1-\theta)\int_{y<q}(y-q)\, dF_Y$$

Nous différentions ensuite par rapport à q :

$$0 = \theta\int_{q}^{+\infty} dF_Y(y) - (1-\theta)\int_{-\infty}^{q}(y-q)\, dF_Y(y)$$

$$=F(q)-\theta$$

Deux cas peuvent alors se présenter :

_ Cette équation présente une solution unique alors $q = F^{-1}(\theta)$

_ Nous obtenons un intervalle de valeurs possibles pour le θ -quantile. Dans ce cas, par convention et pour que la fonction quantile soit continue à gauche, le θ -quantile est fixé à la plus petite valeur de l'intervalle de solutions.

Lorsque F est remplacée par sa fonction empirique : $F_n(y)=\sum_{i=1}^{n} 1_{\{Y_i \le y\}}$ la quantité à minimiser pour obtenir la valeur du θ -quantile est :

$$\int \rho_\theta(y-\theta)\, dF_n(y)=\frac{1}{n}\sum_{i=1}^{n} \rho_\theta(y_i-q)$$

Ceci est le point clé de la régression quantile introduite par Kroenker : le calcul des quantiles ne se ramène plus au tri d'un échantillon mais à un problème d'optimisation.

➢ Définition 3: (θ -quantile conditionnel)

Considérons la distribution conditionnelle de *Y* caractérisée par la fonction de répartition $F_{y|x}$, nous définissons de manière similaire les quantiles conditionnels par :

$$q_{Y|X}(\theta)=\inf\{y : F_{y|x}(y)\geq \theta\}$$

De même que pour l'estimation des quantiles, nous nous ramenons à un problème d'optimisation pour estimer les quantiles conditionnels. La fonction objective à minimiser par rapport à β est la suivante :

$$\mathbf{V_n(\beta,\theta)=\frac{1}{n}\sum_{i=1}^{n}\rho_\theta(y_i - x_i'\beta).}$$

Remarque 1 : Pour θ = 0.5, on se ramène à la méthode LAD (Least Absolute Deviation)

Remarque 2 : Il s'agit de la même généralisation que pour le calcul de la moyenne. La méthode des moindres carrées utilise comme fonction objectif à minimiser : $\sum_{i=1}^{n}(y_i-\mu)^2$. cette fonction est généralisée à $\sum_{i=1}^{n}(x_i^T-\beta)^2$. pour calculer la moyenne conditionnelle.

La fonction objectif n'étant pas dérivable au point $y_i = x_i'\beta$, il n'est pas possible d'obtenir directement une estimation des paramètres. Il faut donc avoir recours à des algorithmes. Pour cela, il est plus pratique de réécrire le problème d'optimisation sous une autre forme équivalente. Etudions le modèle suivant :

$$y_i = x_i'\beta + e_i = \sum_{j=1}^{k} x_{i,j}(\beta_j^+ - \beta_j^-) + (e_i^+ - e_i^-)$$

Où β_j^+ est la partie positive de β_j c'est-à-dire $\beta_j^+=\max(0 ; \beta_j)$.

β_j^- est la partie négative de β_j c'est-à-dire $\beta_j^-= -\min(0 ; \beta_j)$.

Donc $\beta_j^+-\beta_j^- = \beta_j$ et de même $e_i^+ - e_i^- = e_i$

Soit z le vecteur de dimension $2*(k+n)$ défini par : $Z=[\beta^{+\prime}, \beta^{-\prime}, e^{+\prime}, e^{-\prime}]$

Nous pouvons écrire le modèle sous forme matricielle : $y = Az$ avec A la matrice $n*2(k+n)$ définie

Par $A = [X , -X , I_n , -I_n]$

Posons $c = [0',0',\theta\,1', (1-\theta)1']$ où 0 et **1** sont des vecteurs de taille k, respectivement le vecteur nul et le vecteur contenant des 1.

Alors la régression quantile se présente sous le problème d'optimisation suivant :

$$\begin{cases}\min_Z \; c'Z \\ y = AZ\end{cases}$$

De nombreuses méthodes ont été développées pour résoudre ce problème linéaire. Pour en citer quelques-unes : l'algorithme du simplex étendue à la régression quantile en 1987 par Koenker et d'Orey, la méthode du point intérieur adaptée par Koenker et Park en 1996.

2.3.2 Propriétés :

➢ Propriété 3 : Changement d'échelle (Koenker et Basset 1978)

Ces propriétés algébriques peuvent faciliter en pratique le calcul des quantiles conditionnels.
Soit $a \geq 0$ et $\gamma \in \Re^P$

(i) $q(\theta , ay, X) = aq(\theta , y, X)$
(ii) $q(\theta ,- ay, X) = - aq(1-\theta , y, X)$
(iii) $q(\theta , y + X\gamma , X) = q(\theta , y, X)+\gamma$

Preuve :

(ii) $q_{aY|x}(\theta)=\inf \{y,P(-aY < y|X)\geq \theta \}$

$=\inf \{-ay, P(Y>y|X)\geq \theta \}$

$= - a \inf \{ y, 1-P(Y>y|X)\geq 1- \theta \}$

$= - a \inf \{ y, P(Y < y|X)\geq 1- \theta \}$

$= - a * q_{Y|x}(1-\theta)$

➢ Propriété 4 : Invariance

La fonction quantile est invariante par transformation monotone. Soit *f* une fonction monotone.
Alors : $Q_{f(Y|X)}(x_\tau)=f(Q_{Y|X}(x_\tau))$

Preuve :

Cela découle directement de la propriété : $P(Y < y| x)= P(f(Y) < f(y)| x)$

➢ Propriété 5 : Robustesse

Un des avantages de la régression quantile est la robustesse de cette méthode c'est une des raisons qui justifient son utilisation pour traiter des valeurs extrêmes.

2.3.3 Adéquation au modèle

De façon similaire à la régression classique des moindres carrées ordinaires, nous pouvons définir une mesure de la contribution relative de l'ajout d'un régresseur. Ce critère a été développé par Koenker et Machado en 1999.Considérons les modèles suivants :

-le modèle complet : $Q_{Y|X}(\theta)=X_1\beta_1(\theta)+X_2\beta_2(\theta)/$

-le modèle restreint : $Q_{Y|X}(\theta)=X_1\beta_1(\theta)$

Nous pouvons définir un critère d'adéquation du modèle complet relatif au modèle restreint noté R^1

$$R^1(\boldsymbol{\theta}) = \mathbf{1} - \frac{\mathbf{V_n}(\hat{\boldsymbol{\beta}}_1(\boldsymbol{\theta});\hat{\boldsymbol{\beta}}_2(\boldsymbol{\theta});\boldsymbol{\theta})}{\mathbf{V_n}(\tilde{\boldsymbol{\beta}}_1(\boldsymbol{\theta});\mathbf{0};\boldsymbol{\theta})}$$

Où $V_n(\hat{\beta}_1(\theta);0;\theta)$ est calculé sous la contrainte $\beta_2=0$.
Nous pouvons également définir un critère d'adéquation d'un modèle par :

$$R^1(\boldsymbol{\theta}) = \mathbf{1} - \frac{\mathbf{V_n}(\hat{\boldsymbol{\beta}}(\boldsymbol{\theta});\boldsymbol{\theta})}{\mathbf{V_n}(\tilde{\mathbf{q}}(\boldsymbol{\theta});\mathbf{0};\boldsymbol{\theta})}$$

Ce critère nous permet de comparer des modèles entre eux, ainsi celui qui présente le R^1 le plus élevé est le meilleur modèle. Comme pour le coefficient de détermination, R^1 a la propriété : $0 \leq R^1 \leq 1$. En effet, il est évident que : $\hat{V}_n(\theta) \leq \tilde{V}_n(\theta)$. Mais contrairement à R^2 qui permet d'estimer la qualité d'ajustement du modèle dans son ensemble, R^1 est une mesure d'adéquation locale c'est-à-dire pour un quantile précis. En effet, il est facile d'imaginer des situations pour lesquelles une variable 1 X serait significative pour la modélisation des quantiles élevés mais n'apporterait que très peu d'information pour des quantiles faibles.

Remarque : Il existe des processus pour étudier la fonction $\theta \rightarrow R1(\theta)$ dans sa globalité mais ils ne seront pas abordés dans ce mémoire.

2.3.4 Théorie asymptotique :

➢ Propriété 6 : Consistance (El Bantli et Hallin 1999)

Soit *ni F* la fonction de répartition conditionnelle de *Yi* pour $i \in \{1,K,n\}$.

Sous les conditions : $\sqrt{n}\ (ax_n(\epsilon) - \tau) \to \infty$

$$\sqrt{n}\ (\tau - b_n(\epsilon)) \to \infty$$

Avec

$$a_n(\epsilon) = \frac{1}{n}\sum_{i=1}^{n} F_{ni}(x_i^T\beta(\theta) - \epsilon)$$

$$b_n(\epsilon) = \frac{1}{n}\sum_{i=1}^{n} F_{ni}(x_i^T\beta(\theta) + \epsilon)$$

Le coefficient $\hat{\beta}$ est consistent. Nous avons donc :

$$\hat{\beta}_n(\theta) \to \beta(\theta)$$

➢ Propriété 7 : Normalité asymptotique

Sous les hypothèses :
(*i*) Les fonctions de répartition $\{F_i\}$ sont absolument continues, et admet des fonctions de densité $\{f_i\}$ continues et finies dans le voisinage de q_Y (q).
Il existe des matrices définies positives D_0 et D_1 telles que :

i. $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} x_i\, x_i^T = D_0$

ii. $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} f_i(q_y(\theta))x_i\, x_i^T = D_1$

iii. $\frac{1}{\sqrt{n}} max_{i=1\ldots n}\|x_i\| \to 0$

Alors :

$$\sqrt{n}(\hat{\beta}_n(\theta) - \beta_n(\theta)) \to N(0; \theta(1-\theta)D_1^{-1}D_0D_1^{-1}$$

2.3.5 Utilisation :

La méthode des quantiles conditionnels présente certains avantages. Le fait de ne pas faire D'hypothèse sur la distribution des erreurs rend le modèle plus robuste. De plus, contrairement aux modèles linéaires généralisés, la régression quantile ne nécessite pas d'hypothèses sur la distribution de la variable d'intérêt ni de choix de fonction de lien entre les quantiles et les variables explicatives. Une procédure (PROC QUANTREG) a été développée sous le logiciel R pour faciliter son utilisation (voir en annexe la syntaxe de la procédure).

2.4 La théorie des valeurs extrêmes

Un sinistre extrême est caractérisé par une faible probabilité de survenance et un coût élevé. Il est important de modéliser ces sinistres correctement puisqu'ils ont un impact non négligeable sur le résultat du portefeuille. De plus, un modèle qui décrit bien l'ensemble de la distribution n'est pas forcément adéquat pour la modélisation des sinistres extrêmes. Nous utilisons donc une théorie adaptée à la modélisation de tels événements : la théorie des valeurs extrêmes.
Cette partie a pour enjeux de répondre à deux questions :
_ Quels sinistres considérons-nous comme extrêmes ?
_ Comment modéliser ces sinistres extrêmes ?

2.4.1 La méthode par bloc

➢ **Propriété 8 : (Fonction de répartition du maximum)**[2]

Soit $M_n = \max(X_1, \ldots, X_n)$ où $(X_1, \ldots, X_n)$ sont des variables aléatoires réelles i.i.d. de fonction de répartition F.
Alors : $F_{M_n}(x) = P(M_n \leq x) = F(x)^n$

Preuve : $F_{M_n}(x) = P(M_n \leq x)$

$$= P(\max(x_1, \ldots, x_n) \leq x)$$

$$= P\left((\cap_{i=1}^{n} x_i) \leq x\right)$$

$$= \prod_{i=1}^{n} P(x_i \leq x)^n$$

$$= P(x_i \leq x)^n$$

$$= F(x)^n$$

➢ Définition 4 (Distribution GEV)[3]

Trois types de distribution des valeurs extrêmes ont été combinés en une seule famille de loi appelée distribution généralisée des valeurs extrêmes.
Une fonction de répartition appartient à la famille des GEV si elle se met sous une des formes suivantes :

$$H(x) = \exp(-e^{-x})$$

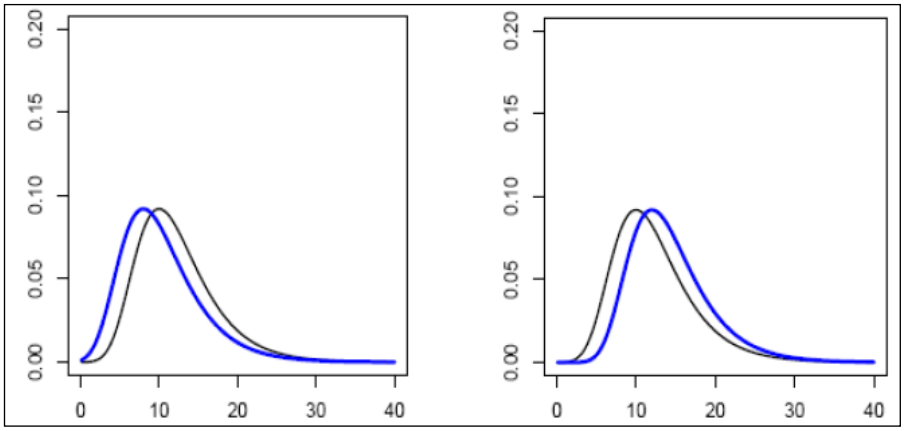
$$H(x) = \begin{cases} 0 & x < 0 \\ \exp(-x^{-\alpha}) & x > 0 \end{cases}$$

$$H(x) = \begin{cases} \exp(-x^{-|\alpha|}) & x < 0 \\ 0 & x > 0 \end{cases}$$

Le premier cas correspond à la loi de Gumbel, le deuxième à la loi de Fréchet et le troisième à la loi de Weibull.
Cette famille peut également être décrite par une seule expression à trois paramètres :

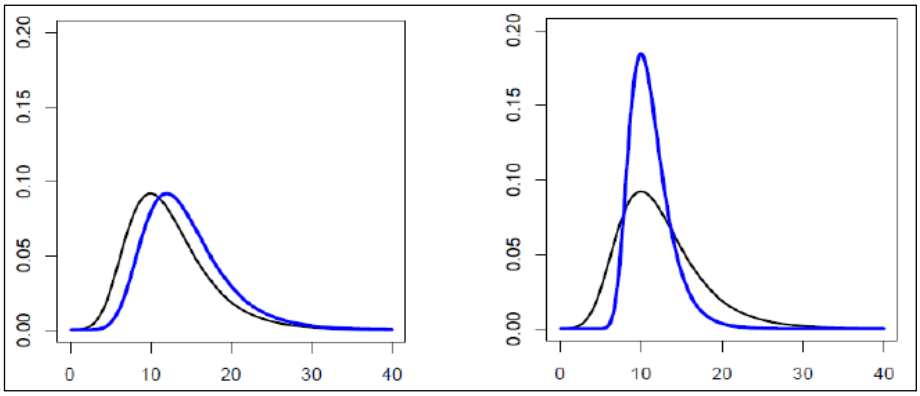
$$H(x) = \exp\left(-\left(1 + \xi\frac{x-\mu}{\psi}\right)_+^{-1/\xi}\right)$$

μ est un paramètre de position qui permet de déterminer où se trouve le poids de la distribution.



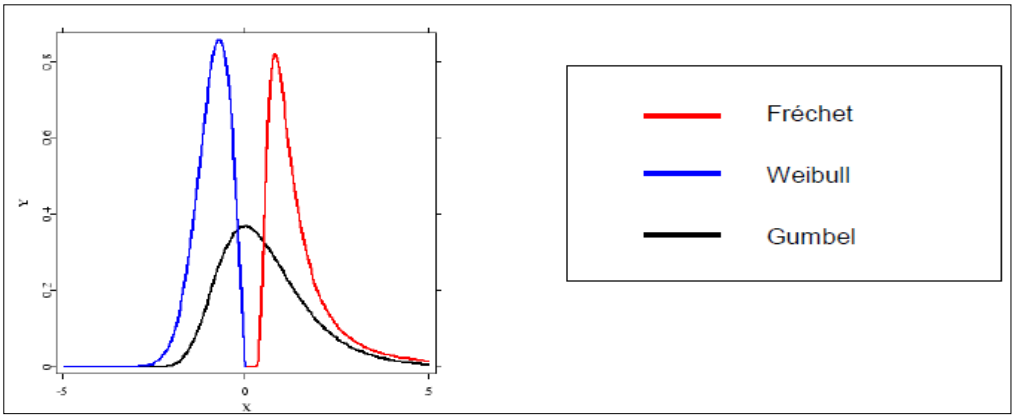
Graphe 2.2 – Impact de la variation du paramètre μ sur la distribution.

ψ est un paramètre d'échelle et détermine la variation des observations extrêmes.



Graphe 2.3 – Impact de la variation du paramètre ψ sur la distribution.

ξ est paramètre de forme qui détermine le type de distribution :



Graphe 2.4 – Impact du signe de ξ sur la distribution.

Nous allons étudier la distribution limite de M_n. Afin d'éviter un problème de dégénérescence : $F_{Mn}(x)$ qui converge vers 0 ou 1 selon la valeur de F(x), la variable M_n est normalisé.

➢ Théorème 1 (Fisher Tippet)

S'il existe deux suites $a_n \geq 0$ et b_n telles que :

$\forall x \in \Re$ $\qquad P\left(\frac{M_n - b_n}{a_n} \leq x\right)^{n \to \infty} \to G(x).$

Où G est une distribution non dégénérée.
Alors G appartient à la famille de loi GEV.

Ainsi, quelque soit la loi de la variable X , ce théorème approche la loi de M_n par une loi GEV.

➢ **Définition 5 : (Domaine d'attraction)**

Le domaine d'attraction maximum de la fonction H, noté *MDA*(*H*), est l'ensemble des lois de fonction de répartition de *F* telles qu'il existe deux suites à $_n \geq 0$ et b $_n$ Telles que $\frac{M_n - b_n}{a_n}$ converge en loi vers une variable aléatoire réelle de fonction de répartition G.

- Le domaine d'attraction maximum de la distribution de Gumbel est composé des distributions qui n'ont pas de queues épaisses comme la loi normale, log-normale, gamma ou de Weibull.
- Le domaine d'attraction maximum de la distribution de Fréchet est composé des distributions à queues épaisses comme la loi de Pareto.
- Le domaine d'attraction maximum de la distribution de Weibull est composé des distributions à support fini comme la loi Beta.

