Faculté d'économie

Analyse des dynamiques liées à l'augmentation générale des primes d'assurances dans un contexte d'incertitude économique.

AHMED DADA Milly - RAË Jolhan

M1 - MBFA - Actuariat - (2024-2025)

"Prediction is very difficult, especially if it's about the future."

— Niels Bohr, physicien danois, prix Nobel de physique.

Remerciements :

Nous tenons à exprimer notre profonde gratitude à notre chargé de projet, **M. Jules Sadefo Kamdem**, pour son accompagnement précieux et ses conseils avisés tout au long de ce projet. Nous adressons également nos sincères remerciements à **Mme Seyte** pour son soutien constant et ses orientations qui ont grandement enrichi notre travail.

Nos remerciements vont également à l'ensemble des professeurs qui ont généreusement partagé leurs connaissances et expériences, nous permettant ainsi de développer une meilleure compréhension des enjeux liés à ce projet. Enfin, nous souhaitons remercier nos camarades pour leurs échanges constructifs, leur solidarité et leur soutien tout au long de cette aventure académique.

Table des matières

1 Introduction

En 2023, les sinistres liés uniquement aux catastrophes naturelles ont atteint les **6,5 milliards d'euros** en France. Un chiffre vertigineux qui illustre une tendance inquiétante : les sinistres coûtent de plus en plus cher aux assureurs. Face à cette montée des risques, comprendre les facteurs influençant ces coûts devient essentiel.

Pourquoi les primes d'assurance augmentent-elles chaque année ? Pourquoi certains départements paient-ils plus cher que d'autres ? Derrière ces questions se cache un enjeu crucial pour les assureurs : le coût des sinistres, qui varie selon la nature des événements et l'évolution du risque. Ce projet s'inscrit donc dans un contexte d'incertitude grandissante.

Un contexte de risques croissants

Plusieurs facteurs aggravent cette vulnérabilité face aux catastrophes naturelles : la croissance économique, l'urbanisation et les effets du réchauffement climatique. La multiplication des phénomènes météorologiques extrêmes témoigne d'un changement d'échelle inquiétant. Les dernières décennies ont mis en lumière un problème autrefois sous-estimé : l'impact économique croissant des catastrophes naturelles. Dans ce contexte, la société du risque s'installe. Aujourd'hui, plus que jamais, l'assurance devient un élément central pour faire face aux conséquences des événements climatiques.

Nous assistons à une tendance préoccupante ; avec la sinistralité qui s'accroît à la fois en fréquence et en intensité, on observe une dérive de *18%* par rapport aux projections des assureurs.

Une tarification des risques à repenser

Les modèles d'évaluation des catastrophes naturelles reposent souvent sur des données historiques. Cependant, face à l'augmentation de la fréquence et de la gravité des sinistres, ces méthodes pourraient ne plus être adaptées aux nouvelles réalités climatiques. Il devient alors nécessaire de repenser la façon dont ces risques sont modélisés afin de mieux refléter leur évolution future.

Une perception du risque encore insuffisante

Bien que la fréquence des catastrophes naturelles soit en augmentation, la prise de conscience du risque reste limitée. En effet, *64%* des français de métropole résidant dans des communes exposées aux inondations ignorent leur vulnérabilité. Chaque année, environ **6 000 communes** font l'objet d'une reconnaissance d'état de catastrophe naturelle.

Une hausse des tarifs d'assurance inévitable

À partir de 2025, les primes d'assurance catastrophes naturelles connaîtront une hausse significative. En cause : des sécheresses plus intenses, des inondations plus fréquentes et une valeur immobilière en hausse, augmentant mécaniquement le coût des réparations. Ainsi, le taux des primes d'assurance passera de *12%* à *20%* sur les contrats dommages aux biens.

Ce régime, qui coûtait en moyenne **25 euros** par an et par foyer, passera à environ **40 euros**. Cette augmentation vise à garantir une capacité d'indemnisation suffisante face à l'augmentation du coût des sinistres.

Une étude ciblée sur la diversité des sinistres

L'objectif de ce projet est d'analyser l'impact financier des catastrophes naturelles en fonction du type de sinistre. L'enjeu est de mieux comprendre comment la sinistralité évolue et quelles adaptations seront nécessaires pour assurer la viabilité du système d'indemnisation à long terme.

Etat des lieux des catastrophes naturelles en France

De 1900 à 2021, plus de **520 événements naturels dommageables**, dont 185 reconnus très graves, sont survenus en France, totalisant un peu plus de **32 000 morts**. Si les inondations représentent les deux tiers des événements naturels survenus, près de *90%* des décès ont été causés par l'éruption volcanique de la montagne Pelée en 1902.

En moyenne, c'est **5700 communes** qui sont concernées par les catastrophes naturelles chaque années, et l'on compte **50 milliards d'euros** d'indemnisation versées en 40 ans par les assurances au titre des catastrophes naturelles.

Le risque d'inondations en France :

On décompte **18,5 millions** d'habitants qui sont exposés aux risques d'inondations par submersion marine ou débordement de cours d'eau. Cela représente *28,9%* de la population totale et **11,8 millions** de logements.

Les dommages indemnisés au titre des inondations ayant fait l'objet d'une reconnaissance catastrophe naturelle varient fortement selon les départements sur la période 1995-2019. Pour neuf départements, les indemnisations annuelle moyennes dépassent les 12 milliards d'euros. Il s'agit des départements du littoral méditerranéen, de la Charente-Maritime ainsi que de la Seine-et-Marne.

