Sommaire

1 - Statistiques descriptives :

L'échantillon prélevé est composé de 100 000 observations d'un portefeuille d'une compagnie d'assurance. On s'intéresse ici uniquement à l'assurance automobile. Il existe 37 variables, divisées en 4 groupes de variables : ID (l'identification du véhicule et des conducteurs) , Policy (les informations concernant les termes du contrat d'assurance), Drivers (les informations concernant les conducteurs), Vehicles (les informations concernant le véhicule).

Nombre de sinistres	Fréquences	Pourcentage (%)	Pourcentage cumulé(%)
0	88187	88,187	88,187
1	10533	10,533	98,720
2	1155	1,155	99,875
3	106	0,106	99,981
4	15	0,015	99,996
5	3	0,003	99,999
6	1	0,001	100,000

Tableau 1: Répartition de la fréquence des sinistres

Le tableau ci-dessous représente la fréquence du nombre de sinistres. On remarque que 88% des conducteurs n'ont pas eu de sinistres cette année-là, et près de 99% de l'échantillon a eu au plus 2 sinistres durant cette année.

	Conducteur Primaire	Conducteur Secondaire
Femme	50,088	50,14042
Homme	49,912	49,85958

Tableau 2 : Répartition des assurés selon le sexe et le type de conducteur

Le tableau ci-dessus représente la part d'hommes et de femmes déclarés en tant que conducteurs primaires et secondaires. Bien que la tarification selon le sexe soit prohibée, il est cependant autorisé de prélever ces informations et de les traiter dans le cadre d'une étude. On remarque que la part d'hommes et la part de femmes déclarés en tant que conducteurs primaires et secondaires sont égales, avec un nombre de femmes légèrement supérieur pour les deux types de conducteurs.

	Min	Moyenne	Médiane	Q1	Q3	Max
Ancienneté du permis	0	32,42	32	23	41	73
Age du conducteur	19	54,7	54	43	65	103
Bonus-Malus	0,5	0,53	0,5	0,5	0,5	1,65
Age du véhicule	1	9,53	8	4	13	63
Puissance du véhicule	15	91,44	87	68	109	555

Tableau 3 : Statistiques descriptives des variables quantitatives

Le tableau ci-dessus propose quelques statistiques pour les informations suivantes : ancienneté du permis, âge du conducteur, bonus-malus, l'âge du véhicule et sa puissance. On remarque tout d'abord que l'âge moyen et l'âge médian se trouvent autour des 54 ans, ce qui est assez élevé, avec pour maximum, un voir des conducteurs âgés de 103 ans. En moyenne l'ancienneté des permis est de 32,42 ans avec un 3eme quartile à 41 ans. Une grande partie des assurés est donc assez jeune.

Type de Formule	Mini	Median1	Median2	Maxi
% des assurés	8,51	9,32	17,316	64,854

Tableau 4 : Répartition des assurés selon le type de formule

D’après le tableau ci-dessus, on voit que l’assureur propose 4 types de formules. Avec près de 65% des assurés ayant choisi la formule maxi, à savoir celle qui assure le plus un conducteur.

Les Annexes 1 et 2 représentent respectivement le nombre de sinistres en fonction de l’âge du véhicule et du conducteur. Tout d’abord, dans les deux graphiques, on ne peut étudier que les assurés ayant subi 0 ou 1 sinistre. En effet, le nombre de conducteurs ayant subi plus d’un sinistre est faible ce qui rend difficile la représentation graphique et donc l’étude de ces observations. Ces graphiques sont en concordance avec nos résultats précédents, car on voit que la majorité des assurés n’ont pas eu de sinistres cette année-là. On voit que les véhicules les plus jeunes déclarent plus de sinistres et que les jeunes ont une fréquence de sinistre plus élevés.

D’après les annexes 3 et 4 qui représentent la proportion du nombre de sinistres en fonction de l’âge du conducteur et du véhicules on apprend que : tous les assurés qui ont eu 5 sinistres ont un véhicule qui a moins de 10 ans et sont âgés de19 à 30 ans ou de 50 à 75 ans.

2 - Modélisation de la fréquence

Dans ce projet, nous essayons de déterminer la corrélation entre différents facteurs et la fréquence de sinistralité dans un portefeuille d'assurance, pour ensuite déterminer une prime pure que l'on proposera aux assurés. On calcule cette prime pure en multipliant la fréquence d'accident estimé au coût moyen d'un sinistre

Nous allons proposer deux modèles pour estimer la fréquence : le modèle Poissonien et le modèle Binomiale-négatif. On comparera ensuite l'indépendance des résidus et on fera une décision en fonction des critères décisionnels.