➢ **Difficultés de la mise en œuvre de la méthode par blocs :**

Une première utilisation du théorème de Fisher et Tippet a été développée en 1958 par Gumbel. Il est difficile de déterminer la loi de M_n à partir d'une unique observation. Gumbel propose donc de découper l'échantillon de base en m échantillons et de calculer pour chaque nouvel échantillon le maximum. En assurance, il serait logique de découper l'échantillon par année de survenance du sinistre, ce qui permettrait de construire des blocs ayant un nombre d'observation relativement important. En pratique, l'historique des données n'est pas assez conséquent pour construire assez de blocs pour modéliser correctement la loi des maximums. Découper les observations par bloc peut également entraîner une perte d'information : plusieurs sinistres extrêmes peuvent être de part le découpage dans le même bloc, or seul le sinistre le plus élevé sera considéré comme extrême. Etant donné la rareté de ces sinistres, nous ne pouvons pas nous permettre d'ignorer certaines observations extrêmes. Nous utiliserons donc une autre méthode dans la suite de cette étude.

2.4.2 La méthode dépassement de seuil

La méthode dépassement de seuil a été introduite pour pallier aux problèmes de la méthode par bloc que nous venons d'évoquer. Par la suite, nous utiliserons cette méthode pour déterminer les sinistres extrêmes et les modéliser.

2.4.2.1 Principe

Cette méthode (POT : Peak Over Threshold en anglais) considère comme extrêmes toutes les Observations supérieures un certain seuil, et modélise les dépassements de seuil ou excès par une loi de Pareto généralisée.
Nous allons désormais nous concentrer sur la distribution des sinistres conditionnellement au fait qu'ils dépassent un seuil noté *u*.

Notons : Y = X –u ≥ 0 la variable des excès et F_u sa fonction de répartition.

Nous avons la relation suivante : $F_u(y) = P(Y \leq y | Y > 0) = \frac{F(u+y)-F(u)}{1-F(u)}$

➢ **Définition 6 : (Distribution de Pareto généralisée)**

La loi de Pareto généralisée est définie par la fonction de répartition :

Si $\xi \neq 0 : G_\xi(x) = 1 - (1 - \xi X)^{-1/\xi}$

Si $\xi = 0 : G_\xi(x) = 1 - \exp(-x)$ c'est-à-dire la loi exponentielle de paramètre 1.
Le support de la loi de Pareto généralisée est $[0,\infty[$ pour $\xi \geq 0$ et $]0; -1/\xi[$ pour $\xi < 0$

➢ **Théorème2 (de Pickands, Balkema, de Haan)**

$F \in MDA(G_\xi)$ si et seulement si $\lim_{u \to x_F} \sup_{[0;x_F-u]} |F_u(x) - G_\xi(x)| = 0$

$x_F = \sup\{x; F(x) < 1\}$

G_ξ est la fonction de répartition d'une loi de Pareto généralisée.

Ce théorème nous montre l'importance du choix du seuil. Si le seuil n'est pas suffisamment grand, la convergence n'a pas lieu.

2.4.2.2 Estimations du seuil

Une des difficultés de la modélisation des sinistres extrêmes est de déterminer le seuil à partir duquel les sinistres sont considérés comme extrêmes. Si le seuil est trop bas alors le caractère extrême perd tout son sens et l'intérêt de modéliser séparément les sinistres est faible. Si le seuil est trop élevé, le nombre d'observation de sinistres extrêmes est alors insuffisant pour les modéliser correctement. Il faut donc jongler entre ces deux critères afin de déterminer un seuil adapté à l'échantillon.
Pour fixer le seuil, nous nous intéresserons à plusieurs méthodes graphiques.

- **La fonction moyenne des excès**

➢ **Définition 7 (Fonction moyenne des excès)**

Soit X une v.a.r. telle que $E(X) < \infty$, la fonction moyenne des excès est définie par :

$$e(u) = \frac{E(X-u|x>u)}{u}$$

Pour chaque seuil u, la moyenne des excès supérieurs à ce seuil est calculée. La fonction moyenne des excès empirique est donc définie par :

$$\hat{e}(u) = \frac{\sum_{i=1}^{n}(X_i - u)}{\sum_{i=1}^{n} I_{\{X_i > u\}}}$$

Il est possible de déterminer graphiquement le seuil le plus adéquat en le prenant égal à la valeur de u à partir de laquelle la fonction $\hat{e}$ est linéaire.

- **Estimateur de Hill**

Notation :

Soit (X_1 ,…, X_n) une variable aléatoire réelle. Nous rangeons ces valeurs par ordre croissant et Introduisons la notation suivante : $X_{1:n} \leq X_{2:n} \leq ... \leq X_{n:n}$

$X_{i:n}$: est la i -ème statistique d'ordre.

L'estimateur de ψ le plus utilisé est l'estimateur de Hill (Hill (1975)) défini par :

$$H_{k,n} = \frac{1}{k}\sum_{i=1}^{k} \log\left(\frac{X_{n-i+1:n}}{X_{n-k:n}}\right)$$

C'est estimateur est défini uniquement pour $\xi > 0$, c'est-à-dire dans le cas où la distribution des valeurs extrêmes correspond à une distribution de Fréchet.

k est le nombre de statistiques d'ordre utilisées dans le calcul de l'estimateur.

Il est également possible de l'écrire sous la forme :

$$H_{k,n} = \frac{1}{k}\sum_{i=1}^{k} \mathrm{i} * \log\left(\frac{X_{n-i+1:n}}{X_{n-k:n}}\right)$$

Le graphique de Hill ($k, H_{k,n}$), est également une méthode permettant de la sélection d'un seuil. La valeur de k à partir de laquelle la fonction est assimilable à une droite horizontale désigne le nombre D'observations à considérer comme extrêmes.

- **Stabilité des coefficients**

Notons u_0 la valeur du seuil recherchée. Si l'approximation par une distribution de Pareto généralisée

GPD (σ_{u_0} , ξ) est valable pour $Y > u_0$ alors la modélisation par un modèle

GPD $(\sigma_{u_0} + \xi(u - u_0), \xi)$ Est également valable pour $Y > u$ avec $u_0 > u$. Au-dessus du seuil u0 le paramètre de forme x est constant contrairement au paramètre d'échelle qui est fonction linéaire de u. Pour étudier la stabilité du paramètre d'échelle, il est plus commode d'étudier $\sigma^* = \sigma_u - \xi^*$. La valeur de *u* à partir de laquelle ce paramètre modifié est stable peut être considérée comme seuil.

En pratique, ces méthodes graphiques permettent le plus souvent de sélectionner plusieurs valeurs Du seuil qui semblent convenir et non pas un seul seuil. Il convient ensuite de tester pour chaque Seuil sélectionné la qualité de la modélisation pour déterminer la valeur finale du seuil.
L'approximation du seuil des valeurs extrêmes est une partie délicate de cette théorie, il est souvent difficile de déterminer un seuil de façon certaine. Cette partie de la théorie de valeurs extrêmes a fait l'objet de nombreuses études et est encore aujourd'hui un sujet d'actualité.

2.4.2.3 Estimations des paramètres

Plusieurs méthodes sont utilisées pour estimer les paramètres de la loi de Pareto généralisée, les plus courantes sont la méthode des moments et le maximum de vraisemblance. Nous utiliserons dans la suite de l'étude la méthode du maximum de vraisemblance.
La log-vraisemblance d'un modèle *GPD*(σ ,ξ) est définie par :

$$L = \begin{cases} -n\, log\sigma - \left(1 + \frac{1}{\xi}\right) \sum_{i=1}^{n} log\left(1 + \frac{\xi * y_i}{\sigma}\right) & \xi \neq 0 \\ -n \log \sigma - \frac{1}{\sigma} \sum_{i=1}^{n} y_i & \xi = 0 \end{cases}$$

En dérivant par rapport à chaque paramètre (en se plaçant donc dans le cas x ¹ 0), le système D'équation est le suivant :

$$\begin{cases} \frac{1}{n} \sum_{i=1}^{n} log\left(1 + \frac{\xi * y_i}{\sigma}\right) = \xi \\ \frac{1}{n} \sum_{i=1}^{n} \frac{1}{1 + \xi \frac{y_i}{\sigma}} = \frac{1}{1+\xi} \end{cases}$$

Le système obtenu n'admet pas de solution explicite, il est donc nécessaire d'avoir recours à des méthodes numériques pour estimer les coefficients.

Remarque : Dans le cas 0 = ξ , on trouve $\hat{\sigma} = \frac{1}{n} \sum_{i=1}^{n} Y_i$

2.4.3 Utilisation

Pour mettre en place cette méthode d'estimation des coûts, nous devons :
_ choisir un seuil adéquat à partir des méthodes graphiques
_ ajuster une loi de Pareto généralisée à l'échantillon des excès
_ estimer empiriquement la probabilité d'être un sinistre extrême
_ calculer le coût des sinistres estimés
La partie la plus délicate dans l'utilisation de la théorie des valeurs extrêmes est l'estimation du seuil.
Nous nous servons de méthodes graphiques pour le déterminer, nous n'obtenons donc pas une Unique valeur possible. Il revient donc à l'utilisateur de tester les différentes possibilités.

Chapitre 3

Mise en œuvre sous R

3.Modélisation du coût des sinistres

3.1 Présentation des données

Afin de modéliser les sinistres extrêmes, on possède quatre bases de données sous forme de quatre tableaux des années 2017, 2018, 2019 et 2020, Ces quatre bases contiennent des informations tels que (ID, nom d'agence, nom de DR, branche, le type de garantie, le cout du sinistre ...).

Définition des données :

1.'ID' : il est composé de ['numéro de police /le code DR / le code agence'].
2. Nom de DR : c'est le nom la direction régionale dans notre cas c'est (Tizi Ouzou).
3. Branche : dans notre cas elle représente la branche automobile.
4. Type du contrat : BDG (brille de glace), DC (dommage collision), DASC (dommage avec et sans contrepartie), RC (risque obligatoire), TR (tous risque) et VI (vol et incendie).
5. Nom d'agence : THENIA, SOUR EL GHOUZLANE, DRAA EL MIZAN, AZAZGA, LAKHDARIA...

3.2 Logiciels utilisés

Les analyses statistiques présentées dans ce rapport ont été réalisées sous R (versions 2.12.2 à 2.14.2)

3.3 Présentation du problème

La distribution des sinistres a une queue lourde difficile à modéliser et de ce fait délicat à étudier, nous avons notre étude sur la prise en compte des sinistres extrêmes en assurance automobiles. Deux types de sinistralités se distinguant : la sinistralité antirationnelle sont les sinistres de forte fréquence et de faible coût, et les sinistres rares mais d'intensité extrême, ces sinistres sont qualifiés comme extrêmes ou graves et sont caractérisés par un coût très élevé ainsi que d'une probabilité de survenance faible par rapport aux autres sinistres, ils ont des répercussions importantes sur le résultat du portefeuille. Le but de la présente démarche est d'établir Une modélisation globale de deuxième type de la sinistralité à fin d'en déterminer un seuil critique au-delà duquel les sinistres puissent être considérés comme un sinistre extrême.
Dans ce dernier chapitre nous présentons une application des notions vues précédemment.

3.4. Analyse préliminaire

3.4.1. Analyse descriptive

La garantie BDG :

```
BDG<-subset(data1, data1$garantie4 == 2)
>summary(BDG)
   garantie4 cout.de.sinis4e
    Min.   :  100
   1st Qu.:  2100
  Median : 13000
    Mean   : 16861
   3rd Qu.: 25000
    Max.   :450000
>sd(BDG$cout.de.sinis4e)
[1] 20815.04
>length(BDG$cout.de.sinis4e)
[1] 29660
```

Figure 3.4 – Statistiques descriptives de la garantie BDG de 2017 à 2020 sous R.

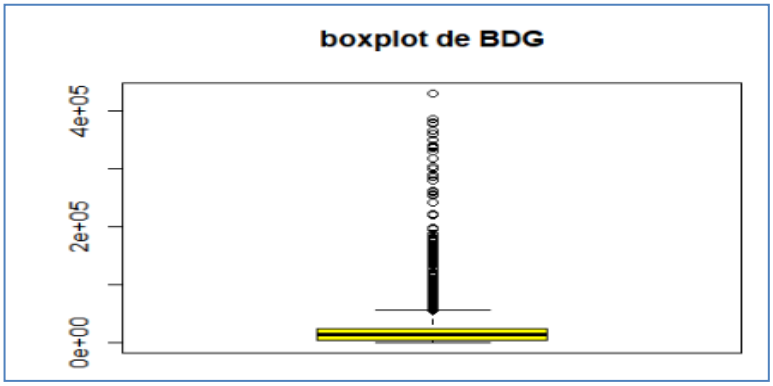
On résume les résultats dans le tableau suivant :

Garantie	Min	Max	Moyenne	Médiane	1Q	3 Q	Fréquence	Ecart type
BDG	100	450000	16861	13000	2100	25000	29660	20815.04

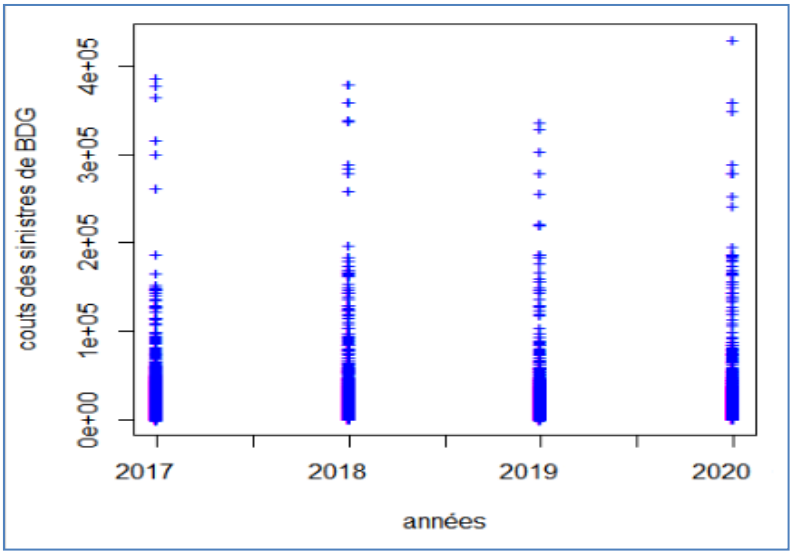
Tableau 3.2 – Statistiques descriptives de la garantie BDG de 2017 à 2020.

Commentaire :

D'après cette table nous constatons que sur les 29660 sinistres. Le maximum du cout de sinistre est 450000 DA es .la moitié à un cout supérieur à 2100DA. Le cout moyen des sinistres est 16861DA avec une dispersion autour de la moyenne estimée à 20815.04 DA, donc les couts des sinistres sont dispersés.



Graphe 3.5 –box plot des couts des sinistres de la garantie BDG.



Graphe 3.6– Distribution des sinistres BDG par année.

Interprétation :

La distribution des sinistres BDG. Le nombre de sinistres impliquant la garantie BDG diminue de 2017 à 2019. En 2020 sont augmentés énormément.

La garantie DASC :

```
DASC<-subset(data1, data1$garantie4 == 5)
>summary(DASC)
 cout.de.sinis4e
  Min.   :    0.01
 1st Qu.:  4260
Median : 13500
 Mean   : 36508
3rd Qu.: 32987
 Max.   :497500
>sd(DASC$cout.de.sinis4e)
[1] 72444.75
>length(DASC$garantie4)
[1] 11712
```

Figure 3.5– statistiques descriptives de la garantie BDG de 2017 à 2020 sous R.

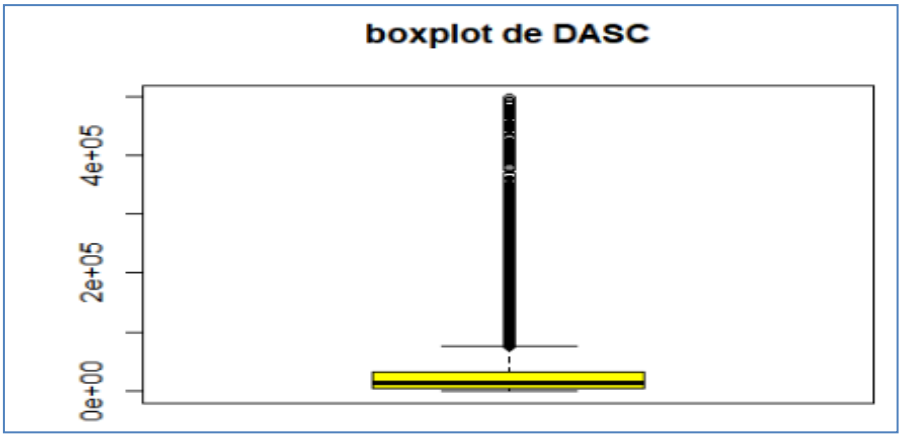
On résume les résultats dans le tableau suivant :

Garantie	Min	Max	Moyenne	Médiane	1Q	3 Q	Fréquence	Ecart type
DASC	0.01	497500	36508	13500	4260	32987	11712	72444.75

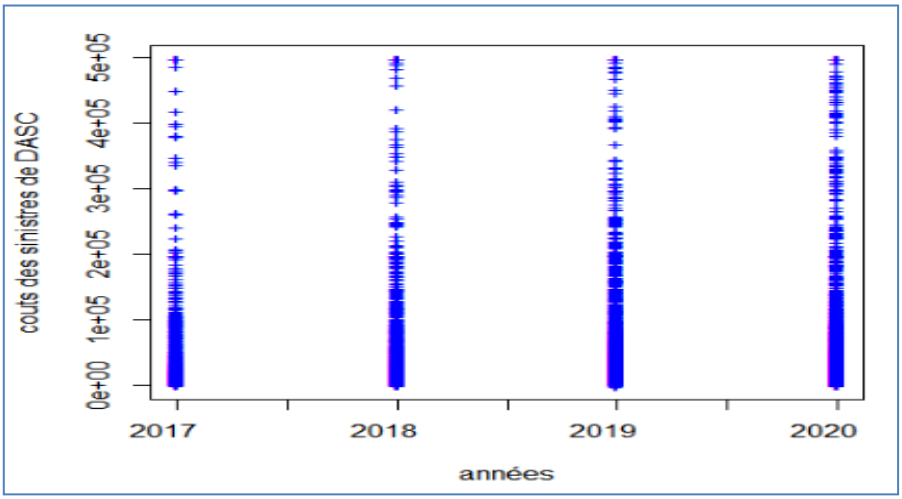
Tableau 3.3– statistiques descriptives de la garantie DASC de 2017 à 2020.

Commentaire :

D'après ce tableau nous constatons que sur les 11712 sinistres. Le maximum du cout de sinistre est 497500 DA. La moitié a un cout supérieur à 4260DA. Le cout moyen des sinistres est 36508DA avec une dispersion autour de la moyenne estimée à 72444.75DA est élevée cela indique que les couts des sinistres sont dispersés.