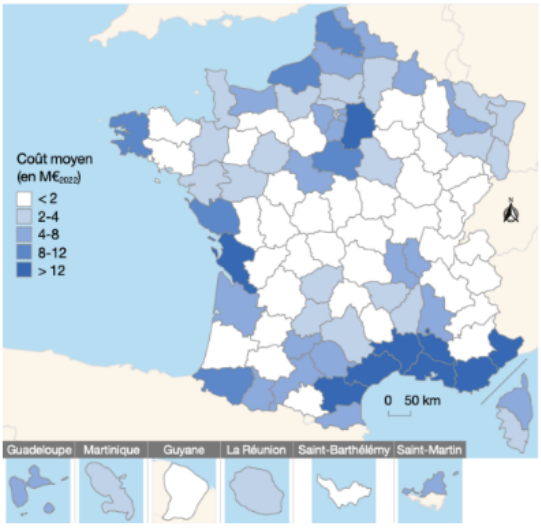

FIGURE 1 – *Carte du coût moyen des sinistres liés aux inondations par département sur la période 1995-2019.*

Les mouvements de terrain en France :

Les mouvements de terrain comprennent plusieurs phénomènes tels que :

— Le retrait-gonflement des argiles, avec *48%* du territoire métropolitain exposé, ce qui correspond à **10.4 millions d'habitations**.

— Les glissements de terrain, chutes de blocs, éboulements, coulées de boue : on décompte **65,000 événements de 1900 à 2019**, les indemnisations s'élevant à **700 millions d'euros**.

— Les séismes.

Six départements (*Alpes-Maritimes, Charente-Maritime, Hautes Pyrénées, Haute Savoie, Var, Vendée*) font l'objet d'**une sinistralité moyenne annuelle supérieure à 1M d'euros** en 2022.

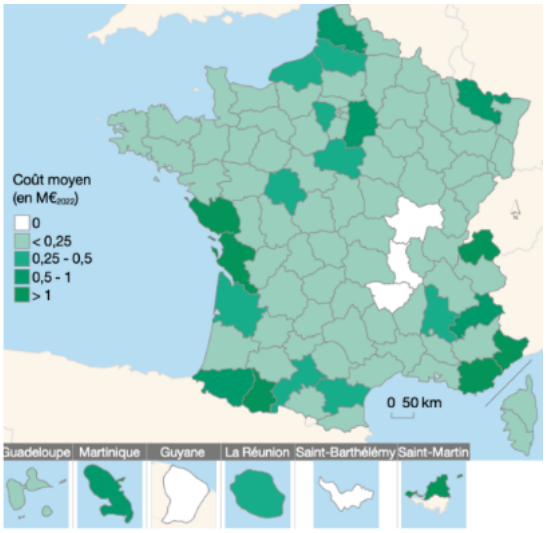

FIGURE 2 – *Le coût annuel moyen des sinistres liés aux mouvements de terrain par département sur la période 1995-2019.*

Le risque lié aux mouvements de terrain est largement répandu sur le territoire et dépend en grande partie des caractéristiques géologiques. Il peut se manifester sous différentes formes, comme les glissements de terrain, les chutes de blocs ou encore les éboulements rocheux.

L'histoire et la géologie de la France expliquent l'existence de **173 800 cavités souterraines**, dont les dimensions varient d'une simple cave à d'anciennes carrières souterraines. Une fois leur exploitation terminée, ces cavités subissent une évolution progressive et inévitable, pouvant provoquer des instabilités du sol.

Par ailleurs, les changements climatiques risquent d'accentuer le phénomène d'effondrement de ces cavités. Les fluctuations des nappes phréatiques influencent la résistance des roches et la stabilité des structures souterraines, augmentant ainsi les risques d'effondrement.

La nature imprévisible des séismes en fait des événements aux conséquences humaines, environnementales et économiques potentiellement les plus meurtrières et dommageables. Ainsi, même si l'on en décompte moins que les autres types de sinistre avec **36 séismes**

décomptés de 1962 à 2020, le coût de ces derniers s'est élevé à **560 millions d'euros** sur cette période.

La sensibilité aux glissements de terrain :

Les glissements de terrain se produisent lorsque des masses de sols ou de roches, fragilisées par des phénomènses naturels, se déplacent sous l'effet de la gravité.

En France métropolitaine, environ *20%* du territoire a une sensibilité élevée aux glissements de terrain, avec les zones montagneuses qui y sont particulièrement sensibles.

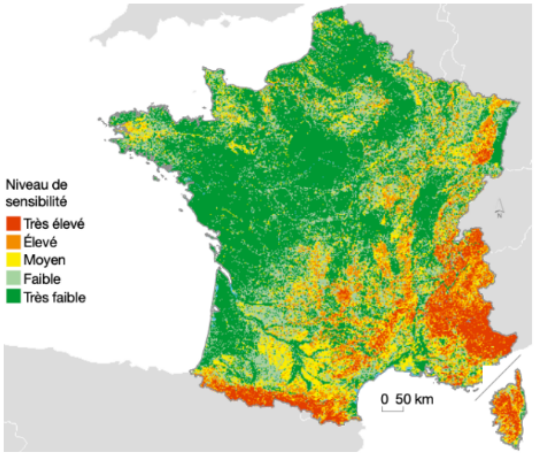

FIGURE 3 – *La sensibilité aux glissements de terrains en France.*

Depuis 1900, **30 glissements de terrain** ont été recensés, dont **24** liés à d'autres aléas naturels, tels que les crues, le ruissellement ou les coulées de boue. Certains glissements de terrain actifs font l'objet d'une surveillance particulière :

— La Clapière, dans les Alpes-Maritimes.