2.1 - La Régression Poissonienne:

Comme vu en classe on commence par utiliser la régression Poissonienne pour modéliser la fréquence. Ce modèle va nous permettre d'identifier la relation qui existe entre le nombre de sinistres et les variables de notre portefeuille. On a utilisé la procédure « stepwise » sous R pour obtenir le modèle le plus précis avec uniquement les variables significatives, le modèle qui minimise le critère AIC (Akaike Information Criteria). Les variables sélectionnées dans la régression Poisson sont : la formule choisie par l'assuré, la durée du contrat, l'âge du conducteur principal et l'ancienneté du véhicule. Si on se réfère à l'annexe **2.1 Coefficients estimés par la Régression de Poisson** on remarque que les coefficients associés aux différentes catégories d'âge sont tous négatifs synonyme d'une relation inverse entre l'âge et la variable d'étude, la fréquence d'accidents : plus je suis âgé moins je fais d'accidents. En calculant l'espérance et la variance on se rend compte le modèle est limité car il y'a un problème de sur-dispersion :

$$E(N) = 0.1324 < V(N)\ 0.14704$$

On effectue un test de significativité pour confirmer le problème de sur-dispersion. On testera : H0 : $Equi\text{-}dispersion$ $H1$: $Sur\text{-}dispersion$. La p-value est inférieure au seuil de significativité, ce qui indique la significativité du résultat qui confirme qu'on rejette H0 et qui confirme le problème de sur-dispersion. On décide donc de s'intéresser aux corrélations qui pourraient expliquer le problème de sur-dispersion (si elles sont assez fortes).

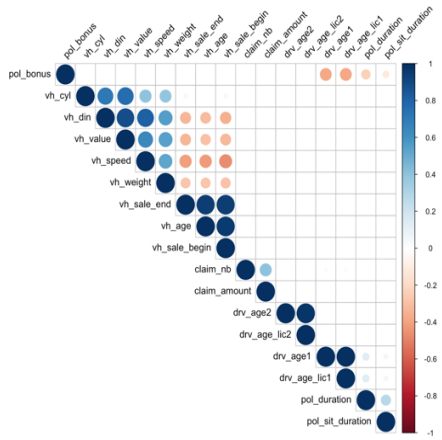


Figure 1 :Matrice des corrélations

On observe directement les fortes corrélations qu'il y'a entre les variables du cylindre de la dynamique, de la valeur et de la vitesse du véhicule. Ces dernières semblent logiques car elles décrivent la qualité (puissance) du véhicule, plus elles sont élevées plus le véhicule est puissant. On a aussi une corrélation forte entre l'âge du conducteur et l'ancienneté du permis qui s'explique par le fait que la plupart des assurés passent leur permis dès leurs majorités (18-19 ans). Enfin la relation qui est liée à l'observation qu'on a fait plus haut sur la régression, plus le conducteur est jeune plus la fréquence d'accident augmente, ce qui explique la relation négative entre le bonus-malus et les variables : âge du conducteur et ancienneté du conducteur dans le portefeuille.

Après ces observations, Le modèle de Poisson ne semble pas adapté pour modéliser la fréquence de nos sinistres, on risque de perdre en efficacité.

2.2- La régression binomiale négative :

Afin de palier à notre problème de sur dispersion on décide d'utiliser la loi binomiale négative. En effet cette dernière dispose d'un paramètre supplémentaire qu'on utilisera pour ajuster la variance sans considérer la moyenne. On gardera les mêmes variables explicatives que celle du modèle Poisson pour cette régression. Lors de cette régression, qui vise à expliquer la fréquence des sinistres, on a remarqué qu'une des catégories d'âge n'était pas significative (tranche de 46 à 50), 0.103997 > 0.100:

```
drv_age1.2(46,50]      -0.064950   0.039950  -1.626 0.103997
```

Il est important que chaque catégorie d'âge ait un coefficient explicite associé pour avoir une régression complétée et une modélisation correcte. Après s'être assurer de la linéarité des catégories d'âges on décide d'affecter cette catégorie d'âge à la suivante. Mis à part le regroupement de catégories on remarque que les coefficients sont similaires (**2.1 Coefficients estimé par la loi binomiale négative).** On décidera d'utiliser ces coefficients pour la régression.