Graphe 3.7–box plot des couts des sinistres de la garantie DASC



Graphe 3.8– Distribution des sinistres DASC par année.

Interprétation :

La distribution des sinistres DASC. Le nombre de sinistres impliquant la garantie DASC s'élèvent progressivement au cours du temps.

La garantie DC :

```
DC<-subset(data1, data1$garantie4 == 3)
>summary(DC)
cout.de.sinis4e
   Min.   :     0.01
 1st Qu.:  4000
Median :  9500
 Mean   : 12664
3rd Qu.: 18000
 Max.   :1142998
>sd(DC$cout.de.sinis4e)
[1] 17255.74
>length(DC$cout.de.sinis4e)
[1] 74074
```

Figure 3.6– statistiques descriptives de la garantie DC de 2017 à 2020 sous R.

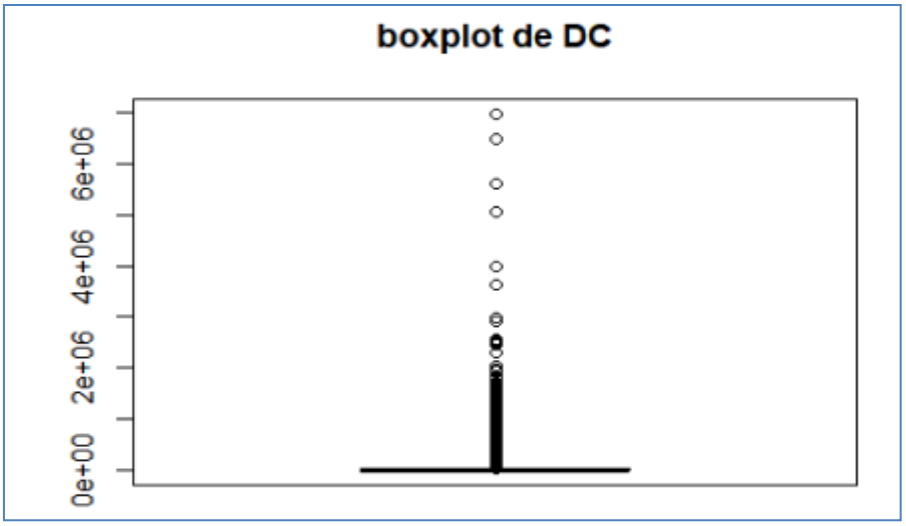
On résume les résultats dans le tableau suivant :

Garantie	Min	Max	Moyenne	Médiane	1Q	3 Q	Fréquence	Ecart type
DC	0.01	1142998	12664	9500	3170.1	18000	74074	17255.74

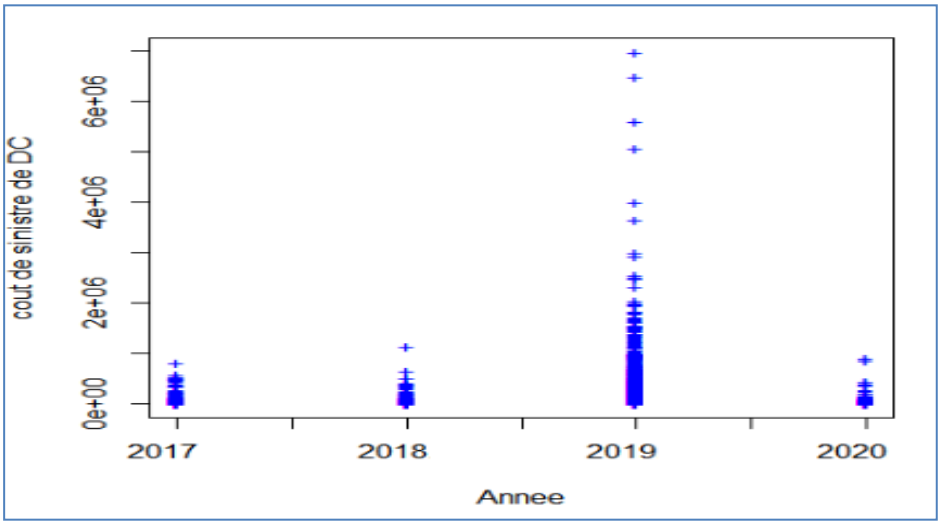
Tableau 3.4 – statistiques descriptives de la garantie DC de 2017 à 2020.

Commentaire :

D'après cette table nous constatons que sur les 74074sinistres. Le maximum du coût de sinistre est 1142998 DA. La moitié à un cout supérieur à 3170.1DA. Le cout moyen des sinistres est 12664DA avec une dispersion autour de la moyenne estimée à 17255.74 DA est très élevée indique que les couts des sinistres sont dispersés.



Graphe 3.9– box plot des couts des sinistres de la garantie DC.



Graphe 3.10 – Distribution des sinistres DC par année.

Interprétation :

La distribution des sinistres DC. Les coûts des sinistres en année de 2017,2018 et 2020 sont faible par contre en 2019 on remarque que le nombre de sinistre est augmenté. Même le coût de sinistre qui a augmenté jusqu'à 6000.

La garantie RC :

```
RC<-subset(data1, data1$garantie4 == 1)
>summary(RC)
  cout.de.sinis4e
  Min.   :    0.01
  1st Qu.:  2300
  Median : 11650
  Mean   : 30287
  3rd Qu.: 37148
  Max.   :6977332
>sd(RC$cout.de.sinis4e)
[1] 74496.39
length(RC$cout.de.sinis4e)
[1] 76559
```

Figure 3.7– statistiques descriptives de la garantie RC de 2017 à 2020 sous R.

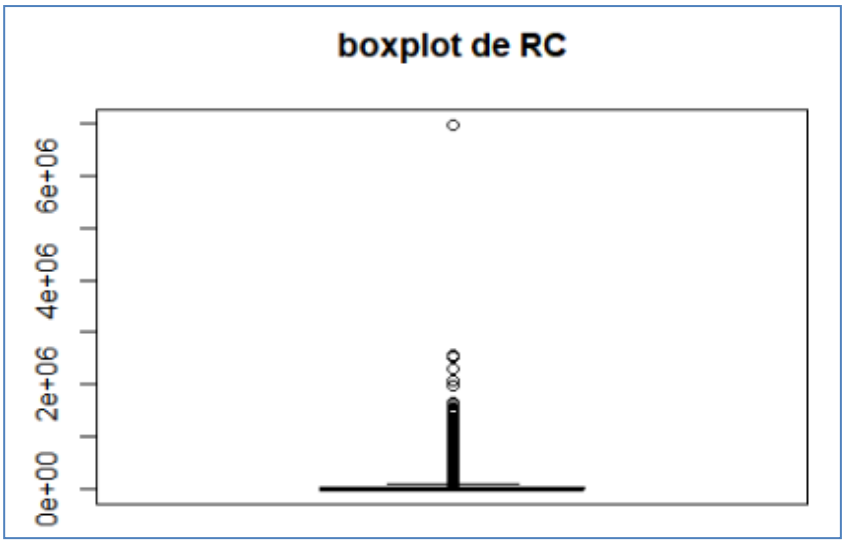
On résume les résultats dans le tableau suivant :

Garantie	Min	Max	Moyenne	Médiane	1Q	3 Q	Fréquence	Ecart type
RC	0.01	6977332	30287	11650	2300	37148	76559	74496.39

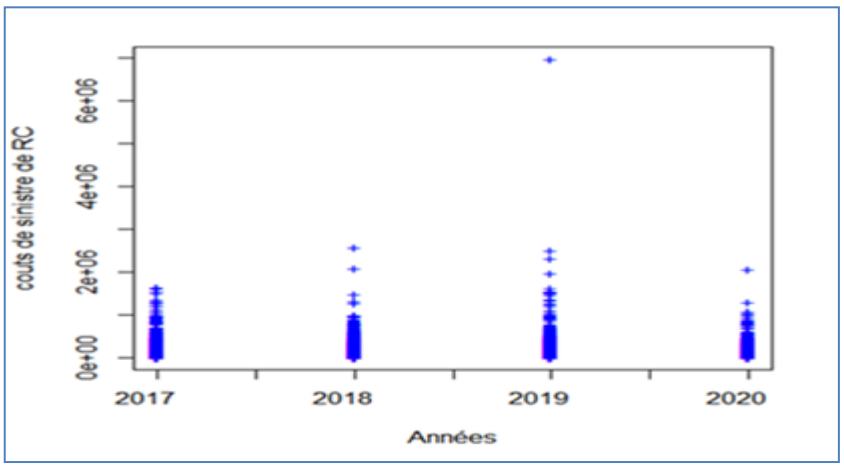
Tableau 3.5– statistiques descriptives de la garantie RC de 2017 à 2020.

Commentaire :

D'après cette table nous constatons que sur les 76559 sinistres. Le maximum du cout de sinistre est 6977332 DA. La moitié à un cout supérieur à 2300.1DA. Le cout moyen des sinistres est 30287DA avec une dispersion autour de la moyenne estimée à 74496.39DA est très élevée indique que les couts des sinistres sont dispersés.



Graphe 3.11 – box plot des couts des sinistres de la garantie RC.



Graphe 3.12 – Distribution des sinistres RC par année.

Interprétation :

La distribution des sinistres RC. Le cout de sinistre est augmenté dans les années 2017, 2018 et 2019. Après ils diminuent dans l'année 2020. On remarque que le nombre de sinistre le plus grand est plus que 6000000 sinistres dans l'année 2019.

La garantie TR :

```
TR<-subset(data1, data1$garantie4 == 4)
>summary(TR)
cout.de.sinis4e
  Min.   :     0.01
 1st Qu.:  3050
Median : 11500
 Mean   : 42781
 3rd Qu.: 30978
  Max.   :7335000
>sd(TR$cout.de.sinis4e)
[1] 148267.2
>length(TR$garantie4)
[1] 44349
```

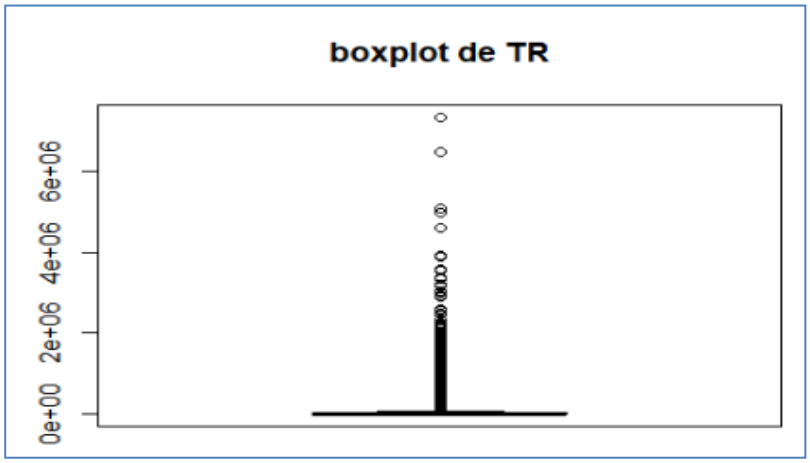
Figure 3.8– statistiques descriptives de la garantie TR de 2017 à 2020 sous R.

On résume les résultats dans le tableau suivant :

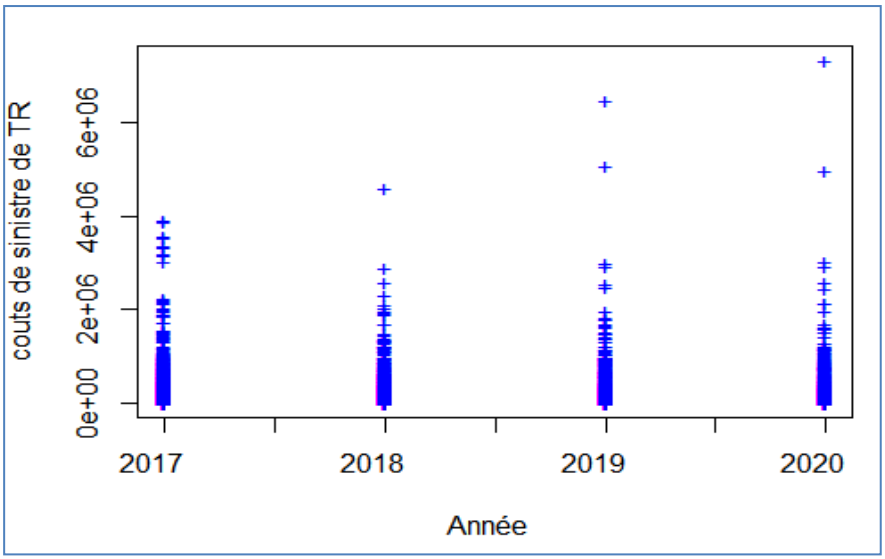
Garantie	Min	Max	Moyenne	Médiane	1Q	3Q	Fréquence	Ecart type
TR	0.01	7335000	42781	11500	3050	30978	44349	148267.2

Tableau 3.6 – statistiques descriptives de la garantie TR de 2017 à 2020.

Commentaire :
D'après cette table nous constatons que sur les 44349 sinistres. Le maximum du coût de sinistre est 7335000 DA .la moitié à un coût supérieur à 3050DA. Le cout moyen des sinistres est 42781DA avec une dispersion autour de la moyenne estimée à 148267.2DA est très élevée indique que les couts des sinistres sont dispersés.



Graphe 3.13 – box plot des couts des sinistres de la garantie TR.



Graphe 3.14– Distribution des sinistres TR par année.

Interprétation :

La distribution des sinistres TR. le out de sinistre est augmenté dans au fil des ans. Dans les années 2017 et 2018 la charge de cout de sinistre est beaucoup et les valeurs sont proches par contre dans les années 2019 et 2020 les valeurs sont très éloignées et la plus grande valeur est plus que 7000000 en 2020.

La garantie VI :

```
VI<-subset(data1, data1$garantie4 == 6)
>summary(VI)
cout.de.sinis4e
  Min.   :    270
  1st Qu.:   7600
 Median :  35148
  Mean   : 432831
  3rd Qu.: 298930
  Max.   :10206000
>sd(VI$cout.de.sinis4e)
[1] 979845.3
>length(VI$garantie4)
[1] 425
```

Figure 3.9– statistiques descriptives de la garantie VI de 2017 à 2020.

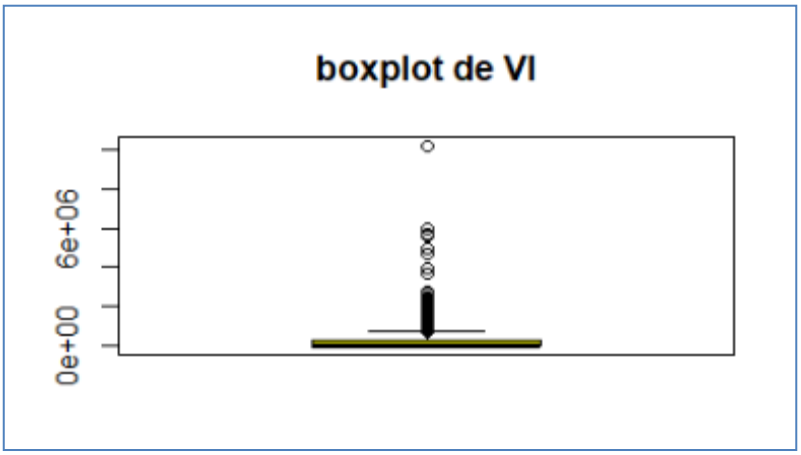
On résume les résultats dans le tableau suivant :

Garantie	Min	Max	Moyenne	Médiane	1Q	3Q	Fréquence	Ecart type
VI	270	10206000	432831	35148	7600	298930	425	979845.3

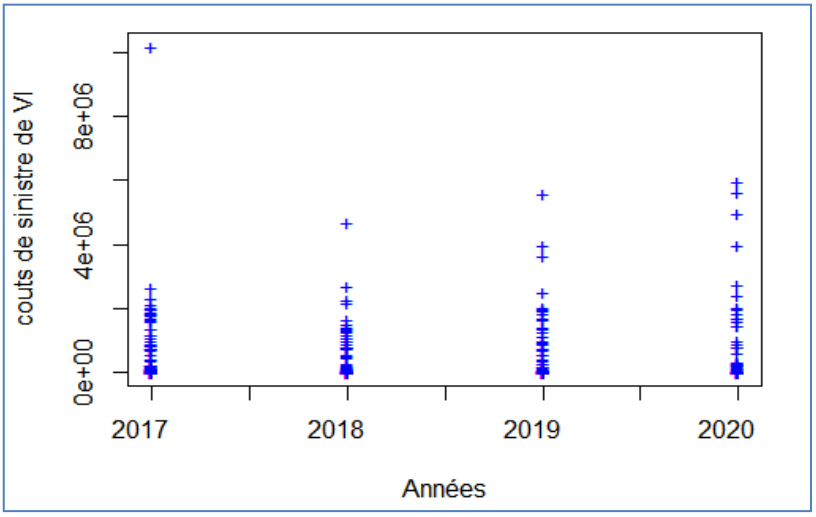
Tableau 3.7 – statistiques descriptives de la garantie VI de 2017 à 2020.

Commentaire :

D'après cette table nous constatons que sur les 425 sinistres. Le maximum du cout de sinistre est 10206000 DA .la moitié à un cout supérieur à 7600DA. Le cout moyen des sinistres est 432831DA avec une dispersion autour de la moyenne estimée à 979845.3DA est très élevée indique que les couts des sinistres sont dispersés.



Graphe 3.15– box plot des couts des sinistres de la garantie VI.



Graphe 3.16 – Distribution des sinistres VI par année.

Interprétation :

La distribution des sinistres VI. Au fur et à mesure, les couts des sinistres augmentent par contre la charge diminue au fil des ans.

Remarque :

Nous remarquons pour les six garanties la présence de points isolés et particulièrement élevés, ce sont des sinistres extrêmes.

3.4.2. Test de Spearman

Corrélation de Spearman est une mesure de dépendance statistique non paramétrique entre deux variables.

La corrélation de Spearman est étudiée lorsque deux variables statistiques semblent corrélées . Elle consiste à trouver un coefficient de corrélation, non pas entre les valeurs prises par les deux variables mais entre les rangs de ces valeurs.

Année	Moyenne	Spearman
2017	27105	-0.005339036
2018	26331	-0.0639597
2019	26085	0.01454183
2020	26232	0.04058397
2017-2020	26463	-0.004426005

Tableau 3.8– Tests de Spearman des couts de sinistre.

Commentaire :

Après avoir le tableau (3.8) effectué le test de Spearman sur nos données, nous obtenons les faible Rho de Spearman qui sont inférieur à (0,1) donc ils convergent vers 0 cela veut dire que nous acceptons l'hypothèse H0.