— Les ruines de Séchilienne, en Isère.

— Super-Sauze, dans les Alpes-de-Haute-Provence.

Les risques météorologiques et climatiques :

Parmi les risques météorologiques et climatiques, les feux de forêts représentent un risque pour **6 870 communes** avec en moyenne **26 400 hectares** de surface brûlées de 1976 à 2022.

Le nombre de tempêtes enregistrées s'élève à **360 de 1980 à 2021**, et les risques sont si élevés que près de **30 milliards d'euros** ont étés versée au titre des garanties tempête, qui peuvent être à l'origine de dommages considérables et de nombreux décès, compte tenu de la difficulté de se protéger contre le vent.

De ce fait, par la diversité de ses territoires — liée aux sols et sous-sols, aux reliefs et au climat — la France est exposée à différents phénomènes naturels, qui peuvent être regroupés en trois types d'aléas naturels principaux :

— Les aléas hydrologiques (**inondations**)
— Les aléas terrestres (**mouvements de terrain, séismes, avalanches, éruptions, etc**)
— Les phénomènes météorologiques (**vents, tempêtes, etc**)

Lorsqu'un phénomène naturel de nature aléatoire est susceptible d'affecter l'intégrité des personnes et des biens et de perturber les activités économiques, il devient un risque. L'ampleur de ce dernier dépend de la vulnérabilité des enjeux exposés.

La France s'est dotée de deux garanties permettant d'indemniser les propriétaires lorsque leurs biens assurés sont endommagés à la suite d'un événement naturel :

— La garantie **Catastrophes naturelles**.
— La garantie **Tempête-Grêle-Neige**.

Garantie Catastrophes Naturelles :

Créée en 1982, la garantie catastrophes naturelles protège les biens matériels contre les événements naturels d'intensité inhabituelle. Progressivement, cette couverture s'est étendue aux dommages sur les véhicules ainsi qu'aux effets de la sécheresse sur les bâtiments, particulièrement problématiques sur certains types de sols en France. Toutefois, ce régime est aujourd'hui menacé par le changement climatique. La **Caisse Centrale de Réassurance** estime que, d'ici 2050, la sinistralité pourrait augmenter de *27%* à *62%* uniquement en raison du réchauffement climatique. Pour faire face à cette situation, les assureurs devront ajuster leur politique tarifaire.

Garantie Tempête-Grêle-Neige :

Les dommages causés par les tempêtes (vents violents), la grêle et la neige sur des biens assurés ne relèvent pas de la catégorie "catastrophe naturelle" (Cat Nat), mais sont couverts par une garantie spécifique. En 2022, les cotisations émises pour cette garantie au niveau national sont estimées à **1,86 milliard d'euros**. Cette année a été marquée par une sinistralité exceptionnelle, avec un total d'indemnisations versées aux assurés atteignant **4,67 milliards d'euros**, soit une hausse de *386%* par rapport à 2021. Mais face au changement climatique, il convient de se demander si ces assurances ne deviennent pas obsolètes.

Les risques sont de plus en plus tangibles, l'impact du changement climatique continue de se renforcer, augmentant ainsi la vulnérabilité des populations et des territoires face aux risques naturels. Parmi ceux qui pourraient être particulièrement affectés figurent les feux de forêt, les vagues de chaleur, les sécheresses, la montée du niveau des mers et des océans, les inondations et submersions marines, ainsi que les cyclones dans les territoires ultramarins. Les conséquences de ce phénomène varieront selon les territoires français, mais on observe une tendance générale à l'augmentation de la fréquence et de l'intensité de certains événements naturels.

Le 6e rapport du GIEC (2021-2023) atteste d'une augmentation des risques. Le réchauffement de la planète va se poursuivre à court terme et devrait atteindre **1,5°C** au plus tard au début des années 2030. Selon certain scénarios, il apparaît très probable que la température moyenne à la surface du globe à l'horizon de 2100 augmente de **4,4°C**.

Une manière de se prémunir du risque : le catastrophe modeling

La mesure du risque est fondamentale à l'industrie de l'assurance. En mettant de coté les complexités inhérentes à la modélisation du risque, un problème fondamental réside : les techniques communes de mesure de risque adoptées dans l'industrie sont elles appropriées ? En calculant un risque, les compagnies d'assurance doivent considérer tous les dangers possibles. La prime payée par un client doit couvrir le dommage potentiel tout en amenant un profit suffisant à la compagnie d'assurance. C'est dans ce contexte que naît la modélisation des catastrophes, une science du secteur de l'assurance qui vise à aborder les catastrophes naturelles et leurs conséquences d'un point de vue probabiliste. Les assureurs et réassureurs utilisent ces modèles pour estimer la probabilité et le coût potentiel des sinistres liés aux catastrophes naturelles (*inondations, ouragans, séismes, sécheresses...*). Ces modèles combinent plusieurs éléments :

— Des bases de données historiques.
— Des modèles scientifiques et physiques (simulation basée sur des scénarios climatiques futurs, prise en compte des effets du changement climatique).
— L'exposition des biens assurés (localisation des bâtiments et infrastructures, valeur des biens couverts, vulnérabilité des bâtiments aux catastrophes).

Le risque peut être calculé en terme de perte annuelle, par excès de probabilité (la probabilité qu'une perte d'une certaine quantité ait lieu dans les années à suivre), ou encore en utilisant la tail value at risk (la valeur attendue de perte au delà d'un dépassement de probabilité).