3- Modélisation pour la sévérité

On passe maintenant à la modélisation des sinistres. Les lois que l'on va utiliser sont définies sur R+, ce qui implique qu'on ne va pas considérer les montant négatif, c'est à dire les recours. Les modèles Log-Normal et Gamma seront utilisés pour modéliser la sévérité. Sous R, on a utilisé la procédure « stepwise » pour sélectionner les variables les plus significatives (celle qui minimise le critère AIC) : le type d'alimentation : hybride ou fuel et l'âge des conducteurs. Le coût moyen d'un accident a moins de variables significatives que celle de la fréquence ce qui fait sens car le coût d'un accident n'est pas aussi dépendent des variables concernent le type de conducteur et d'assurances que la fréquence l'est. Les variables contiennent moins d'information sur la sévérité.

3.1- Le modèle Log-Normal

La log-linéarisation des variables dans le modèle log-normal permet d'atténuer la représentation des sinistres avec un coût élevé. On remarque dans les résultats de la régression que le coefficient âge est positif 0,004512 (3.1 Coefficients estimés par les régressions Log-Normal et Gamma) ce qui indique une relation positive entre l'âge et le coût du sinistre. Le tableau de corrélation nous indique une relation positive entre la variable âge et les variables qui caractérisent la qualité et donc le prix du véhicule (valeur, vitesse, puissance).

	drv_age1
vh_din	0,001214911
vh_speed	0,0019199
vh_value	0,00288483

Tableau 5 : Corrélation entre l'âge et les caractéristiques du véhicule

Cela indique que les individus âgés possèdent les véhicules les plus coûteux donc ceux qui ont un coût élevé lors d'un accident. De plus lorsque l'on s'intéresse au type d'alimentation, les véhicules alimentés à l'essence ont un montant de sinistre plus élevés. On ne peut pas conclure pour les véhicules hybrides car la valeur n'est pas significative.

3.2- Le modèle Gamma

Pour le modèle gamma on observe le même effet pour l'âge que dans le modèle Log-Normal. La variable hybride n'est toujours pas significative et nous avons la même conclusion pour les véhicules à essence. La variable hybride n'étant pas significative et la variable diesel pas prise en compte dans la modélisation on s'intéresse au prix moyen d'un véhicule en fonction de son alimentation. Le prix des véhicules hybrides est plus élevé que celui des véhicules à essence et diesel. Les véhicules essence étant les véhicules les moins couteux en moyenne. Néanmoins les véhicules à essence ont une sévérité plus élevée malgré leurs prix plus faibles ce qui peut expliquer le coefficient positif dans la régression (3.1- Coefficients estimés par les régressions Log-Normal et Gamma)

Type d'alimentation	**Prix Moyen du véhicule**	**Prédiction de la sévérité** moyenne
Diesel	21184,41	1568.50
Gasoline	14232,64	1752.98
Hybride	28581,24	730.07

Tableau 6 : Prix moyen du véhicule en fonction du type d'alimentation

4- Étude des sinistres graves

Pour avoir une méthode de tarification correcte, nous devons traiter les sinistres graves. Comme les modèles peuvent être sensibles aux valeurs extrêmes, il est plus pertinent de les analyser à part. On remarque les sinistres les plus graves correspondent aux assurés les plus âgé. Pour savoir s'il est pertinent d'étudier les sinistres graves il faut analyser le montant cumulés des sinistres. On remarque que le premier sinistre représente 1.314% de la totalité des pertes de l'assureur. Les deux premiers sinistres représentent plus de la moitié des sinistres graves en montant cumulés. Il n'est donc pas pertinent de modéliser les sinistres grave à part.

Montant des sinistres	Nombre de sinistres	Age du conducteur	Type de véhicule	Montant cumulés de sinistres
253236	1	69	Commercial	1,314
173167	1	60	Tourism	2,213
116845	1	77	Commercial	2,819
90393	2	77	Tourism	3,288
80687	1	66	Tourism	3,707
50121	1	59	Commercial	3,967

Tableau 7 : Montant des sinistres graves

5-Prime Pure

5.1 - Choix des modèles

Critére/Modèle	Poisson	Binomiale-Nègative	Log-Normal	Gamma
AIC	82083	81659	37657	198305
Déviance	57588	47521	16758	15146
Log-Vraisemblance			-18819	-99147

Tableau 8: Critères des décisions des modèles

Fréquence : Les résidus des lois binomiales-négative et poisson sont similaires (voir en annexe), on décide de sélectionner la loi binomiale-négative car c'est celle qui minimise le critère AIC.
Sévérité : La régression log-normal minimise très largement le critère AIC. Il est donc préféré au model Gamma.