Conclusion :

Nous acceptons l'hypothèse H0, nous concluons donc qu'il n'ya pas de corrélation entre la fréquence et le coût moyen de sinistre.

- **Le test de Kolmogorov-Smirnov**

En appliquant le test de Kolmogorov-Smirnov **sous** R pour les coûts des sinistres dans chaque Exercice et la période totale de 2017 au 2020.nous obtenons les résultats suivants :

Année	D (valeur d'écart max)	p. value
2017	0.99952	2.2e-16
2018	0.9993	2.2e-16
2019	0.99973	2.2e-16
2020	0.99981	2.2e-16
2017-2020	0.99958	2.2e-16

Tableau 3.9–teste kolmogrov-smirnov de chaque année et la période totale

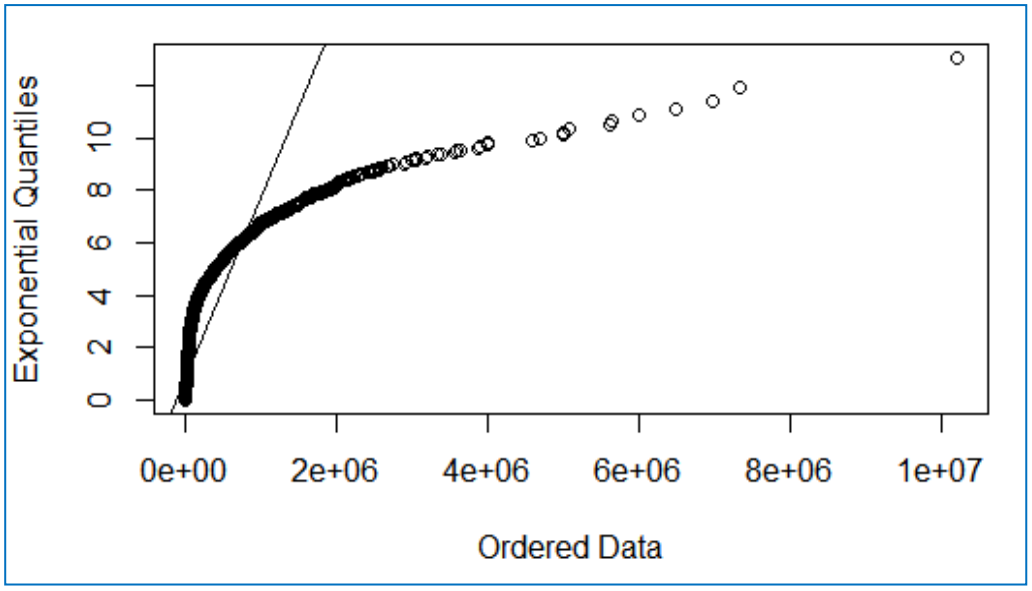
De 2017 à 2020

Commentaire :
Les résultats du tableau (3.9) montrent que le coût de sinistre n'est pas normalement distribué dans chaque exercice et dans le total des exercices (2017 à 2020), car (p. value est inférieur à 0,05).

3.4.2. QQ-Plot

Le QQ-plot ou Quantile-Quantile graphique permet de comparer la distribution d'un échantillon avec une distribution théorique. L'abscisse représente les quantiles empiriques et l'ordonnée les quantiles Théoriques de la loi considérée. Si les données suivent la loi théorique, alors les points doivent être alignés en ligne droite.
-Ce graphique permet de répondre rapidement à la question : la modélisation des données par la loi Considérée est-elle plausible ?
-Si les points ont une forme convexe, alors la queue de la loi est légère, si les points sont

Présentés de manière concave, la queue est épaisse.



Graphe 3.17 – QQ-plot du cout de sinistre total (2017/2020)

Interprétation :

Ici, l'adéquation des données à une loi exponentielle, pour les sinistres bas est très bonne. Par contre La modélisation des sinistres extrêmes n'est pas très adaptée. À partir de ce graphe nous choisissons le seuil à **150 0000.**

Conclusion :

La régression quantile pour déterminer les valeurs extrêmes. Le principal avantage de cette méthode est sa robustesse mais un inconvénient est la modélisation de ces sinistres extrêmes par une loi de la famille exponentielle qui n'est pas la plus adaptée à ce type de données.

3.5. Méthode linéaire généralisée

3.5.1. Identification du Modèle :

Identifier le modèle ajusté en fonction de la moyenne et de la variance de la variable du coût de sinistre.

```
> summary(data1$cout.de.sinis4e)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
       0     3004    10100    26463    24477 10206000
> var(data1$cout.de.sinis4e)
[1] 8465166611
> sd(data1$cout.de.sinis4e)^2
[1] 92006.34
```

Figure 3.10– Moyenne et variance du coût de sinistre de 2017 à 2020 sous R.

Moyenne	26462.52
Variance	8465166611
Ecart	92006.34

Tableau 3.10– Moyenne et variance du coût de sinistre2017 à 2020.

Commentaire :

Sur la base des résultats du tableau (3.10), la variance (8465166611) du coût de sinistre est plus forte que la moyenne (92006.34) de sorte que la régression de la loi binomiale négative est la plus adaptée nous données que la régression de la loi Poisson.

On a calculé les tests de Wald en utilisant l'erreur standard de sandwich.

3.5.2. Régression binomiale négative

```
Call:
glm(formula = data1$cout.de.sinis4e ~ data1$garantie4, family = negative.binomial(2),
    data = data1, weights = data1$gcout.de.sinis4)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.6682  -1.2661  -0.6507  -0.0467  19.1087

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     9.968311   0.005833 1709.07   <2e-16 ***
data1$garantie4 0.042347   0.002070   39.78   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(0.6201) family taken to be 1)

    Null deviance: 293434  on 236778  degrees of freedom
Residual deviance: 291177  on 236777  degrees of freedom
AIC: 5248744

Number of Fisher Scoring iterations: 1


              Theta:  0.62012
          Std. Err.:  0.00152

 2 x log-likelihood:  -5248738.39800
```

Figure 3.11–résultat de GLM (régression binomiale négative) sous R.

Interprétation de la vraisemblance :

On remarque que **le max de la vraisemblance** dans les résultats au-dessus est égale a

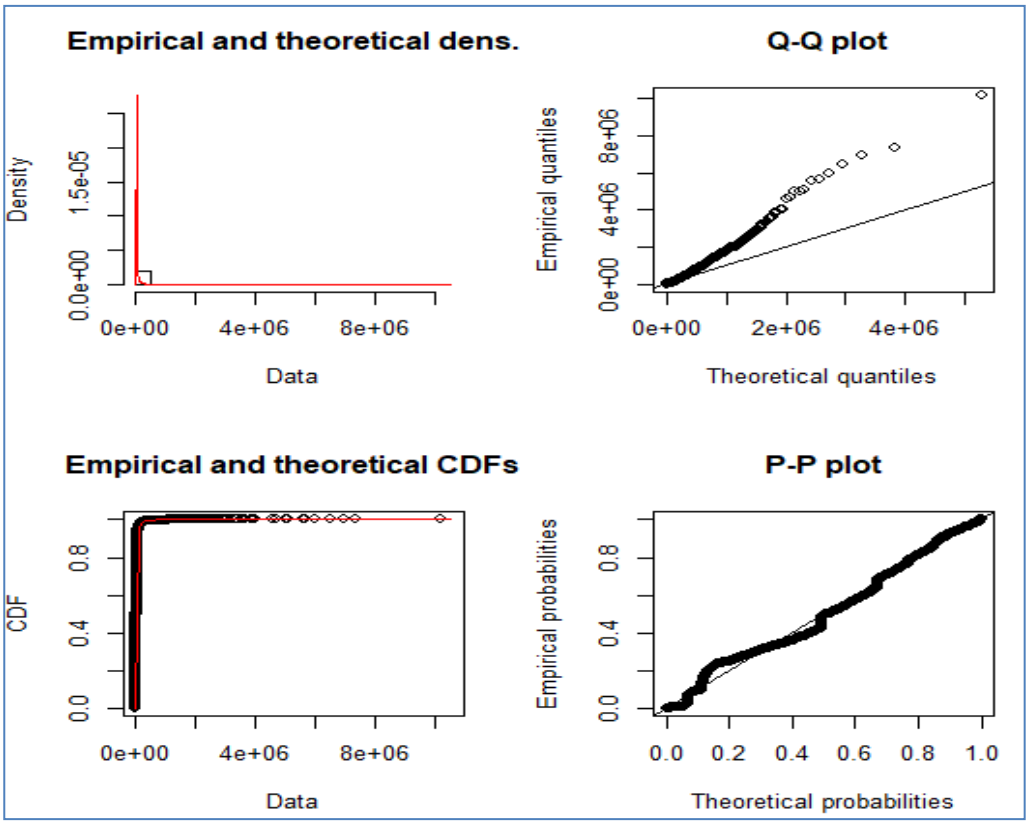
Log(L)=-5 246 738.39 800 ce qui correspond à une valeur infime de la vraisemblance : $L=e^{(-5\ 246\ 738.39\ 800)}$

Régression	Paramètres	Estimations	Std. Error	T value	Pr (>|t|)
Binomiale	Constant	9.96839	0.01473	676.52	<2e-16 ***
Négative	Type garantie	0.04232	0.00523	15.74	<2e-16 ***

Tableau 3.11– GLM (régression binomiale négative).

Commentaire :

D'après le tableau (3.11) on remarque que la variable type garantie est statiquement significative (P-value<0.05).



Graphe 3.18 – Distribution Négative binomiale.

Conclusion :

On conclut donc que le modèle de régression négative binomiale est très utile pour la modélisation de nos données d'après la représentation graphique et le tableau **(3.10)**

3.6. Méthode de Régression Quantile

Ce qui nous intéresse dans cette partie c'est d'avoir l'impact des types de garanties sur les couts du sinistre et les modéliser.

```
Call: rq(formula = data1$cout.de.sinis4e ~ data1$garantie4, tau = c(0.5,
    0.75, 0.9, 0.995), data = data1)

tau: [1] 0.5

Coefficients:
                 Value       Std. Error  t value     Pr(>|t|)
(Intercept)     11361.73333   102.84936   110.46965     0.00000
data1$garantie4  -465.43333    31.72134   -14.67256     0.00000

Call: rq(formula = data1$cout.de.sinis4e ~ data1$garantie4, tau = c(0.5,
    0.75, 0.9, 0.995), data = data1)

tau: [1] 0.75

Coefficients:
                 Value       Std. Error  t value     Pr(>|t|)
(Intercept)     30200.64000   216.68045   139.37870     0.00000
data1$garantie4 -2300.64000    71.78387   -32.04954     0.00000

Call: rq(formula = data1$cout.de.sinis4e ~ data1$garantie4, tau = c(0.5,
    0.75, 0.9, 0.995), data = data1)

tau: [1] 0.9

Coefficients:
                 Value       Std. Error  t value     Pr(>|t|)
(Intercept)     44617.51500   288.31686   154.75167     0.00000
data1$garantie4   127.49500    77.25732     1.65026     0.09889

Call: rq(formula = data1$cout.de.sinis4e ~ data1$garantie4, tau = c(0.5,
    0.75, 0.9, 0.995), data = data1)

tau: [1] 0.995

Coefficients:
                 Value         Std. Error   t value      Pr(>|t|)
(Intercept)     251300.00000  13637.14883     18.42761      0.00000
data1$garantie4  91700.00000   6849.73053     13.38739      0.00000
```

Figure 3.12– Régression quantile sous R.

Voici le résumé des résultats de la méthode de la régression quantile dans le tableau ci-dessus :

Taux		Coefficients	Std. Error	T value	Pr(>|t|)
Tau : [1] 0.5	**Intercepte**	11361.73333	102.84936	110.46965	0.00000
	Garantie	-465.43333	31.72134	-14.67256	0.00000
Tau : [1] 0.75	**Intercepte**	30200.64000	216.68045	139.37870	0.00000
	Garantie	-2300.64000	71.78387	-32.04954	0.00000
Tau : [1] 0.90	**Intercepte**	44617.51500	288.31686	154.75167	0.00000
	Garantie	127.49500	77.25732	1.65026	0.09889
Tau : [1] 0.995	**Intercepte**	251300.00000	13637.14883	18.42761	0.00000
	Garantie	91700.00000	6849.73053	13.38739	0.00000

Tableau 3.12– Régression quantile

Commentaire :

D'après le tableau (3.12), les résultats les plus intéressants sont les coefficients de garantie,

Il y a plusieurs choses notables à propos de ces résultats :

- La garantie est statistiquement moins significative au niveau de quantile 99.5% (p-value>0.05).

- La garantie est statistiquement significative pour les quantiles 50%,75 %, 90% et 99,5% (P-value<0.05)
- L'impact de garantie sur les coûts de sinistre augmente selon l'augmentation du quantile 75 % au quantile 99,5 %. Parce que la valeur des coefficients augmente de (-2300.64 jusqu'à 251300).

3.7. Méthode de la théorie des valeurs extrêmes

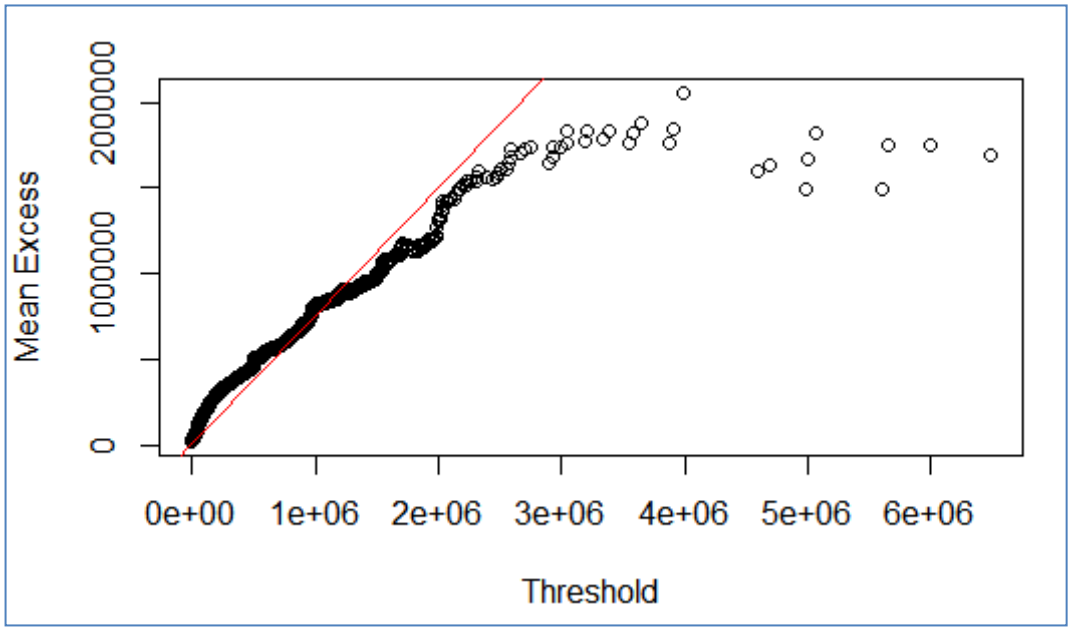
Nous appliquerons cette méthode aux toutes garanties et e particuliers à chaque garantie. Pour sélectionner un seuil nous utilisons les méthodes graphiques développées dans la partie 2.4.1

3.7.1. La fonction moyenne des excès

C'est une estimation graphique qui se base sur la visualisation des points qui se présentent sous la forme d'une droite donc Si à un moment donné, elle devient linéaire, les données suivent la distribution GPD. En utilisant sous R la commande "meplot" qui se trouve dans le package "evir".

Le $\mathbf{k_{min}}$ et le $\mathbf{k_{max}}$ sont lus à partir du graphe comme étant la projection des deux extrémités De la partie linéaire du graphe sur l'axe des abscisses.
En appliquant la méthode sur les couts des sinistres : de 2017 à 2020 pour différent type du garantie PDG, DASC, DC, RC, TR et VI On obtient les graphes suivants :

➢ **La fonction moyenne des excès pour toutes les Garantie**

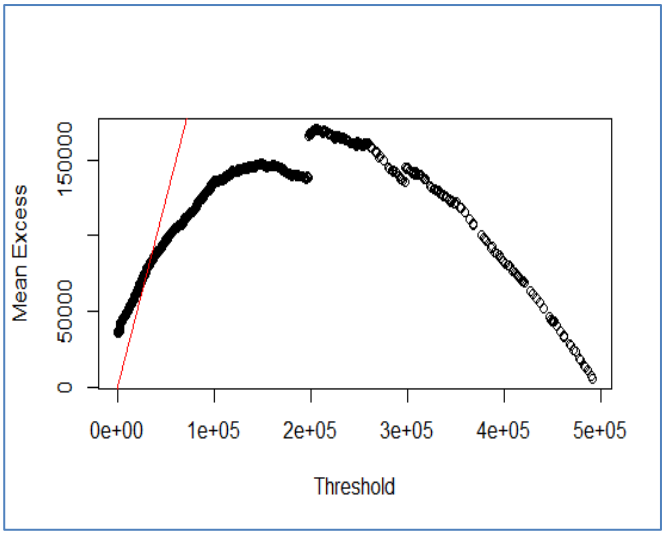


Graphe 3.19– La fonction moyenne des excès du cout de sinistre (pour toutes les garanties).

Interprétation :

Au vu de ce graphe, la linéarité du coût de sinistre commence à 2 000 000 et au-delà du seuil de 3 000 000 cette linéarité n'est plus assurée. Ce graphe nous montre un seuil à partir duquel on considère que les coûts de sinistre sont graves. On peut donc dire que le seuil sera dans l'intervalle **A = [2 000 000 ; 3 000 000].**

➢ **La fonction moyenne des excès pour chaque Garantie**

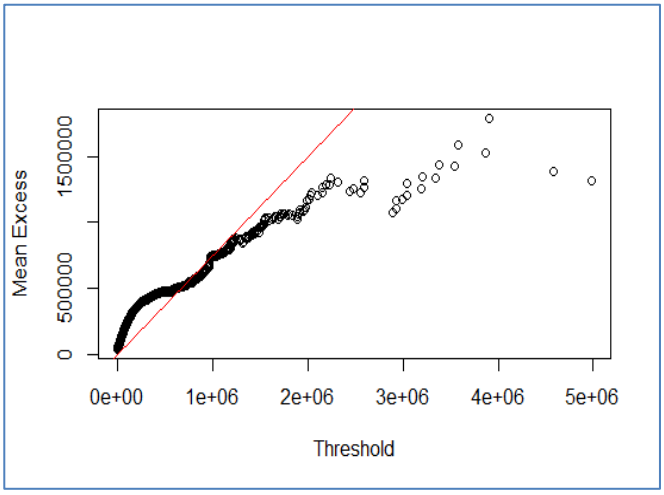


Interprétation :

La linéarité du cout de sinistre pour la garantie **DASC** commence à 100 000 et au-delà du seuil de 200 000 cette linéarité n'est plus assurée. Le graphe DASC nous montre un seuil à partir duquel on considère que les coûts sont sérieux. On peut donc dire que le seuil sera dans l'intervalle

b= [100 000 ; 200 000].