Des tendances récentes voient les « **CatModeling** » appliqués à des événements autres que les catastrophes naturelles, sur des événements moins frequents tel que les épidémies, les crises financières, les agitations politiques... Le défi réside dans le fait qu'avec l'augmentation de la fréquence et de l'intensité des événements imprévisibles, les méthodes de tarification traditionnelles, surtout celles basées sur des données historiques deviennent moins efficaces.

2 Revue de la littérature

Dans le cadre de l'analyse du coût moyen des sinistres en France, il est fondamental d'examiner en profondeur la littérature scientifique et technique existante portant sur les différents types de catastrophes naturelles couvertes par le régime CatNat (inondations, sécheresse, mouvements de terrain, séismes, etc.). Cette démarche est indispensable pour identifier les spécificités économiques et géographiques de chaque type de sinistre, mais aussi pour comprendre les méthodes économétriques mobilisées pour leur modélisation. En effet, plusieurs travaux empiriques, thèses universitaires et mémoires d'actuariat permettent d'éclairer les dynamiques de coûts associées à chaque événement naturel. Ces analyses s'appuient très souvent sur des modèles économétriques avancés, notamment les modèles linéaires généralisés (`GLM`), qui permettent de capturer les variations du coût des sinistres à partir de multiples variables explicatives, telles que la typologie du sinistre, la densité de population, la valeur assurée, les caractéristiques géologiques ou encore les données climatiques locales.

Les GLM sont particulièrement pertinents dans ce contexte, car ils permettent d'intégrer des distributions non normales (comme la gamma, la log-normale ou la binomiale négative) pour modéliser des variables positives, asymétriques et dispersées, ce qui correspond parfaitement à la nature des données de coût (McCullagh & Nelder, 1989). Ils sont également adaptables à différentes fonctions de lien (logarithmique, identité, inverse, etc.), offrant ainsi une souplesse précieuse dans l'ajustement du modèle aux données empiriques observées. Ces modèles sont devenus une référence dans le domaine de l'assurance et de l'analyse actuarielle appliquée aux risques climatiques (Fox, 2016).

2.1 Inondations : fréquence élevée et coûts extrêmes localisés

La thèse de *Bourguignon* (2014) souligne que les inondations représentent environ 50% des indemnisations totales du régime CatNat entre 1982 et 2009. Cette part importante s'explique par leur forte fréquence et leur répartition sur des zones densément urbanisées. Ce sinistre affecte souvent plusieurs milliers de logements ou d'infrastructures simultanément, ce qui entraîne une accumulation rapide des montants indemnisés. Le mémoire de Hamza El Hassani (2017) ajoute que dans les cas extêmes, le coût d'un sinistre d'inondation peut excéder les 6 millions d'euros notamment dans les zones urbaines à forte valeur foncière ou industrielle. Il montre également que la distribution des coûts est très hétérogène et sensible à des facteurs comme la durée de submersion, la hauteur d'eau atteinte, et la qualité des dispositifs de protection en place.

Le mémoire *Modélisation dynamique du coût des inondations historiques* renforce cette idée en identifiant la branche professionnelle (entreprises, exploitations agricoles) comme le secteur le plus vulnérable en termes de coût moyen. Le modèle économétrique utilisé repose sur un GLM avec distribution gamma, jugée pertinente pour les variables de type montant positif, et fonction de lien logarithmique. Ce modèle montre que les variables les plus explicatives du coût sont la densité d'exposition assurée, la proximité du centre de l'événement et les précipitations enregistrées. L'approche dynamique permet également d'intégrer des effets temporels comme les saisons ou les cycles de retour de crue.

2.2 Sécheresse : faible fréquence, mais coûts moyens très élevés

La thèse *Contributions des données de l'assurance à l'étude des risques naturels* (2022) démontre que la sécheresse est un sinistre d'une gravité particulière à l'échelle unitaire. Bien qu'elle soit moins fréquente que les inondations, elle concentre à elle seule jusqu'à `45% des indemnisations totales`, en raison des effets différés et structurels qu'elle produit. Les fissurations de fondations, les tassements différentiels et les dommages aux réseaux souterrains sont typiques des impacts durables de la sécheresse sur le bâti. Son coût moyen par sinistre est systématiquement supérieur à celui des autres types d'aléas, notamment parce que les réparations sont longues, coûteuses et complexes techniquement. L'atteinte des structures est progressive, ce qui retarde souvent la déclaration du sinistre, mais en amplifie la gravité finale.

Les modèles mobilisés dans cette étude sont également des GLM avec fonction de lien log et distribution gamma, afin de modéliser à la fois la probabilité d'occurrence et la gravité des sinistres. La composante géographique est essentielle dans ces analyses, car les zones argileuses sont particulièrement sensibles aux épisodes prolongés de sécheresse. Le bassin parisien, le sud-ouest et certaines zones du centre de la France sont identifiés comme les territoires les plus exposés. Enjolras (2008), dans sa thèse appliquée au secteur agricole, montre que les pertes agricoles dues à la sécheresse peuvent dépasser `plusieurs centaines de millions d'euros` selon la période de l'année, avec des effets de seuil critiques entre les stades de semis, de floraison et de récolte. Les modèles économétriques reposent ici sur des données agroclimatiques croisées avec des bases de sinistres CatNat, permettant une estimation robuste des pertes potentielles.