5.2 - Prime Pure

Pour obtenir notre tarif, on multiplie les prédictions de fréquence de sinistre (donné par le modèle binomial négatif) par le montant des sinistres (donné par le modèle log-normal). D'après le tableau ci-dessous, on retrouve une prime pure moyenne de 231,8 euros, avec une formule minimum à 118,5 euros et une formule maximum à 289 euros. On peut retrouver les détails pour chaque assuré dans le fichier base_tarif.csv.

	Min	Moyenne	Médiane	Q1	Q3	Max
Tarif	118,5	231,8	231,3	222,2	240,5	289

Tableau 9: Statistiques descriptives du tarif de notre prime pure

On remarque que la moyenne de notre prime pure est supérieure à celle de l'exemple étudié en classe, on peut déduire que notre portefeuille d'assuré commet soit plus d'accident (fréquence plus élevés) soit le coût moyen de chaque accident est plus élevé soit les deux. Cela indique un portefeuille plus risqué d'où une prime plus élevée.

	Min	Moyenne	Médiane	Q1	Q3	Max
Tarif	48.0	156.8	156.4	150.2	163.0	200.00

Prime pur de l'exemple vu en classe

Annexe :

1- Statistiques descriptives

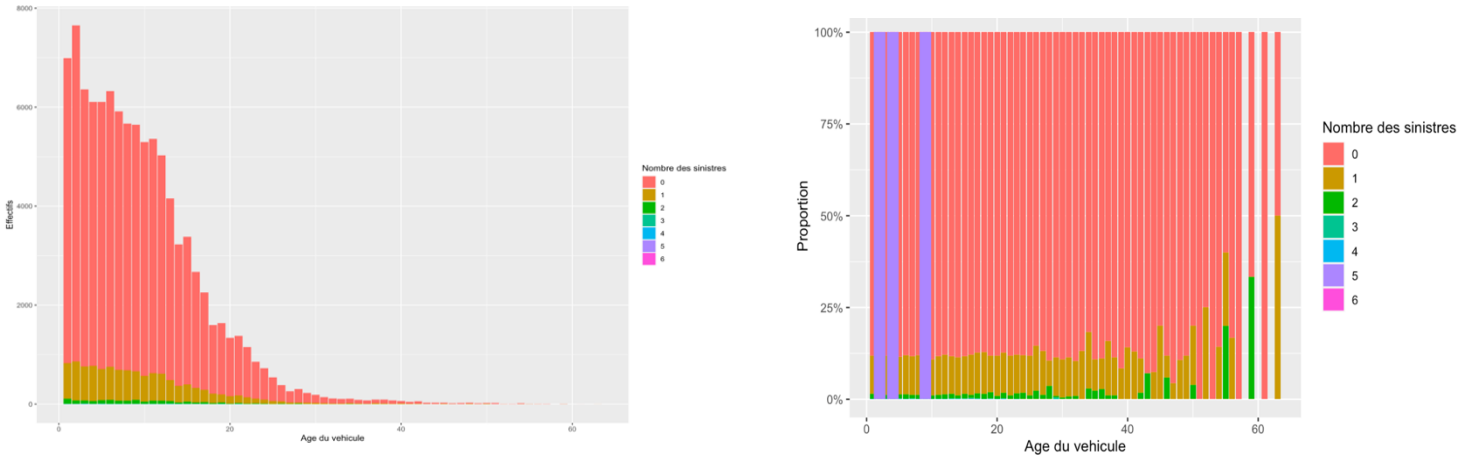


Figure 2: Nombre de sinistres en fonction de l'âge du véhicule

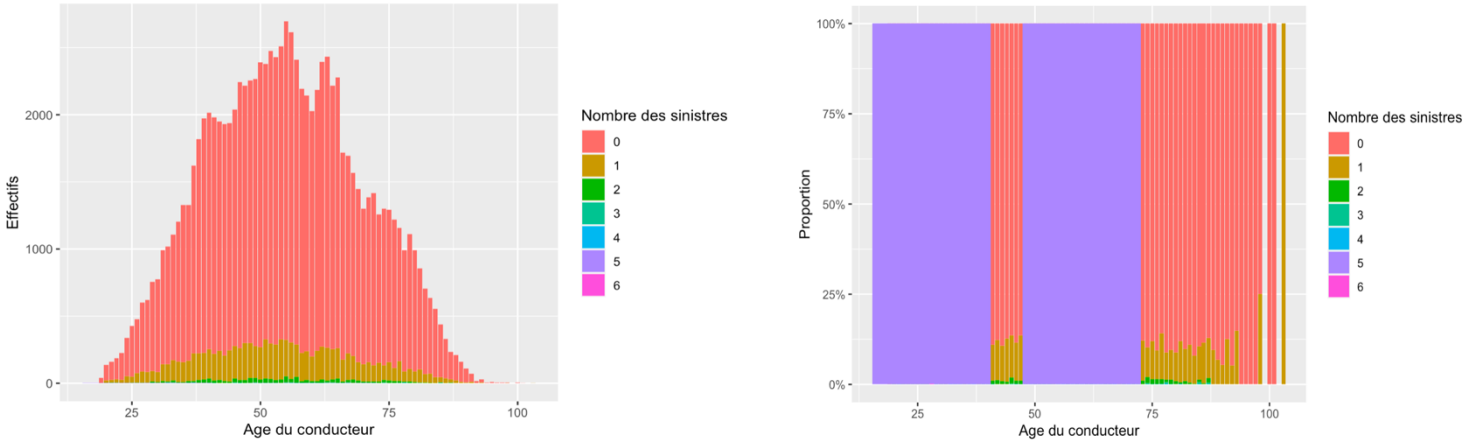


Figure 3: Nombre de sinistres en fonction de l'âge du conducteur

2. Régression pour la Fréquence

2.1 Coefficients estimés Régression de Poisson / Négative Binomial