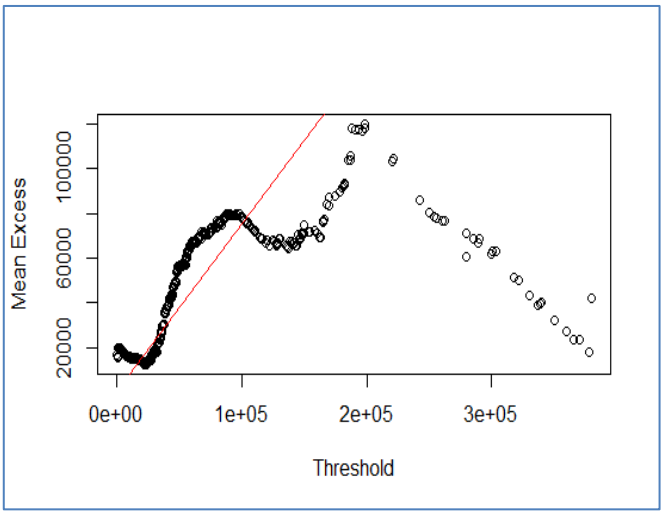
Graphe 3.20 –La fonction moyenne des excès de DASC.



Interprétation :

La linéarité du cout de sinistre pour la garantie **TR** commence à 1000 000 et au-delà du seuil de 1800 000 cette linéarité n'est plus assurée. Le graphe TR nous montre un seuil à partir duquel on considère que les coûts sont sérieux. On peut donc dire que le seuil sera dans l'intervalle c= [1000000 ; 1800000].

Graphe 3.21 –La fonction moyenne des excès de TR.

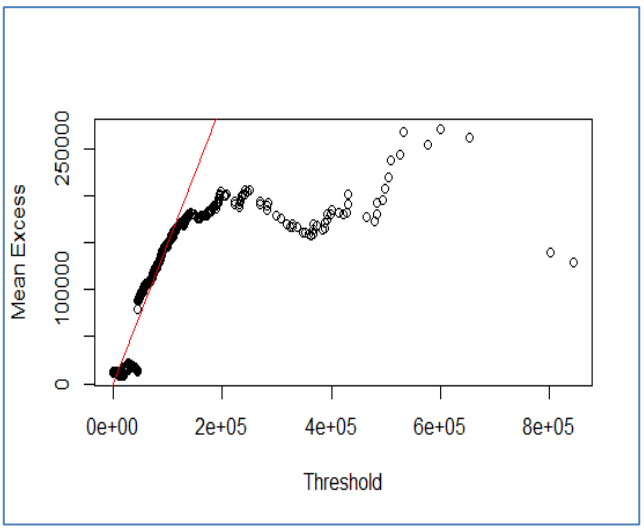


Interprétation :

La linéarité du cout de sinistre pour la garantie **BDG** commence à 100 000 et au-delà du seuil de 180 000 cette linéarité n'est plus assurée. Le graphe BDG nous montre un seuil à partir duquel on considère que les coûts sont sérieux. On peut donc dire que le seuil sera dans l'intervalle

d= [100 000 ; 180 000].

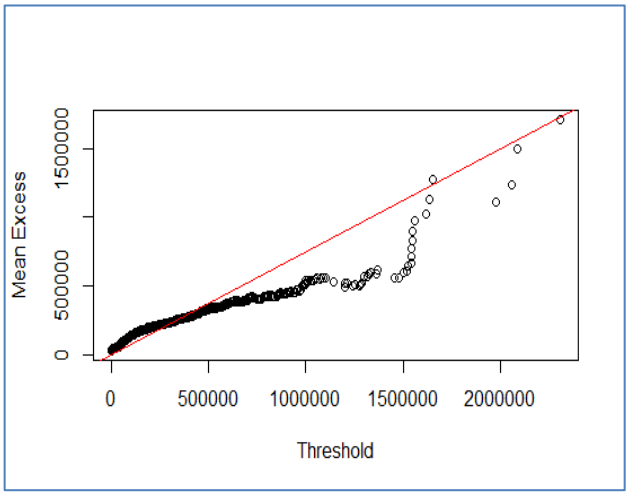
Graphe 3.22– La fonction moyenne des excès de BDG.



Interprétation :

La linéarité du cout de sinistre pour la garantie **DC** commence à 1000 000 et au-delà du seuil de 300 000 cette linéarité n'est plus assurée. Le graphe DC nous montre un seuil à partir duquel on considère que les coûts sont sérieux. On peut donc dire que le seuil sera dans l'intervalle e= [100 000 ; 300 000].

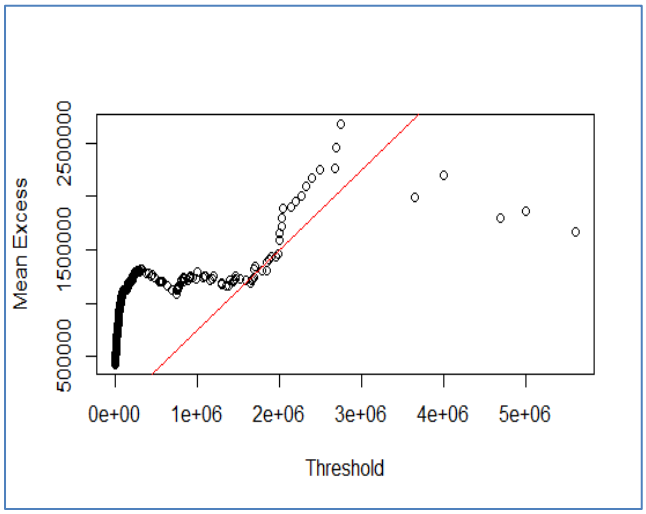
Graphe 3.23 –La fonction moyenne des excès de DC.



Interprétation :

La linéarité du cout de sinistre pour la garantie **RC** commence à 200 000 et au-delà du seuil de 500 000 cette linéarité n'est plus assurée. Le graphe RC nous montre un seuil à partir duquel on considère que les coûts sont sérieux. On peut donc dire que le seuil sera dans l'intervalle f = [200 000 ; 500 000].

Graphe 3.24–La fonction moyenne des excès de RC.



Interprétation :

La linéarité du cout de sinistre pour la garantie **VI** commence à 800 000 et au-delà du seuil de 2 000 000 cette linéarité n'est plus assurée. Le graphe VI nous montre un seuil à partir duquel on considère que les coûts sont sérieux. On peut donc dire que le seuil sera dans l'intervalle g = [800 000 ; 2 000 000].

Graphe 3.25 –La fonction moyenne des excès de VI

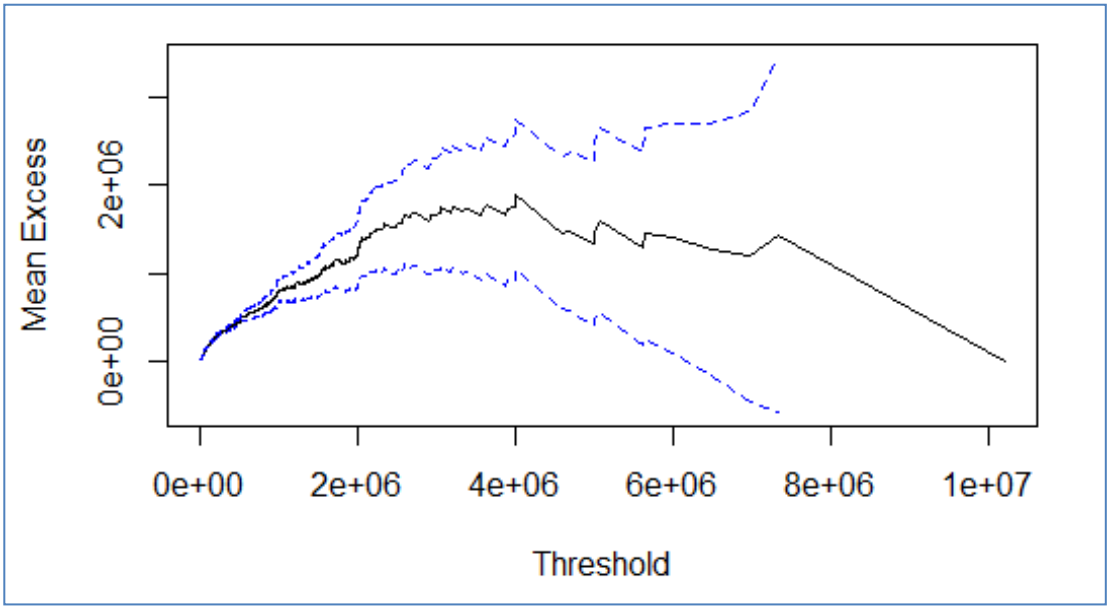
Remarque :

Cette méthode graphique ne nous donne pas beaucoup d'information sur la valeur du seuil.

3.7.2. Stabilité des coefficients

- **Stabilité des coefficients pour toutes les garanties**

Les coefficients sont estimés pour plusieurs valeurs de seuil. Nous étudions la stabilité du paramètre d'échelle modifié défini par : $\sigma^* = \sigma_u - \xi^*$.

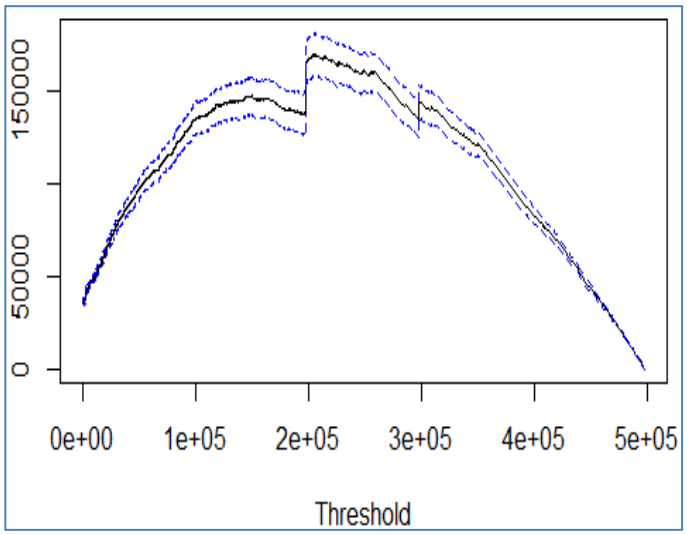


Graphe 3.26– Estimations du coefficient d'échelle modifié d'une loi de Pareto généralisée en Fonction de la valeur du seuil (pour toutes les garanties).

Interprétation :

Le coefficient d'échelle modifié pour toutes les garanties (2017 /2020) est stable dans l'intervalle [2 500 000 ; 3 000 000], le seuil des valeurs extrêmes se situerait donc dans cet intervalle.

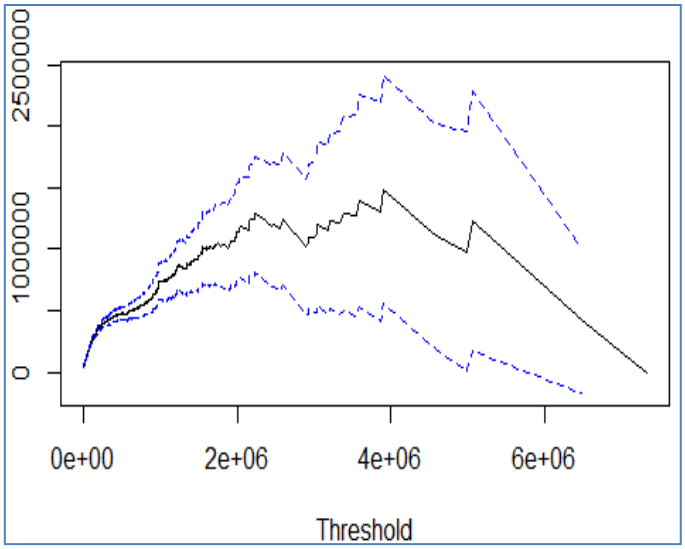
- **Stabilité des coefficients pour chaque garantie**



Interprétation :

Le coefficient d'échelle modifié de **DASC** est stable dans l'intervalle [150 000 ; 200 000], le seuil des valeurs extrêmes se situerait donc dans cet intervalle.

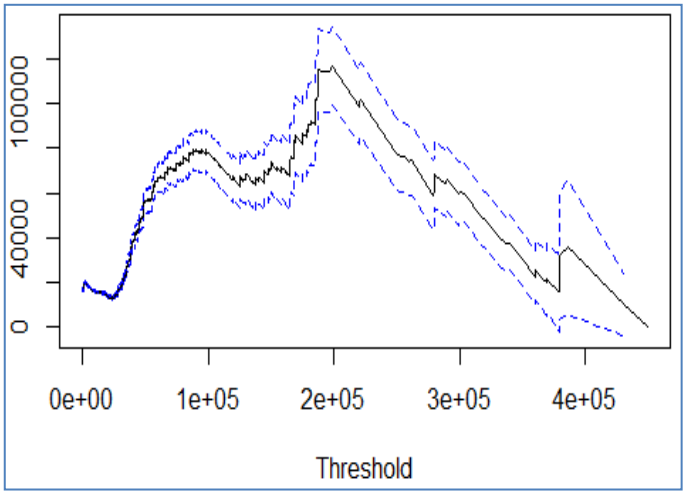
Graphe 3.27– Estimations du coefficient d'échelle modifié d'une loi de Pareto généralisée en Fonction de la valeur du seuil de DASC.



Interprétation :

Le coefficient d'échelle modifié de **TR** est stable dans l'intervalle [1 800 000 ; 1900 000], le seuil des valeurs extrêmes se situerait donc dans cet intervalle.

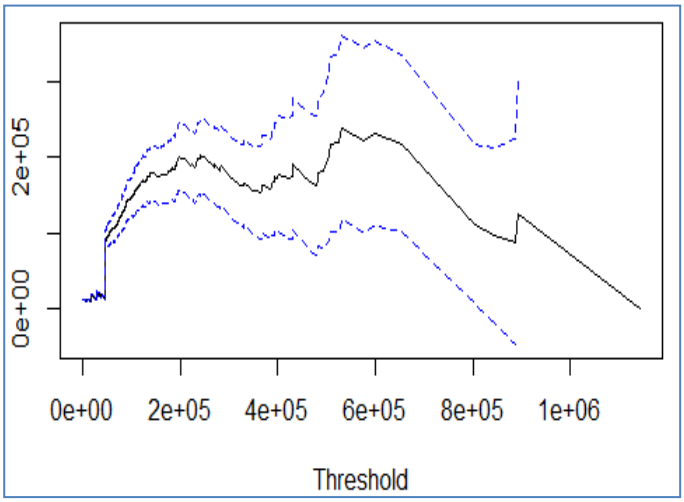
Graphe 3.28– Estimations du coefficient d'échelle modifié d'une loi de Pareto généralisée en Fonction de la valeur du seuil de TR.



Interprétation :

Le coefficient d'échelle modifié de **BDG** est stable dans l'intervalle [130 000 ; 160 000], le seuil des valeurs extrêmes se situerait donc dans cet intervalle.

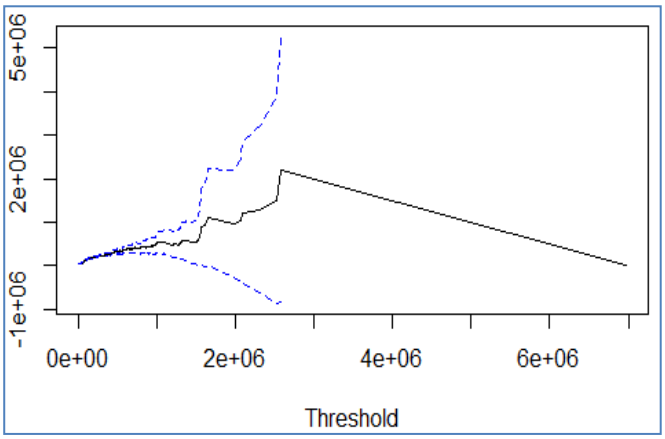
Graphe 3.29– Estimations du coefficient d'échelle modifié d'une loi de Pareto généralisée en Fonction de la valeur du seuil de BDG.



Interprétation :

Le coefficient d'échelle modifié de **DC** est stable dans l'intervalle [150 000 ; 200 000], le seuil des valeurs extrêmes se situerait donc dans cet intervalle.

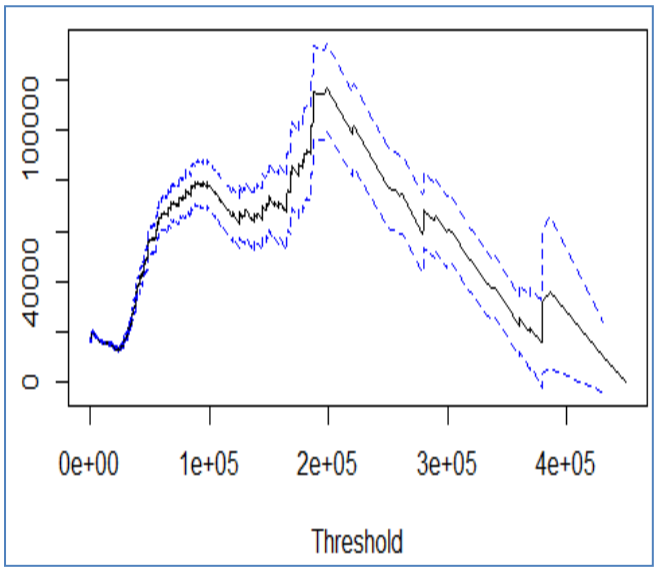
Graphe 3.30– Estimations du coefficient d'échelle modifié d'une loi de Pareto généralisée en Fonction de la valeur du seuil de DC.



Interprétation :

Le coefficient d'échelle modifié de **RC** est stable dans l'intervalle [200 000 ; 400 000], le seuil des valeurs extrêmes se situerait donc dans cet intervalle.

Graphe 3.31– Estimations du coefficient d'échelle modifié d'une loi de Pareto généralisée en Fonction de la valeur du seuil de RC.



Interprétation :

Le coefficient d'échelle modifié de **VI** est stable dans l'intervalle [1000 000 ; 2000 000], le seuil des valeurs extrêmes se situerait donc dans cet intervalle.