2.3 Séismes et mouvements de terrain : sinistres rares mais à impact potentiellement très fort

Les séismes et les mouvements de terrain constituent des sinistres rares à l'échelle nationale, mais peuvent générer des pertes massives localement, en particulier dans les zones montagneuses mal préparées. La thèse *Contribution* (2022) mentionne que ces événements représentent une part marginale des indemnisations totales, mais que leur intensité unitaire peut dépasser celle des autres aléas. Dans les Alpes ou les Pyrénées, certaines communes peu peuplées mais très vulnérables peuvent enregistrer des coûts très importants en raison de glissements de terrain ou de secousses sismiques localisées. L'absence de données systématiques sur ces événements rend leur modélisation difficile. Toutefois, lorsqu'un historique minimal est disponible, les chercheurs utilisent soit des GLM, soit des modèles semi-paramétriques intégrant des facteurs spatiaux, et géotechniques.

Certains travaux explorent également des approches mixtes, combinant la modélisation de la fréquence (via une régression logistique) et celle de la sévérité (via un GLM avec fonction de lien adaptée). Ces approches sont particulièrement utiles pour modéliser les sinistres rares, dont la distribution est souvent très dissymétrique.

2.4 Conclusion de la revue de la littérature :

Les travaux analysés montrent une diversité marquée dans la fréquence et l'intensité financière des sinistres selon leur nature. Les inondations se distinguent par leur fréquence et leur ampleur potentielle en zone urbaine, tandis que la sécheresse représente le sinistre au coût moyen le plus élevé, avec des effets différés et structurels lourds. Les séismes et mouvements de terrain, bien que rares, constituent des menaces locales graves. Cette hétérogénéité justifie pleinement une modélisation différenciée par type de sinistre.

Les modèles **GLM** se révèlent particulièrement adaptés pour modéliser les coûts des sinistres, grâce à leur flexibilité statistique, leur compatibilité avec des distributions asymétriques et leur capacité à intégrer des variables spatiales, climatiques et structurelles. Leur usage récurrent dans la littérature démontre leur pertinence pour accompagner la tarification, la prévention et l'analyse des politiques publiques de couverture des catastrophes naturelles.

3 Méthodologie : Modèle Linéaire Généralisé (GLM)

3.1 Présentation du modèle GLM

Le **modèle linéaire généralisé** (GLM) est une **extension** des **modèles linéaires classiques**. Il permet d'analyser les relations entre des variables dépendantes et explicatives, même lorsque les hypothèses de **normalité des résidus** et **d'homoscédasticité** ne sont **pas respectées**. Ce modèle est particulièrement adapté lorsque la variable dépendante suit une distribution asymétrique.

Le GLM repose sur trois composantes principales :

a) La loi de probabilité

On suppose que la variable dépendante Y suit une loi appartenant à la famille exponentielle, incluant :

— La loi *Normale* (régression linéaire classique),
— La loi *de Poisson* (variables de comptage),
— La loi *Binomiale* (données dichotomiques ou proportionnelles),
— La loi *Gamma* (données strictement positives continues),
— La loi *Inverse-Gaussienne* ...

Dans cette étude, deux modèles sont comparés :

— Un modèle avec la *loi de Poisson*,
— Un modèle avec la *loi Gamma*, adaptée à la variable `moytotal` représentant un coût moyen strictement positif.

La loi de *Poisson* est souvent utilisée pour modéliser des événements rares dans des données de comptage, mais elle est fondée sur une hypothèse fondamentale : **la variance est égale à la moyenne** (les données suivent une disposition de type équidistribution). C'est-à-dire que si λ est le paramètre de la loi de *Poisson*, alors :

$$E[Y] = \mathrm{Var}(Y) = \lambda$$

Cependant, dans le cas de certaines données économiques ou environnementales, **la variance peut être plus grande que la moyenne**, un phénomène appelé *surdispersion* (Gelman & al.,2013). Cette *surdispersion* peut survenir lorsque les événements observés sont **plus variables** que ce que la loi de *Poisson* permet de modéliser.

Dans le cas de la *surdispersion*, la **variance devient supérieure à la moyenne**, ce qui signifie que les données montrent une **variabilité plus importante** que ce que la loi de *Poisson* permet de modéliser.

Pour traiter cette situation, la loi *Gamma* a été choisie, permettant de modéliser une variance qui dépend de la moyenne à travers un paramètre de forme. α :

$$\mathrm{Var}(Y) = \alpha \cdot \mu^2$$

Cette approche permet une **meilleure flexibilité** pour expliquer la variabilité des données, notamment dans les cas où la *dispersion* est plus grande que ce qui est prévu par la loi de *Poisson*.

b) La fonction de lien $g(\mu)$:

Le modèle linéaire généralisé (GLM) relie la moyenne conditionnelle de la variable dépendante Y, notée $\mu = E(Y \mid X)$, à une combinaison linéaire des variables explicatives via une fonction de lien g :

$$g(\mu) = \eta = X\beta$$

où :

— η est le *prédicteur linéaire* (comme dans une régression linéaire classique),
— $X\beta$ est la *combinaison linéaire des variables explicatives* $X_1, X_2, \ldots, X_k$,
— g est une fonction appelée *fonction de lien*, choisie en fonction de la loi suivie par Y.

Les fonctions de lien usuelles sont :

— *Identité* : $g(\mu) = \mu$
— *Logarithme* : $g(\mu) = \log(\mu)$
— *Logit* : $g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$
— *Inverse* : $g(\mu) = \frac{1}{\mu}$

La fonction de lien logarithmique (lien *log*) garantit que $\mu > 0$, ce qui est essentiel pour des données telles que les coûts ou les nombres d'occurrences, qui ne peuvent pas être négatifs (Agresti, 2013). De plus, elle permet une interprétation multiplicative des coefficients, c'est-à-dire en termes d'effets relatifs (en pourcentage), ce qui est souvent plus parlant dans des domaines comme l'économie ou l'assurance.