Coefficients:	Coefficents Estimés Modèle de Poisso	Ecart Type
(Intercept)	-1,91344	0,02682
pol_coverageMedian1	-0,05445	0,03129
pol_coverageMedian2	0,04957	0,0231
pol_coverageMini	-0,0103	0,03196
pol_duration2(21, Inf]	0,06078	0,02308
drv_age1.2(35,41]	-0,09612	0,03649
drv_age1.2(41,46]	-0,1162	0,03729
drv_age1.2(46,50]	-0,06511	0,03782
drv_age1.2(50,54]	-0,07678	0,03723
drv_age1.2(54,58]	-0,13392	0,03769
drv_age1.2(58,63]	-0,22036	0,03734
drv_age1.2(63,68]	-0,12454	0,03808
drv_age1.2(68,75]	-0,1987	0,03894
drv_age1.2(75, Inf]	-0,27155	0,03923
vh_age2(17, Inf]	0,04998	0,02637

Tableau 10: Coefficients estimés par la Régression de Poisson

Coefficients:	Coefficents Estimés Modèle Negatif Binomial	Ecart Type
(Intercept)	-1,913429	0,028456
pol_coverageMedian1	-0,053899	0,03289
pol_coverageMedian2	0,049341	0,024414
pol_coverageMini	-0,009862	0,033661
pol_duration2(21, Inf]	0,061083	0,024369
drv_age1.2(35,41]	-0,096329	0,038636
drv_age1.2(41,46]	-0,116184	0,039458
drv_age1.2(46,54]	-0,071426	0,034042
drv_age1.2(54,58]	-0,134107	0,039854
drv_age1.2(58,63]	-0,220306	0,039382
drv_age1.2(63,68]	-0,124723	0,040273
drv_age1.2(68,75]	-0,198871	0,041078
drv_age1.2(75, Inf]	-0,271498	0,41289
vh_age2(17, Inf]	0,050047	0,027862

Tableau 11: Coefficients estimé par la loi binomiale négative

2.2 Comparaison des coefficient Modèle Poisson/Binomial Négatif

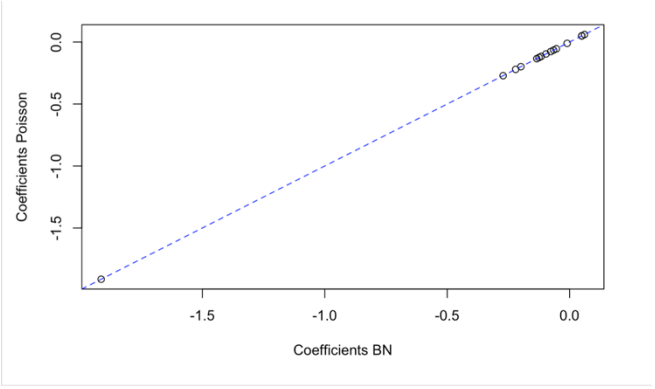
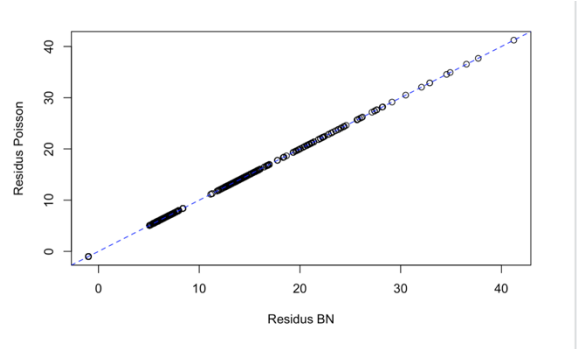