Graphe 3.32– Estimations du coefficient d'échelle modifié d'une loi de Pareto généralisée en Fonction de la valeur du seuil de VI.

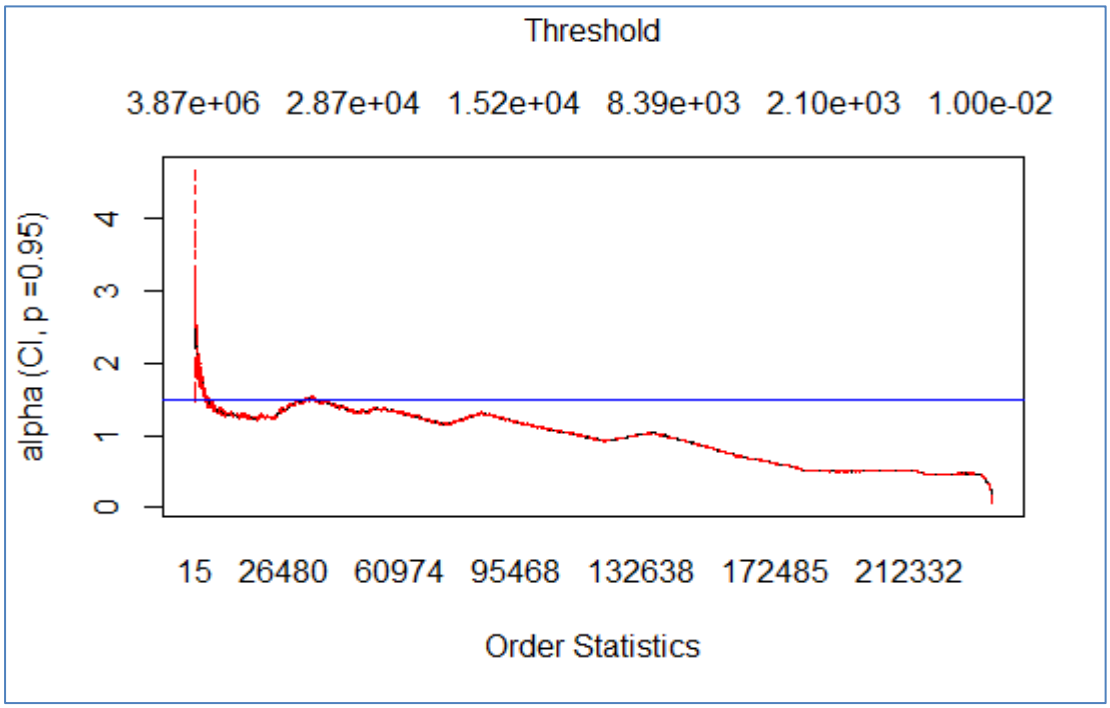
Remarque :

On remarque que par rapport à la méthode de la fonction moyenne des excès les intervalles sont réduits, il reste encore trop étendu pour choisir le seuil de manière précise pour cela on va réduit encore les intervalles par l'estimateur de Hill.

3.7.3. Hill-plot

Le graphe Hill-plot nous permet d'avoir des estimations du paramètre α en fonction de la statistique. En observant ce paramètre en fonction du seuil, on remarque qu'en dessous d'un certain seuil, la stabilité de ce paramètre est atteinte, c'est-à-dire la représentation de la queue de la répartition des sinistres graves à partir de ce seuil est stable.

- **Hill plot pour toutes les garanties:**



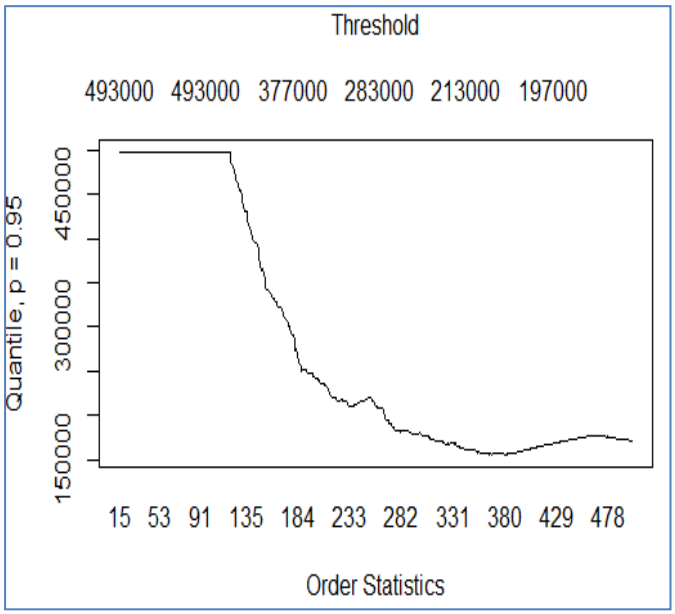
Graphe 3.33– Hill Plot du cout de sinistre (pour toutes les garanties)

Interprétation :

L'allure de la courbe de l'estimateur de Hill semble horizontale dans L'intervalle [2900 000 ; 2950 000] qui est donc réduit.

Nous allons donc fixer le seuil à 2 925 000 qui est la valeur la plus proche des estimations basées sur la méthode de la stabilité des coefficients et de l'estimateur de Hill.

- **Hill plot pour chaque Garantie :**

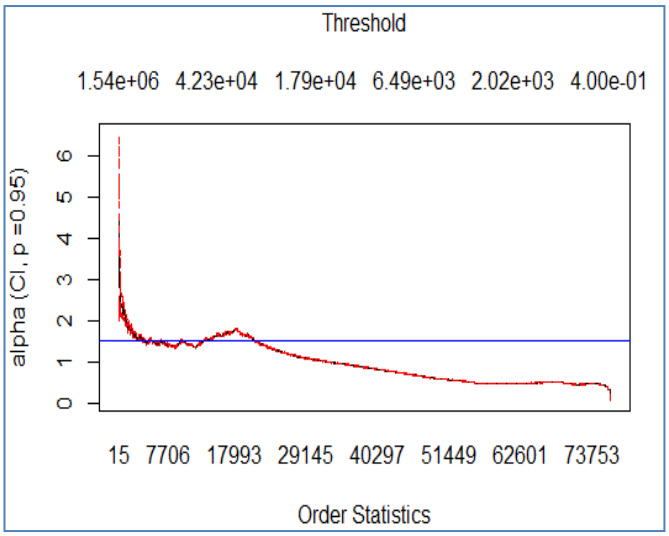


Interprétation :

L'allure de la courbe de l'estimateur de Hill semble horizontale dans L'intervalle [190 000 ; 200 000] qui est donc réduit

Nous allons donc fixer le seuil à 195 000 qui est la valeur la plus proche des estimations basées sur la méthode de la stabilité des coefficients et de l'estimateur de Hill.

Graphe 3.34– Hill Plot du cout de sinistre de DASC.

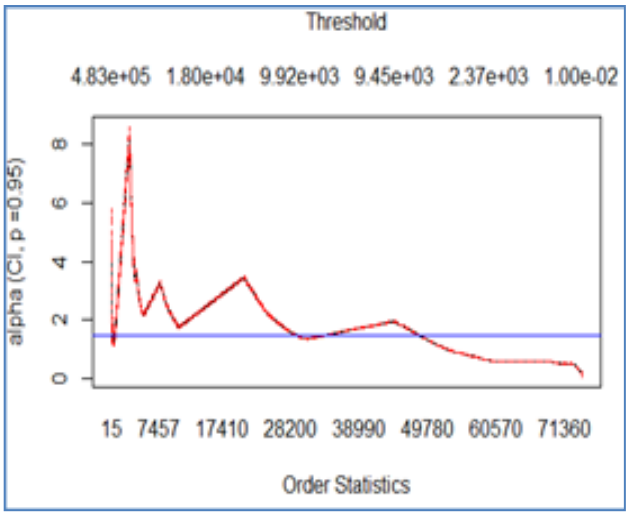


Interprétation :

L'allure de la courbe de l'estimateur de Hill semble horizontale dans L'intervalle [1650 000 ; 1700 000] qui est donc réduit.

Nous allons donc fixer le seuil à 1680000qui est la valeur la plus proche des estimations basées sur la méthode de la stabilité des coefficients et de l'estimateur de Hill.

Graphe 3.35– Hill Plot du cout de sinistre de TR.

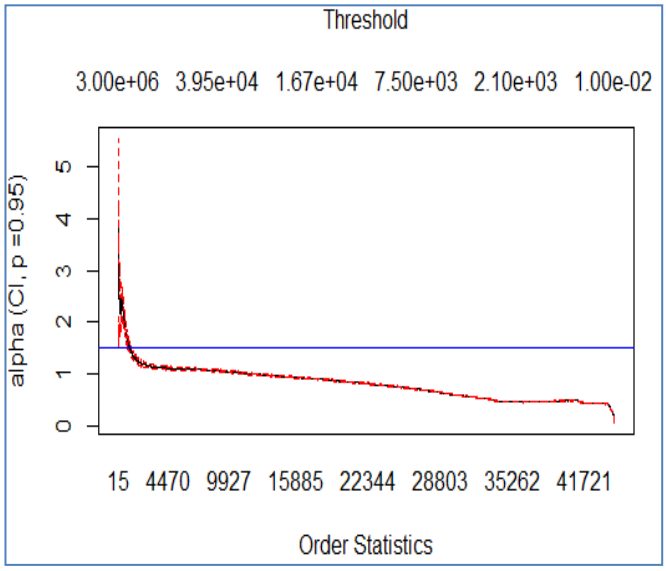


Interprétation :

L'allure de la courbe de l'estimateur de Hill semble horizontale dans L'intervalle [9 500 ; 10 000] qui est donc réduit.

Nous allons donc fixer le seuil à 98000qui est la valeur la plus proche des estimations basées sur la méthode de la stabilité des coefficients et de l'estimateur de Hill.

Graphe 3.36– Hill Plot du cout de sinistre de BDG.

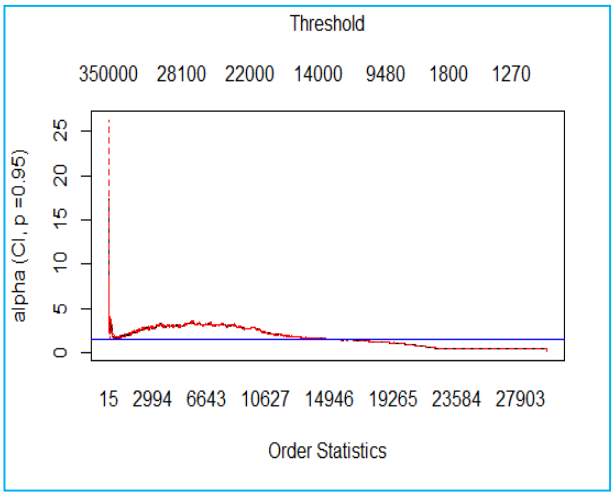


Interprétation :

L'allure de la courbe de l'estimateur de Hill semble horizontale dans L'intervalle [40 000 ; 50 000] qui est donc réduit

Nous allons donc fixer le seuil à 45000qui est la valeur la plus proche des estimations basées sur la méthode de la stabilité des coefficients et de l'estimateur de Hill.

Graphe 3.37– Hill Plot du cout de sinistre de DC.

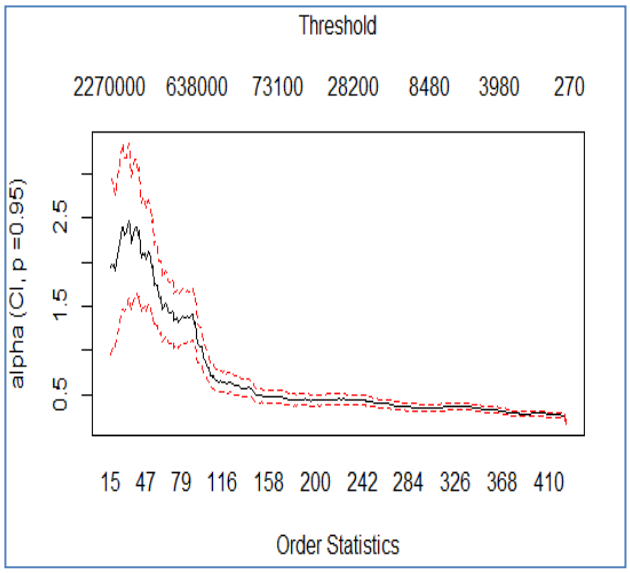


Interprétation :

L'allure de la courbe de l'estimateur de Hill semble horizontale dans L'intervalle [80 000 ; 100 000] qui est donc réduit

Nous allons donc fixer le seuil à 90000qui est la valeur la plus proche des estimations basées sur la méthode de la stabilité des coefficients et de l'estimateur de Hill.

Graphe 3.38– Hill Plot du cout de sinistre de RC.



Interprétation :

L'allure de la courbe de l'estimateur de Hill semble horizontale dans L'intervalle [1700 000 ; 1800 000] qui est donc réduit.

Nous allons donc fixer le seuil à 1750000qui est la valeur la plus proche des estimations basées sur la méthode de la stabilité des coefficients et de l'estimateur de Hill.

Graphe 3.39– Hill Plot du cout de sinistre de VI.

Conclusion :

La détermination du seuil d'écrêtement des sinistres est très délicate surtout que nous travaillons sur des sinistres d'assurance automobile. Les coûts extrêmes sont donc moins élevés que pour des sinistres de type catastrophes naturelles pour lesquels cette théorie est régulièrement utilisée. De plus la modélisation des sinistres élevés par une loi de Pareto généralisée présuppose que nous nous plaçons dans le domaine de convergence des valeurs extrêmes et donc le seuil est correctement évalué.

3.7.4. Choix des distributions basé sur la méthode d'estimateur du maximum de vraisemblance (EMV)

Le tableau ci-dessus représente quelques distributions avec leur (moyenne, variance et écart type) basé sur l'estimateur du maximum de vraisemblance (EMV).

DISTRINUTION	Méthode d'estimation	Moyenne	Variance	Ecart type
Weibull	EMV	34918.14	1 219 276 641	34918.14
Log-Normale	EMV	9731.312	3.928 658	1.982084
Pareto II	EMV	64639.22	4 178 228 148	64639.22
Pareto Généralisée	EMV	23031.5	529 892 895	23019.4

Tableau 3.13– Moyenne, var et l'écart type des distributions.

Commentaire :

D'après le tableau (3.13) on peut dire que la distribution de Pareto généralisée et log normal semble mieux adaptés à nos données (car les valeurs des écarts types sont inférieures aux autres).

3.7.5. Test de Qualité d'ajustement des modèles

Score	Distribution	Estimations	Rejet ou non Rejet de H0	Statistique	P-value
1	**Pareto Généralisée**	EMV	Non rejet	467.8239157	0.1143383
2	**Log-Normale**	EMV	Non rejet	241.0665781	0.0774051
3	**Weibull**	EMV	Non rejet	174.7533709	0.0733026
4	**Pareto**	EMV	Rejet	0.1072963	0.0001

Tableau 3.14 – Test de Qualité d'ajustement des modèles.

Commentaire :

Dans le tableau (3.14) On voie que La distribution Pareto généralisée avec les estimateurs du maximum de vraisemblance (EMV) a la valeur de p supérieure à celles des autres distributions. Cela signifie que la probabilité de se tromper en rejetant l'hypothèse H0 est la plus élevée.

On accepte les trois premiers modèles Pareto généralisée, log-normale et weibull car les P-values sont supérieur à 0.05 et on rejet Pareto (p-v=0.0001<0.05).

Conclusion :

D'après le tableau La distribution de Pareto Généralisée et Log-Normal sont cependant préférables à la distribution de weibull car ils ont les p-values plus grandes que celle de weibull.

3.7.6. Estimations des paramètres

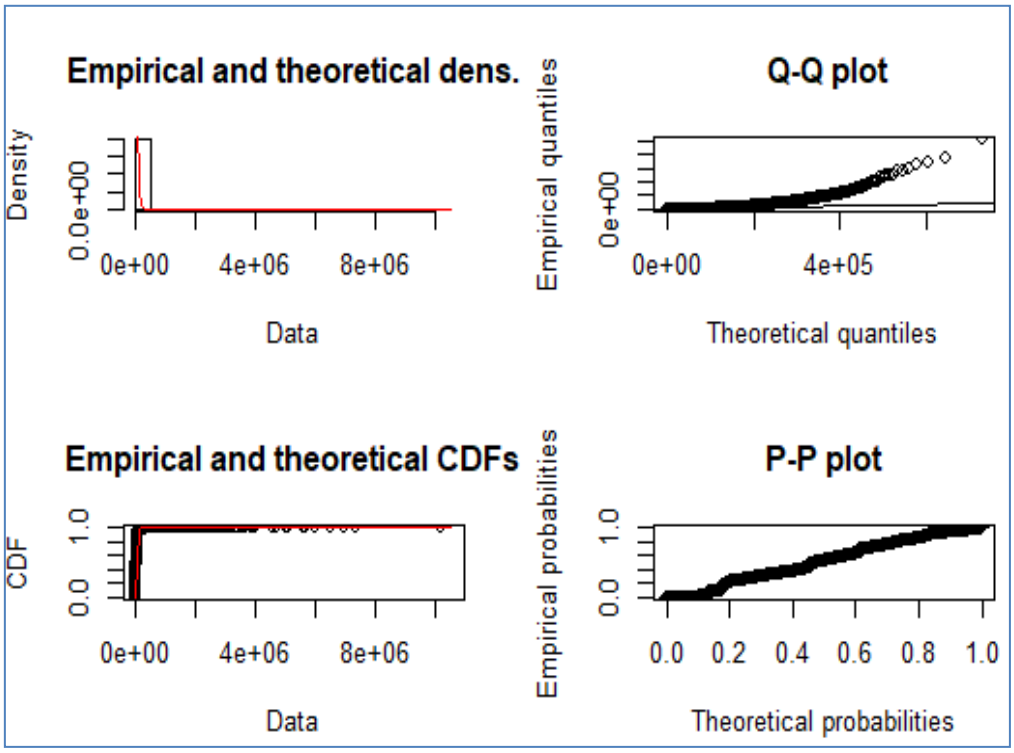
Distribution	Estimation	Parameter	Value	Parameter	Value
Weibull	EMV	Scale	1.92e+04	Shape	7.07e-01
Log-normal	EMV	Mu	24815.32	Sigma	1.368
Generlized Pareto	EMV	Scale	8.258e+05	Shape	1.343e-01

Tableau 3.15 – Estimations des paramètres pour les différentes lois.