Dans cette étude, c'est donc cette fonction de lien qui sera utilisée, implémentée des lois de *Poisson* et *Gamma*.

c) La structure linéaire

La **structure linéaire** constitue **le noyau d'un modèle linéaire généralisé** (GLM). Elle correspond à l'équation qui relie les variables explicatives $X_1, X_2, \ldots, X_k$ à la variable dépendante Y, au moyen d'un prédicteur linéaire noté η.

La forme générale du prédicteur linéaire η s'écrit :

$$\eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

ou, de manière plus compacte, sous forme matricielle :

$$\eta = X\beta$$

où :

— X est la *matrice des variables explicatives* (de dimension $n \times (k+1)$),
— β est *le vecteur des coefficients* (de dimension $(k+1) \times 1$),
— η est un vecteur de dimension $n \times 1$, contenant les prédicteurs linéaires pour chaque observation.

Chaque coefficient β_j mesure **l'effet marginal** de la variable X_j sur η, toutes choses égales par ailleurs.

Cependant, il est important de noter que dans un modèle linéaire généralisé (GLM), η n'est pas directement égal à la variable dépendante Y, ni même à sa moyenne μ, mais à une fonction de cette moyenne :

$$g(\mu) = \eta$$

Ainsi, la variable dépendante est reliée de manière indirecte aux variables explicatives, via une fonction de lien. Cette transformation rend le modèle plus souple, permettant d'exprimer des relations non-linéaires sur les données d'origine, tout en conservant la linéarité des effets dans l'espace transformé.

3.2 Tests de normalité et d'hétéroscédasticité des résidus

Bien que les hypothèses classiques (résidus i.i.d) ne soient pas nécessaires pour un GLM, certains tests sont utilisés pour évaluer l'ajustement du modèle.

a) Tests de Shapiro-Wilk :

H_0 : les résidus suivent une distribution normale.

H_1 : les résidus ne suivent pas une distribution normale.

Le test de Shapiro-Wilk utilise la statistique W, qui est définie comme suit :

$$W = \frac{\left(\sum_{i=1}^{n} a_i e_i\right)^2}{\sum_{i=1}^{n}(e_i - \bar{e})^2}$$

où :

— a_i est un coefficient spécifique pour chaque observation,

— e_i est le résidu de l'observation i,

— $\bar{e}$ est la moyenne des résidus.

Un coefficient W proche de 1 indique que la normalité des résidus est respectée.

Décision :

Si la p-value < 0.05 : rejet de H_0.

Si la p-value ≥ 0.05 : on ne rejette pas H_0.

b) Test de Kolmogorov-Smirnov :

H_0 : les résidus suivent une distribution normale.

H_1 : les résidus ne suivent pas une distribution normale.

Le test de Kolmogorov-Smirnov utilise la statistique D, qui est définie comme suit :

$$D = \max_{1 \leq i \leq n} \left| \frac{i}{n} - F(e_i) \right|$$

où :

— i est l'indice de l'observation dans l'échantillon ordonné (de $i = 1$ à $i = n$),

— n est la taille de l'échantillon,

— e_i est le résidu de l'observation i,

— $F(e_i)$ est la fonction de répartition théorique (ici la loi normale) évaluée en e_i.

Un D élevé indique que la différence entre la distribution empirique des résidus et la loi normale est importante, suggérant que les résidus ne suivent pas une distribution normale.

Décision :

Si la p-value < 0.05 : rejet de H_0 (les résidus ne suivent pas une distribution normale).
Si la p-value ≥ 0.05 : on ne rejette pas H_0 (les résidus suivent une distribution normale).

c) Test d'hétéroscédasticité

Une procédure en 5 étapes est utilisée :

1. **Estimation du modèle GLM** : La première étape consiste à ajuster un modèle linéaire généralisé (GLM) aux données, en choisissant une loi de probabilité adéquate (par exemple, la loi Gamma) et une fonction de lien appropriée (par exemple, le lien logarithmique). Cela signifie que nous ajustons un modèle pour prédire la variable dépendante en fonction des variables explicatives. La fonction de lien logarithmique garantit que les prédictions ($\hat{\mu}_i$) sont strictement positives, ce qui est essentiel pour des données telles que les coûts, qui ne peuvent pas être négatifs. Le modèle peut être formulé sous la forme :

   $$g(\mu_i) = \eta_i = X_i \beta$$

   où $g(\mu_i) = \log(\mu_i)$ est la fonction de lien logarithmique et $X_i\beta$ représente la combinaison linéaire des variables explicatives pour l'observation i.

2. **Extraction des résidus** e_i **et valeurs ajustées** $\hat{\mu}_i$: Après avoir ajusté le modèle,les **résidus** e_i sont extraits, mesurant la différence entre les valeurs observées et les valeurs ajustées. Les résidus sont calculés comme suit :

   $$e_i = Y_i - \hat{Y}_i$$

   où Y_i est la valeur observée de la variable dépendante et $\hat{Y}_i$ est la valeur prédite par le modèle. Les **valeurs ajustées** $\hat{\mu}_i$ sont également extraites, permettant les prédictions du modèle pour chaque observation i, c'est-à-dire $\hat{\mu}_i = \exp(\eta_i)$.

3. **Transformation des résidus et valeurs ajustées** : Dans cette étape, les transformation des résidus au carré e_i^2 et des valeurs ajustées $\hat{\mu}_i$ sont effectuées en prenant leur logarithme :

$$\log(e_i^2) \quad \text{et} \quad \log(\hat{\mu}_i)$$

Cela est fait afin de linéariser la relation entre la variance des résidus et les valeurs ajustées. La transformation logarithmique est utilisée pour rendre plus évidente la relation de proportionnalité entre les résidus et les valeurs ajustées, et ainsi faciliter la détection d'hétéroscédasticité (c'est-à-dire de variabilité non constante des erreurs en fonction des prédictions du modèle).