Figure 4: Comparaison des coefficients

2.4 Résidus des Modèles Poisson/Binomial Négatif



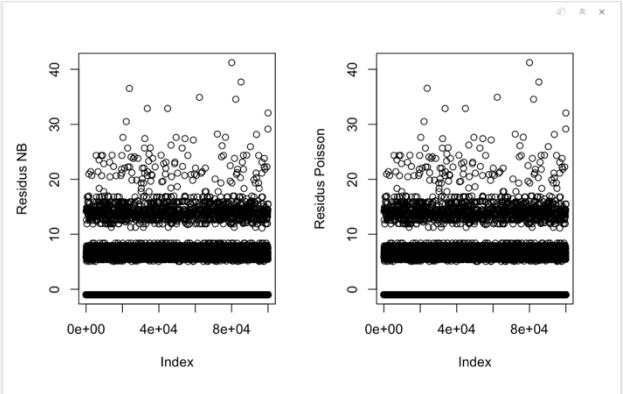


Figure 5 : Résidus Poisson/Négative Binomial

Figure 6 : Comparaison des résidus Poisson/Négative Binomial

3.Régression pour la Sévérité

3.1. Coefficients estimés par la régression Log-Normal et Gamma

	Coefficents Estimés Modèle Log-Normal	Ecart-type	Coefficents Estimés Modèle Gamma	Ecart-type
(Intercept)	6,49715	0,042555	7,081436	0,075963
vh_fuelGasoline	0,030061	0,022084	0,089458	0,03922
vh_fuelHybrid	-0,564778	0,308031	-0,785899	0,549856
drv_age1	0,004512	0,000741	0,00509	0,001323

Tableau 12: Comparaison des coefficients estimés par les modèles Log-Normal et Gamma

3.2. Résidus des régressions Log-Normal et Gamma

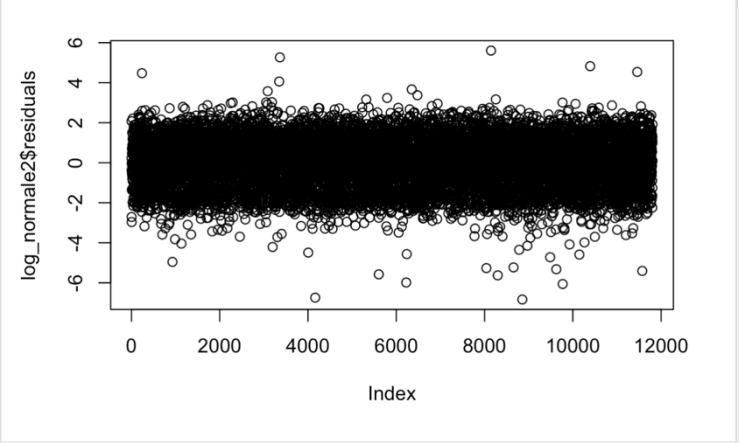


Figure 7 : Résidus Log-Normal

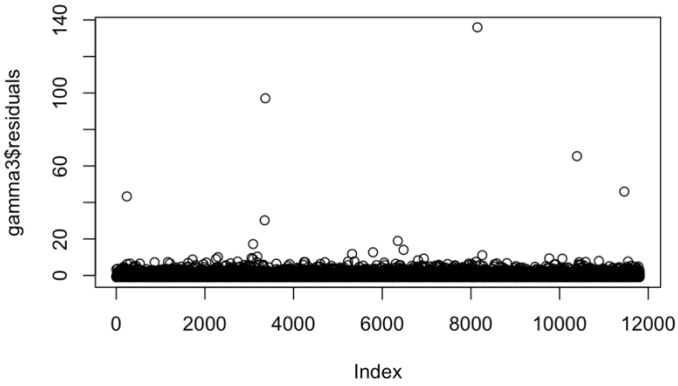


Figure 8 : Résidus Gamma