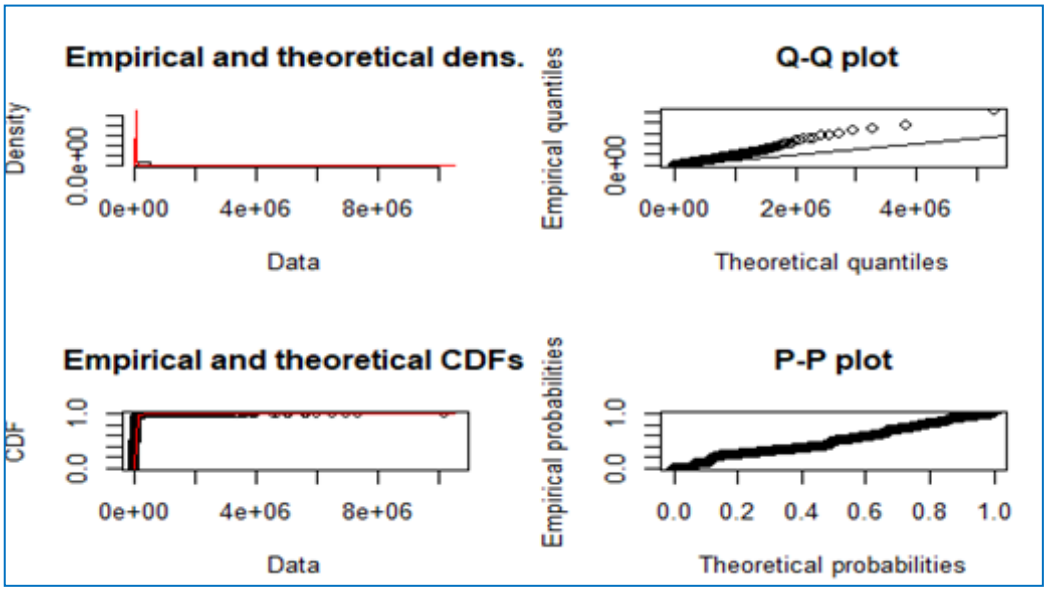
Commentaire :

Les résultats du tableau (3.15) montrent les paramètres estimés par La méthode d'estimateur du maximum de vraisemblance (EMV).

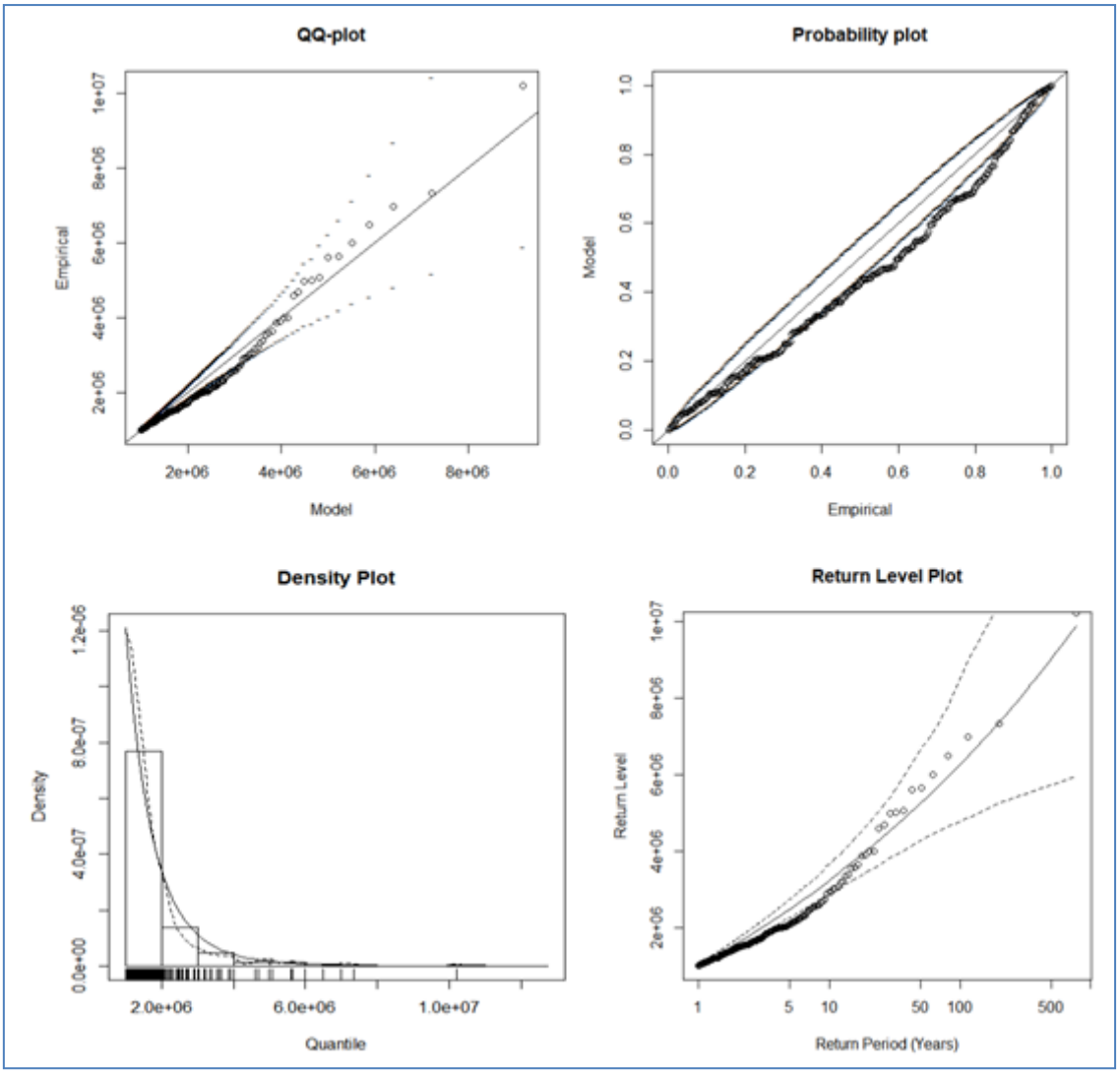
Représentation graphique :



Graphe 3.40 –Distribution de weibull.



Graphe 3.41 –Log-normal Distribution.



Graphe 3.42 –Distribution de Pareto généralisée.

Interprétation :

D'après les graphes (**3.42, 3.41 et 3.40**) d'une part nous constatons que les trois distributions semblent correspondre à notre distribution ce qui confirme les résultats de tableau **3.14** et d'autre part on remarque que les points qui représentent nos données sont mieux alignés avec Pareto généralisée.

Conclusion :

La distribution de **Pareto généralisée** est mieux correspondre à nos données.

Cette recherche de seuil est réalisée pour chaque garantie, voici le tableau récapitulatif :

Garanties	Seuils	Au-dessous de seuil
Tous	2000000	0.9988512
RC	90000	0.9963558
BDG	98000	0.9297707
DC	45000	0.9469449
TR	1680000	0.9495366
DASC	195000	0.8277835
VI	1750000	0.9482353

Tableau 3.16– Les seuils par la méthode des valeurs extrêmes

Remarque :

Les résultats sont cohérents : les garanties VI et TR ont des seuils assez proches et assez grands par rapports aux autres seuils des autres garanties cela veut dire que les couts de sinistre de VI et TR sont les plus intéressants.

Conclusion générale

Conclusion générale

Dans cette étude, nous avons compris l'importance de modéliser les coûts des sinistres en assurance. Après avoir constaté le poids et l'impact des sinistres causés, nous avons décidé d'axer notre problématique sur le prix en compte de la sinistralité extrême en assurance automobile. Pour cela, nous avons considéré plusieurs modèles mathématiques que nous avons appliqués à nos données.

Comme une première étape de modélisation nous avons commencé par la méthode linéaire généralisée en utilisant la distribution de la loi binomiale négative basé sur la fonction de la vraisemblance, après cette étape nous avons passé par la méthode régression quantile, pour savoir L'impact des différents types des garanties sur les couts du sinistre et après nous avons comparé nos données avec la distribution exponentielle en se servant de la fonction

qq-plot.

Le principe est différent pour la méthode utilisant la théorie des valeurs extrêmes qui était basé sur trois étapes (fonction moyenne des excès ensuite l'estimation de coefficients d'échelle basée sur La distribution de la loi Pareto généralisée et enfin l'estimateur de Hill) pour cette dernière un unique seuil est déterminé pour chaque garantie pendant toutes les exercices.

On espère que cette étude ajoutera un plus dans le domaine assurantiel en Algérie, et qu'elle sera la base pour d'autres études plus approfondies.

Annexe

Annexes

A1 : Programme des densités de la distribution des lois de valeurs extrêmes (chapitre 2):

```
# Density of the GEV distribution "DensitÈ des Lois de Valeurs ExtrÍmes"
op= par(mfrow=c(1,2), mar=c(3,2,4,2)+.1)
library(evd)
curve(dfrechet(x,shape=1),xlim=c(-4,4),ylim=c(0,1),ylab="",col='red',main="DensitÈ",lty=2)
curve(dgumbel(x),add=T,lty=1)
curve(drweibull(x,shape=1),col="blue",add=T, lty=4)
legend("topleft", c("Frechet","Gumbel","Weibull"),pt.bg="white", lty=c(2,1,4),col = c('red','black','blue')
```

# Density of the GEV distribution "DensitÈ des Lois de Valeurs ExtrÍmes"
op= par(mfrow=c(1,2), mar=c(3,2,4,2)+.1)
library(evd)
curve(dfrechet(x,shape=1),xlim=c(4,4),ylim=c(0,1),ylab="",col='red',main="DensitÈ",lty=2)
curve(dgumbel(x),add=T,lty=1)
curve(drweibull(x,shape=1),col="blue",add=T, lty=4)
legend("topleft", c("Frechet","Gumbel","Weibull"),pt.bg="white", lty=c(2,1,4),col =
c('red','black','blue')

A2 : les donnes de la société SAA :

Les donnes initiales de la société SAA

Les données de 2017 :

	1	2	3	4	5	6	7	8	9	10	11	12
1	ID	code_agence	Nom_agence	code_DR	Nom_DR	type_agence	produit	branche	Nume_police	garantie	cout de sinistre	
2	1206-11000	1206	THENIA	20	Direction Ré	Agence direc	Automobile I	AUTO	1,1E+09	RC	34885,07	
3	1206-11000	1206	THENIA	20	Direction Ré	Agence direc	Automobile I	AUTO	1,1E+09	RC	1150	
4	1206-11000	1206	THENIA	20	Direction Ré	Agence direc	Automobile I	AUTO	1,1E+09	RC	1820	
5	1206-11000	1206	THENIA	20	Direction Ré	Agence direc	Automobile I	AUTO	1,1E+09	BDG	5018,8	
6	1206-11000	1206	THENIA	20	Direction Ré	Agence direc	Automobile I	AUTO	1,1E+09	RC	1900	
7	1206-11000	1206	THENIA	20	Direction Ré	Agence direc	Automobile I	AUTO	1,1E+09	RC	3450	
8	1206-11000	1206	THENIA	20	Direction Ré	Agence direc	Automobile I	AUTO	1,1E+09	BDG	25500	
9	1206-11000	1206	THENIA	20	Direction Ré	Agence direc	Automobile I	AUTO	1,1E+09	DC	6184,32	
10	1206-11000	1206	THENIA	20	Direction Ré	Agence direc	Automobile I	AUTO	1,1E+09	RC	1780	
11	1206-11000	1206	THENIA	20	Direction Ré	Agence direc	Automobile I	AUTO	1,1E+09	RC	1780	

Les données de 2018 :

	ID	code_agen	Nom_agen	code_DR	Nom_DR	type_agenc	produit	branche	Nume_poli	garantie	cout de sinistre
2	1206-11000	1206	THENIA	20	Direction R	Agence dire	Automobile	AUTO	1,1E+09	TR	30460
3	1206-11000	1206	THENIA	20	Direction R	Agence dire	Automobile	AUTO	1,1E+09	TR	18500
4	1206-11000	1206	THENIA	20	Direction R	Agence dire	Automobile	AUTO	1,1E+09	DC	9237,8
5	1206-11000	1206	THENIA	20	Direction R	Agence dire	Automobile	AUTO	1,1E+09	DC	20700
6	1206-11000	1206	THENIA	20	Direction R	Agence dire	Automobile	AUTO	1,1E+09	TR	21500
7	1206-11000	1206	THENIA	20	Direction R	Agence dire	Automobile	AUTO	1,1E+09	DC	21240
8	1206-11000	1206	THENIA	20	Direction R	Agence dire	Automobile	AUTO	1,1E+09	DC	9500
9	1206-11000	1206	THENIA	20	Direction R	Agence dire	Automobile	AUTO	1,1E+09	DASC	8500
10	1206-11000	1206	THENIA	20	Direction R	Agence dire	Automobile	AUTO	1,1E+09	TR	16500
11	1206-11000	1206	THENIA	20	Direction R	Agence dire	Automobile	AUTO	1,1E+09	DC	26251,26
12	1206-11000	1206	THENIA	20	Direction R	Agence dire	Automobile	AUTO	1,1E+09	TR	6500

Les données de 2019 :

	ID	code_agence	Nom_agence	code_DR	Nom_DR	type_agence	produit	branche	Nume_police	garantie	cout de sinistre
2	1206-11000	1206	THENIA	20	Direction Ré	Agence direc	Automobile f	AUTO	1,1E+09	RC	0
3	1206-11000	1206	THENIA	20	Direction Ré	Agence direc	Automobile f	AUTO	1,1E+09	DC	9500
4	1206-11000	1206	THENIA	20	Direction Ré	Agence direc	Automobile f	AUTO	1,1E+09	DC	35167,27
5	1206-11000	1206	THENIA	20	Direction Ré	Agence direc	Automobile f	AUTO	1,1E+09	DC	24619,5
6	1206-11000	1206	THENIA	20	Direction Ré	Agence direc	Automobile f	AUTO	1,1E+09	DC	18067,5
7	1206-11000	1206	THENIA	20	Direction Ré	Agence direc	Automobile f	AUTO	1,1E+09	DC	19575
8	1206-11000	1206	THENIA	20	Direction Ré	Agence direc	Automobile f	AUTO	1,1E+09	RC	0
9	1206-11000	1206	THENIA	20	Direction Ré	Agence direc	Automobile f	AUTO	1,1E+09	TR	4500
10	1206-11000	1206	THENIA	20	Direction Ré	Agence direc	Automobile f	AUTO	1,1E+09	DASC	23500
11	1206-11000	1206	THENIA	20	Direction Ré	Agence direc	Automobile f	AUTO	1,1E+09	DC	18000

Les données de 2020 :

	ID	code_agence	Nom_agence	code_DR	Nom_DR	type_agence	produit	branche	Nume_police	garantie	cout de sinistre
	1206-11000	1206	THENIA	20	Direction Ré	Agence direc	Automobile f	AUTO	1,1E+09	TR	112669,87
	1206-11000	1206	THENIA	20	Direction Ré	Agence direc	Automobile f	AUTO	1,1E+09	RC	0
	1206-11000	1206	THENIA	20	Direction Ré	Agence direc	Automobile f	AUTO	1,1E+09	DASC	26000
	1206-11000	1206	THENIA	20	Direction Ré	Agence direc	Automobile f	AUTO	1,1E+09	DC	12700,63
	1206-11000	1206	THENIA	20	Direction Ré	Agence direc	Automobile f	AUTO	1,1E+09	DASC	411747,74
	1206-11000	1206	THENIA	20	Direction Ré	Agence direc	Automobile f	AUTO	1,1E+09	TR	352125,35
	1206-11000	1206	THENIA	20	Direction Ré	Agence direc	Automobile f	AUTO	1,1E+09	TR	551736,33
	1206-11000	1206	THENIA	20	Direction Ré	Agence direc	Automobile f	AUTO	1,1E+09	DC	18000
10	1206-11000	1206	THENIA	20	Direction Ré	Agence direc	Automobile f	AUTO	1,1E+09	RC	4170

Les donnes après traitement :

Les données après traitement pour analyse descriptive :

	année	couts des sinistres de BDG
2	2017	8000
3	2017	8020
11361	2018	26000
11362	2018	7500
16365	2019	29000
16366	2019	29400
23818	2020	188000
23819	2020	196300,1

BDG

	années	couts des sinistres de DASC
2	2017	77173,11
3	2017	13640,7
1846	2018	7247,9
1847	2018	46252,1
1848	2018	7809
7303	2019	32700
7304	2019	1860
7305	2020	23500
7306	2020	19515

DASC

	Annee	couts des sinistres de DC
2	2017	6184,32
3	2017	18753,47
41162	2018	7000
41163	2018	8064,8
41164	2019	12700,63
41165	2019	411747,74
110033	2020	17268,57
110034	2020	18000

DC

	Années	couts de sinistre de RC
2	2017	34885,07
3	2017	1150
43156	2018	26151,2
43157	2018	13607,1
43158	2019	4170
43159	2019	2180
73301	2020	66452,26
73302	2020	54307,92

RC

	Années	couts de sinistre de VI
2	2017	185773,1
3	2017	174975,26
239	2018	1860
240	2018	37421,85
241	2019	165582,84
242	2019	1991067
369	2020	34031,99
370	2020	981500

VI

	Années	couts de sinistre de TR
2	2017	20715,1
3	2017	6000
25053	2018	3770
25054	2018	84177,18
25055	2019	112669,87
25056	2019	352125,35
37036	2020	43306,89
37037	2020	63075,25

TR

Les donnes après traitement pour appliquer un programme sur les trois méthodes:

1	Date	garantie4	cout de sinis4e
2	2017	1	34885.07
3	2017	1	1150
4	2017	1	1820
5	2017	2	5018.8
6	2017	1	1900
7	2017	1	3450
8	2017	2	25500
9	2017	3	6184.32
10	2017	1	1780
11	2017	1	1780
12	2017	1	1940
13	2017	2	16949.46
14	2017	3	18753.47
15	2017	3	27000
16	2017	1	1780
17	2017	3	45000
18	2017	3	22050
19	2017	4	20715.1
20	2017	3	13050
21	2017	3	4610
22	2017	1	500
23	2017	1	1550
24	2017	4	6000
25	2017	1	1450.36
26	2017	4	23137.78
27	2017	3	11049.54
28	2017	3	9873.82
29	2017	3	8000

89970	2018	4	3170
89971	2018	3	2020
89972	2018	3	1820
89973	2018	3	1900
89974	2018	5	2500
89975	2018	4	2220
89976	2018	2	9500
89977	2018	2	32000
89978	2018	2	11000
89979	2018	2	28000
89980	2018	2	9500
89981	2018	2	20000
89982	2018	2	28000
89983	2018	2	27000
89984	2018	2	28000
89985	2018	2	9000
89986	2018	4	3080.08
89987	2018	3	9500
89988	2018	5	2460
89989	2018	5	54939.97
89990	2018	4	2300
89991	2018	3	3490
89992	2018	3	2220
89993	2018	3	1270
89994	2018	4	1820
89995	2018	3	9500
89996	2018	1	31471.67
89997	2018	3	18000
89998	2018	3	7500