4. **Régression auxiliaire** : La régression auxiliaire est ensuite effectuée entre les résidus transformés $\log(e_i^2)$ et les valeurs ajustées $\log(\hat{\mu}_i)$:

$$\log(e_i^2) = \alpha + \beta \log(\hat{\mu}_i) + \varepsilon_i$$

Cette régression permet d'étudier l'effet des valeurs ajustées $\hat{\mu}_i$ sur la variance des résidus. Si la variance des résidus dépend linéairement des valeurs ajustées, cela peut être un signe d'hétéroscédasticité. - α est l'ordonnée à l'origine, - β est le coefficient de pente, - ε_i est l'erreur du modèle auxiliaire.

5. **Interprétation** : Enfin, l'interprétation des résultats de la régression auxiliaire permet de tester l'hypothèse d'homoscédasticité :

   — **Hypothèse nulle** H_0 : Il n'y a pas d'hétéroscédasticité, c'est-à-dire que $\beta = 0$. Cela signifie que la variance des résidus est constante, indépendamment des valeurs ajustées $\hat{\mu}_i$.
   — **Hypothèse alternative** H_1 : Il existe de l'hétéroscédasticité, c'est-à-dire que $\beta \neq 0$. Cela suggère que la variance des résidus dépend des valeurs ajustées, ce qui peut signifier que certaines observations sont plus variables que d'autres.

Décision :

— Si la p-value du test de la pente β dans la régression auxiliaire est inférieure à 0.05, nous rejetons l'hypothèse nulle H_0, indiquant ainsi une présence significative d'hétéroscédasticité dans le modèle.
— Si la p-value est supérieure ou égale à 0.05, nous ne rejetons pas H_0, ce qui suggère qu'il n'y a pas de preuve suffisante d'hétéroscédasticité.

4 Analyse de Données

4.1 Traitement des données

Le traitement des données a été une étape capitale pour garantir leur qualité et leur compatibilité avec l'analyse économétrique. Il s'est déroulé en plusieurs phases :

— Téléchargement des données
— Harmonisation des formats
— Consolidation des fichiers
— Calcul de la variable dépendante

Téléchargement des données

Les données sur les nombres de sinistres ont été collectées via Géorisques, pour quatre types de catastrophes naturelles : mouvements de terrain, inondations, sécheresses et séismes, sur la période 1995–2023 et pour toutes les communes de France métropolitaine.

Les données sur le coût moyen des sinistres ont été récupérées séparément, mais sur la même période et avec le même découpage communal, garantissant un maximum de cohérence au niveau des données.

La récupération des données étant effectuée, leur traitement pouvait être amorcé en commençant par une simple visualisation des données.

Harmonisation des formats

À ce moment-là, un premier problème a été détecté dans les fichiers de coût des sinistres : les montants étaient exprimés sous forme d'intervalles monétaires, format impraticable pour une étude économétrique.

Pour rendre ces données exploitables, une feuille de conversion a été créée, associant à chaque intervalle une valeur médiane correspondante.

La formule Excel utilisée pour la conversion des intervalles en valeurs numériques est basée sur la fonction `RECHERCHEV()` :

```
=RECHERCHEV(E2;Conversion!$A$1:$B$7;2;FAUX)
```

Le problème est donc résolus en remplaçant chaque intervalle par sa valeur médiane. Cette opération a été répétée pour chaque type de sinistre puisque ce problème concernait tous les fichiers de coûts, peu importe le type.

Consolidation des fichiers

Une fois l'harmonisation des formats effectuée, la fusion de toutes les informations dans un seul fichier permit d'avoir un fichier unique recensant toutes les variables explicatives.

Comme les données étaient toutes associées aux communes, une simple opération de copier-coller a permis de les regrouper.

Afin de tester plusieurs modèles, la création de la base de données par département s'est faite en quelques étapes simples :

— Création d'une colonne `département` comprenant les deux premiers chiffres du `code_insee`

— Mise en place d'un tableau croisé dynamique pour obtenir le `nombre de sinistres par département`

— Création d'un autre tableau croisé dynamique pour obtenir la `moyenne des coûts par département`

Calcul de la variable dépendante

Un fichier contenait donc les coûts des sinistres par type, ainsi que le nombre d'occurrences, par communes françaises, et un autre les mêmes variables mais cette fois-ci par département.

Pour chaque observation, la moyenne des coûts moyens des sinistres a été calculée. Cette valeur a été retenue comme **variable dépendante** de la régression.

4.2 Présentation des modèles

L'étude économétrique s'est basée sur la construction de deux modèles bien distincts : dans un premier temps, un modèle testé avec des données à l'échelle départementale, puis dans un second temps, un modèle utilisé uniquement avec les données communales. Ces derniers utilisent les mêmes variables dépendantes et explicatives, mais seul le nombre d'observations change.

Pour chacun des deux modèles, l'étude des statistiques descriptives et de la corrélation entre variables est effectuée, permettant à la fin de retenir le plus pertinent.