170041	2019	2	21500
170042	2019	2	9800
170043	2019	2	45000
170044	2019	3	27000
170045	2019	4	6500
170046	2019	4	14348,6
170047	2019	4	96698,97
170048	2019	5	40274,05
170049	2019	3	9500
170050	2019	5	97850
170051	2019	4	32946,4
170052	2019	3	4000
170053	2019	3	27000
170054	2019	3	27000
170055	2019	4	10372,5
170056	2019	2	101687,34
170057	2019	2	24500
170058	2019	2	12000
170059	2019	2	28000
170060	2019	2	34000
170061	2019	2	31000
170062	2019	2	32000
170063	2019	2	26000
170064	2019	2	24500
170065	2019	2	18000
170066	2019	2	14500
170067	2019	2	26500

204651	2020	1	19100
204652	2020	4	9500
204653	2020	5	23400
204654	2020	6	2220
204655	2020	1	2420
204656	2020	1	3290
204657	2020	2	1860
204658	2020	4	1980
204659	2020	5	3250
204660	2020	6	2220
204661	2020	4	1980
204662	2020	6	3450
204663	2020	6	1900
204664	2020	6	1860
204665	2020	5	2300
204666	2020	3	1980
204667	2020	4	1900
204668	2020	3	1150
204669	2020	3	1270
204670	2020	3	1900
204671	2020	3	3250
204672	2020	3	1980
204673	2020	3	2100
204674	2020	4	1940
204675	2020	4	1350
204676	2020	3	2340
204677	2020	4	2480

A3 : programme sous R

```r
#packages
library(readxl)
library(VGAM)
library(evir)
library(ReIns)
library(EnvStats)
library(POT)
library(eva)
library(quantreg)
library(fdaoutlier)
library(extRemes)
library(goft)
library(MASS)
data1<-read.csv(file.choose(),header=TRUE)
data2017<-read.csv(file.choose(),header=TRUE)
data2018<-read.csv(file.choose(),header=TRUE)
data2019<-read.csv(file.choose(),header=TRUE)
data2020<-read.csv(file.choose(),header=TRUE)

guarantee 1
g1<-subset(data1, data1$garantie4 == 1)
guarantee 2
g2<-subset(data1, data1$garantie4 == 2)
guarantee 2
g3<-subset(data1, data1$garantie4 == 3)
guarantee 2
g4<-subset(data1, data1$garantie4 == 4)
guarantee 2
g5<-subset(data1, data1$garantie4 == 5)
guarantee 6
g6<-subset(data1, data1$garantie4 == 2)

```

```r
#analyse discriptive
#BDG
BDG <- subset(data1,data1$garantie4==2)
summary(BDG)
sd(BDG$cout.de.sinis4e)
length(BDG$cout.de.sinis4e)
boxplot(BDG,col=c("yellow"),main="boxplot de BDG ",ylabl="cout sinistre")
plot(BDG ,pch="+",col=c("blue"))
#DASC
DASC <- subset(data1,data1$garantie4==5)
summary(DASC)
sd(DASC$cout.de.sinis4e)
length(DASC$cout.de.sinis4e)
boxplot(DASC,col=c("yellow"),main="boxplot de de DASC ",ylabl="cout sinistre")
plot(DASC ,pch="+",col=c("blue"))
#DC
DC<- subset(data1,data1$garantie4==3)
summary(DC)
sd(DC$cout.de.sinis4e)
length(DC$cout.de.sinis4e)
boxplot(DC,col=c("yellow"),main="boxplot de DC ",ylabl="cout sinistre")
plot(DC ,pch="+",col=c("blue"))
#RC
RC<- subset(data1,data1$garantie4==1)
summary(RC)
sd(RC$cout.de.sinis4e)
length(RC$cout.de.sinis4e)
boxplot(RC,col=c("yellow"),main="boxplot de RC",ylabl="cout sinistre")
plot(RC ,pch="+",col=c("blue"))
#TR
TR <- subset(data1,data1$garantie4==4)
summary(TR)
sd(TR$cout.de.sinis4e)
length(TR$cout.de.sinis4e)
boxplot(TR,col=c("yellow"),main="boxplot de TR",ylabl="cout sinistre")
plot(TR ,pch="+",col=c("blue"))
```

```r
#VI
VI <- subset(data1,data1$garantie4==6)
summary(VI)
sd(VI$cout.de.sinis4e)
length(VI$cout.de.sinis4e)
boxplot(VI,col=c("yellow"),main="boxplot de VI",ylabl="cout sinistre")
plot(VI ,pch="+",col=c("blue"))
#Analyse préliminaire

#Test de Spearman
cor.test(data1$garantie4, data1$cout.de.sinis4e, method=c("spearman"))
cor.test(data2017$garantie4, data1$cout.de.sinis4e, method=c("spearman"))
cor.test(data2018$garantie, data1$cout.de.sinistre, method=c("spearman"))
cor.test(data2019$garantie, data1$cout.de.sinistre, method=c("spearman"))
cor.test(data2020$garantie, data1$cout.de.sinistre, method=c("spearman"))
#QQPLOT
#qplot(data1$cout.de.sinis4e,xi=0)
emplot(data1$cout.de.sinis4e,'xy')
#Méthode linéaire généralisée
#Moyenne et variance du coût de sinistre2017 à 2020.
summary(data1$cout.de.sinis4e)
var(data1$cout.de.sinis4e)
sd(data1$cout.de.sinis4e)^2
#régression binomiale négative
myglm=glm(formula=data1$cout.de.sinis4e ~ data1$garantie4,data1$gcout.de.sinis4,
          family =negative.binomial(2),data = data1)
summary(myglm)
# Méthode regression quantil
qr<-rq(formula = data1$cout.de.sinis4e ~ data1$garantie4, tau = c(0.5,
                                    0.75, 0.9, 0.995), data = data1)
summary(qr)

#Méthode linéaire généralisée
#Moyenne et variance du coût de sinistre2017 à 2020.
  summary(data1$cout.de.sinis4e)
  var(data1$cout.de.sinis4e)
```

```r
  sd(data1$cout.de.sinis4e)^2
#régression binomiale négative
 myglm=glm(formula=data1$cout.de.sinis4e ~ data1$garantie4,data1$gcout.de.sinis4,
               family =negative.binomial(2),data = data1)
 summary(myglm)
# Méthode regression quantil

qr<-rq(formula = data1$cout.de.sinis4e ~ data1$garantie4, tau = c(0.5,
                                      0.75, 0.9, 0.995), data = data1)
summary(qr)
#Méthode valeur exetrmes
# abline(h=alpha,col="blue")
# abline(a =0, b = 0.75,col="red")
 fit<-gpd(data1$cout.de.sinis4e,1e6)
msplot(data1$cout.de.sinis4e)
meplot(data1$cout.de.sinis4e, omit = 3, labels = TRUE)
hill(data1$cout.de.sinis4e)
msplot(DASC$cout.de.sinis4e)
meplot(DASC$cout.de.sinis4e, omit = 3, labels = TRUE)
hill(DASC$cout.de.sinis4e)
msplot(TR$cout.de.sinis4e)
meplot(TR$cout.de.sinis4e, omit = 3, labels = TRUE)
hill(TR$cout.de.sinis4e)
msplot(BDG$cout.de.sinis4e)
meplot(BDG$cout.de.sinis4e, omit = 3, labels = TRUE)
hill(BDG$cout.de.sinis4e)
msplot(DC$cout.de.sinis4e)
meplot(DC$cout.de.sinis4e, omit = 3, labels = TRUE)
hill(DC$cout.de.sinis4e)
msplot(RC$cout.de.sinis4e)
meplot(RC$cout.de.sinis4e, omit = 3, labels = TRUE)
hill(RC$cout.de.sinis4e)
msplot(VI$cout.de.sinis4e)
meplot(VI$cout.de.sinis4e, omit = 3, labels = TRUE)
hill(VI$cout.de.sinis4e)
```

```r
#goodness fit of distribution
#weibull
eweibull(data1$cout.de.sinis4e,method = "mle")
 #
eqweibull(data1$cout.de.sinis4e, p =c(0.75,0.90,0.995), method = "mme", digits = 0)
 #l  ognormal
 fitw<-fitdist(data1$cout.de.sinis4e, "lnorm", method = "mle")
 plot(fitw)
 gofstat(fitw)
 #pareto
 fitg <- fitdist(data1$`cout de sinis4e`, "pareto")
  summary(fitg)
  #
 w <- rep(1, length(data1$`cout de sinis4e`))

  mmedist(data1$`cout de sinis4e`, "pareto", order=c(1, 2), memp=memp2,
              weights=w,start=list(shape=2.01058, scale=24944.16739), lower=1, upper=Inf)
 #generlized pareto
 egevd(data1$cout.de.sinis4e, method = "mle")
 gofTest(data1$cout.de.sinis4e, distribution = "gev",test = "ks")
 egevd(data1$cout.de.sinis4e, method = "mle")
 gofTest(data1$cout.de.sinis4e, distribution = "gev",test = "ks")
 model<-fit.GPD(data1$cout.de.sinis4e,1e6)
 model1<-fit.GPD(g1$cout.de.sinis4e,500000)
 model2<-fit.GPD(g2$cout.de.sinis4e,35000)
 model3<-fit.GPD(g3$cout.de.sinis4e,35000)
 model4<-fit.GPD(g4$cout.de.sinis4e,150000)
 model5<-fit.GPD(g5$cout.de.sinis4e,500000)
 mode6<-fit.GPD(g6$cout.de.sinis4e,2000000)
 #VARselect(data1$cout.de.sinis4e,lag.max = 12, type = "const")
 #bv.est <- VAR(data1, p = 2, type = "const", season = NULL,exog = NULL)
 bv.est
#teste de kolmogorov
ks.test(data1$garantie4,data1$cout.de.sinis4e)
ks.test(data2017$garantie4,data1$cout.de.sinis4e)
ks.test(data2018$garantie4,data1$cout.de.sinis4e)
ks.test(data2019$garantie4,data1$cout.de.sinis4e)
ks.test(data2020$garantie4,data1$cout.de.sinis4e)

```

A4 : Calcul de l'espérance et de la variance d'une loi de type exponentielle

Y suit une loi de la famille exponentiel :$f_{\emptyset,\theta} = \exp\left(\frac{y^{\theta}-b(\theta)}{a(\emptyset)} + c(y,\emptyset)\right)$

Posons $u(\theta,\emptyset) = \int_y f_{\emptyset,\theta}\,\partial v = 1$ car il s'agit d'une densité.

Calculons la dérivée partielle de u par rapport à θ :

$$\frac{\partial u(\theta,\emptyset)}{\partial\theta} = \lim_{h\to 0}\frac{u(\theta+h,\emptyset)-u(\theta,\emptyset)}{h}$$

$$= \lim_{h\to 0}\frac{\int_y(f_{\theta+h,\emptyset}-f_{\theta,\emptyset})}{h}$$

$$= \liminf_{h\to 0}\frac{\int_y(f_{\theta+h,\emptyset}-f_{\theta,\emptyset})}{h}$$

$$= \limsup_{h\to 0}\frac{\int_y(f_{\theta+h,\emptyset}-f_{\theta,\emptyset})}{h}$$

En appliquant le lemme de Fatou nous obtenons :

$$\liminf_{h\to 0}\frac{\int_y(f_{\theta+h,\emptyset}-f_{\theta,\emptyset})}{h}\partial v \le 0 \le= \limsup_{h\to 0}\frac{\int_y(f_{\theta+h,\emptyset}-f_{\theta,\emptyset})}{h}\partial v$$

Or la dérivée partielle de $f_{\theta,\emptyset}$ par rapport à θ existe et donc :

$$= \int_y \liminf_{h\to 0}\frac{(f_{\theta+h,\emptyset}-f_{\theta,\emptyset})}{h}\partial v \le 0 \le \int_y \limsup_{h\to 0}\frac{(f_{\theta+h,\emptyset}-f_{\theta,\emptyset})}{h} = \int_y \frac{\partial f_{\theta,\emptyset}}{\partial\theta}\partial v = 0$$

D'après l'expression de$f_{\theta,\emptyset}$:$\frac{\partial f_{\theta,\emptyset}}{\partial\theta} = \frac{y-b'(\theta)}{a(\emptyset)}f_{\theta,\emptyset}(y)$

Donc :$\frac{1}{a(\emptyset)}\left[\int y f_{\theta,\emptyset}(y)\partial v - b'(\theta)\int f_{\theta,\emptyset}(y)\partial v\right] = 0$

Or $\frac{1}{a(\emptyset)}$est non nul. Nous obtenons donc :

$$E(Y) = b'(\theta)$$

Calcul de la variance de Y (GLM)

En utilisant le même type de raisonnement que pour le calcul de l'espérance de Y nous obtenons :

$$\frac{\partial^2 u(\theta,\emptyset)}{\partial\theta^2} = 0 = \int \frac{\partial^2 f_{\theta,\emptyset}}{\partial\theta^2}\partial v$$

A partir de l'expression de$f_{\theta,\emptyset}$ nous avons :

$$\frac{\partial^2 f_{\theta,\emptyset}}{\partial\theta^2} = -\frac{b''(\theta)}{a(\emptyset)}\int f_{\theta,\emptyset}(y)\,\partial v + \frac{y - b'(\theta)}{a(\emptyset)}\frac{\partial f_{\theta,\emptyset}}{\partial\theta}$$

$$= -\frac{b^{"}(\theta)}{a(\emptyset)}\, f_{\theta,\emptyset}(y) + \frac{1}{a(\emptyset)^2}\left[\left(y - b'(\theta)\right)^2 f_{\theta,\emptyset}(y)\right]$$

Or d'après le résultat précédent : $E(Y) = b'(\theta)$ nous obtenons donc en intégrant:

$$-\frac{b^{"}(\theta)}{a(\emptyset)}\int f_{\theta,\emptyset}(y)\,\partial v \; + \frac{1}{a(\emptyset)^2}\int \left(y - E(Y)\right)^2 f_{\theta,\emptyset}(y)\,\partial v = 0$$

Soit :

VAR(Y)= $b^{"}(\theta)\, a(\emptyset)$

A5: Définition d'un modèle régulier

Soit $(P_{\theta,}\Theta)$ un modèle paramétrique. On note $f(x,\theta)$ la densité de P_θ relativement à la mesure dominante υ (mesure de comptage ou mesure de Lebesgue). Le modèle $(P_{\theta,}\Theta)$ est régulier si les quatre hypothèses suivantes sont satisfaites :

(H1) L'ensemble des paramètres Θ est un ouvert de R_d pour d fini et
$f(x,\theta > 0) \Leftrightarrow f(x,\theta') > 0, \forall\theta, \theta' \in \Theta$

(H2) Pour υ presque tout x , les fonctions $\theta \to f(x,\theta)$ et $\theta \to \log f(x,\theta)$ sont deux fois continûment dérivables sur Θ.

(H3) Pour tout $\theta^* \in \Theta$ il existe un ouvert $U_{\theta^*} \subseteq \Theta$ contenant θ^* et une fonction borélienne $\Delta(x)$
Tels que
$\|\nabla_\theta (\log f(x,\theta))\| \le \Delta(x)\|$ et $\| H_\theta (\log f(x,\theta))\| \le \Delta(x)$

Pour tout $\theta \in U_{\theta^*}$ et υ -presque tout $x \in X$, et

$$\int \Delta(x) sup f(x,\theta) dv(x) < \infty$$

(H4) La matrice $-E\left[H_\theta(\log f(x,\theta))\right]$ de taille $d * d$ est symétrique définie positive pour tout $\theta \in \Theta$

Bibliographie

Bibliographie

En plus des articles et ouvrages et sites que nous avons cités, au fur et à mesure, au niveau de chaque chapitre, nous avons utilisé aussi les articles et ouvrages et sites suivants:

Articles et Ouvrages

- A.J. Dobson, A.G. Barnett. An introduction to generalized linear models, third edition, Chapmann and Hall, 2008.
- Code des assurances d'Algérie.
- Jan Beirlant, Yuri Goegebeur, Jozef Teugels. Statistics of extreme Theory and application, John Wiley&Sons, 2004.
- L'argus de l'automobile, Juillet 2011 'magasine'.
- Noureddine Benlagha, Michel Grun-Réhomme, Olga Vasechko. Les sinistres extremes en assurance automobile : une nouvelle approche par la théorie des valeurs extrêmes, Revue MODULAD, Numero 39, 2009.
- P.M.E. Altham, Introduction to linear generalized modeling. Statistical Laboratory, University of Cambridge.
- Sverre Grevskott, Sten Sture. Using PROC GENMOD to find a fair house insurance rate for the Norvegian market, SAS global forum 2008.
- Roger Koenker. Quantile regression, Econometric society monographs, 2005.
- Roger Koenker, Kevin F.Hallock. Quantile regression an introduction.
- Roger Koenker, José A.F.Machado. Goodness of fit and related inference processes for quantile regression, Journal of the American Statistical Association, Vol. 94, No 448, 1999.

Sites internet

www.saa.dz

. [1] François couibault, (2011), constant eliashberg, « les grands principes de l'assurance », 10 ème éditions, p57.
[1] Jérôme Yeatman, (1998). Manuel international de l'assurance, édition : Economica, p1.
[1] Lambert Faivre, (1986), Y Droit des assurances, édition : Précis Dalloz, p12.
[1] François Ewald-Jean Hérné Lorenzi, (1997), Encyclopédie d'assurance, Economica, p432, 433.
[1] James Landel, Lexique des termes d'assurance, Éditions L'Argus de l'assurance, p472.
[1] <https://www.index-assurance.fr/dictionnaire/assureur>. Consulté le 20 août 2020.
[1] François Ewald-Jean Hérné Lorenzi, (1997), Encyclopédie d'assurance, Economica, p432, 433.
[1] Bekkicha Abdelaziz, (2016), Estimation de la prime d'assurance Automobile via le modèle
[1] SylvieC.jean.p, (2016), « manuel des l'assurance automobile »,5eme éditions, l'agrus, paris, p40.

[1] Les figures 1.5 et 1.6 :http://www.cna.dz.Le portail d'assurance en Algérie.

[1] www.saa.dz

[1] <https://www.uar.dz/assurances-automobiles>. Consulté le 20 août 2020

[1] Source : univ-paris8.fr

[1] Indépendantes et Identiquement Distribuée

[1] Generalized Extreme Value distribution