Modèle 1 : échelle départementale

Avant cela, un rapide rappel de la structure des données est effectué afin de comprendre plus en profondeur l'étude :

Structure des données

Variables explicatives :

- `sinmvt` : Nombre d'occurrences de sinistres liés aux mouvements de terrain
- `sininond` : Nombre d'occurrences de sinistres liés aux inondations
- `sinsech` : Nombre d'occurrences de sinistres liés aux sécheresses
- `sinseisme` : Nombre d'occurrences de sinistres liés aux séismes
- `Département` **: Départements associés aux nombres de sinistres**
- `moysech` : Moyenne du coût des sinistres de sécheresse par département
- `moyinond` : Moyenne du coût des sinistres d'inondation par département
- `moyséisme` : Moyenne du coût des sinistres de séisme par département
- `moymvt` : Moyenne du coût des sinistres de mouvements de terrain par département
- `sintotal` : Nombre total d'occurrences de sinistres par département, tout type de sinistre confondu

Variable dépendante :

- `moytotal` : Moyenne totale des coûts des sinistres par département, tout type de sinistre confondu

Statistiques Descriptives des Variables :

Variable	Obs	Moyenne	E.T.	Min	Max
Département	0				
sinmvt	98	697.541	3424.093	1.00	34156
sininond	98	2791.643	13617.8	16.00	135962
sinsech	98	592.031	2927.979	0.00	29009
sinseisme	98	12.857	67.886	0.00	630
sintotal	97	2677.297	1622.149	18	7813.56
moymvt	95	8110.705	4687.385	1250	26458.33
moyinond	98	6241.352	1881.304	3500	17011.23
moysech	95	12941.02	4156.927	3750	25000
moyseisme	21	3790.431	2717.139	1250	11875
moytotal	97	28084.75	9516.901	9128.836	668469.56

Table 1 – Statistiques descriptives Modèle 1, Observations 1-98

A l'aide de la fonction **summarize** de stata, le tableau ci-dessus est obtenu, permettant de montrer les statistiques descriptives de chacune des variables. Il est observé que certaines variables n'ont pas un grand nombre d'observations (`moyseisme`), ce qui est normal puisque nombreux sont les départements qui n'ont pas subi de séisme. La moyenne du nombre d'occurrence n'est donc pas très élevée, malgré un coût moyen important (les dégâts des séismes font partie des plus dévastateur et des plus coûteux).

Les résultats obtenus étant au regard des données, nous sommes en droit de nous demander si ces derniers sont pertinents et interprétables, étant donné le faible nombre d'observations.

Test d'Indépendance et multicolinéarité :

Les statistiques descriptives des données étant probablement biaisées compte tenu de la taille de l'échantillon, il est capital d'effectuer les tests de multicolinéarité puisque le manque de fiabilité des résultats pourrait s'y refléter.

La matrice de corrélation est calculée via la commande **correlate** de stata, faisant apparaître la table de corrélation suivante :

	moytotal	sinmvt	sininond	sinsech	sinseisme
moytotal	1.000				
sinmvt	-0.2181	1.000			
sininond	0.0902	0.5196	1.000		
sinsech	0.0432	0.2225	0.3251	1.000	
sinseisme	0.0504	-0.1266	0.0599	-0.0905	1.000

Table 2 – Matrice des coefficients de corrélation, modèle 1

Elle permet d'identifier, dans ce modèle, les relations linéaires entre les variables explicatives et dépendantes. Seul les variables `moytotal, sinmvt, sinseisme, sinsech et sininond` sont utilisés puisque uniquement elles seront retenues pour la mise en place de la régression. La forte corrélation entre les variables `sininond` et `sinmvt` (=**0.5196**) ou encore entre `sininond` et `sinsech` (=**0.3251**) montre une redondance d'information, ce qui est sans doute le reflet d'un manque de fiabilité du modèle.

Un test de **VIF** est ensuite effectué, permettant de quantifier la colinéarité entre variable :

Variable	VIF	1/VIF
sininond	1.51	0.664062
sinmvt	1.42	0.703593
sinsech	1.14	0.88014
sinseisme	1.05	1.951452
Mean VIF	1.28	

Table 3 – Tableau des VIFs, modèle 1

Aucune valeur supérieure à 10 n'indique la nécessité de supprimer ou de transformer une ou plusieurs variables. Les résultats sont satisfaisants.

Régression GLM :

Variable dépendante : **moytotal**

moytotal	Coefficient	éc. type	z	P. critique
sinmvt	-0.0004677	2.77e-06	-168.65	0.000
sininond	0.0000955	8.08e-07	118.27	0.000
sinsech	0.0000346	1.59e-06	21.76	0.000
sinseisme	-0.0001359	0.0000242	-5.60	0.000
_cons	10.25764	0.0012372	8291.09	0.000
AIC		2787.76	BIC	268815.5
Déviance		269236.3766	Pearson	268659.4062

Table 4 – Modèle 1 : GLM, 98 observations

L'ensemble des coefficients est significatif, mais certains d'entre eux semblent avoir un impact négatif sur la variable dépendante (`sinmvt` et `sinseisme`), ce qui laisserait imaginer qu'une augmentation d'une unité de sinistre lié au mouvement de terrain ou au séisme entraînerait une diminution du coût total des sinistres,toutes choses égales par ailleurs. Un critère **AIC** à **2787.76** permettra de comparer ce modèle au modèle 2.

Modèle 2 : échelle communale

Structure des données

Variables explicatives :

— `sinmvt` : Nombre d'occurrences de sinistres liés aux mouvements de terrain
— `sininond` : Nombre d'occurrences de sinistres liés aux inondations
— `sinsech` : Nombre d'occurrences de sinistres liés aux sécheresses
— `sinseisme` : Nombre d'occurrences de sinistres liés aux séismes
— `code_insee` **: Code associé à la commune concernée**
— `moysech` : Moyenne du coût des sinistres de sécheresse par commune
— `moyinond` : Moyenne du coût des sinistres d'inondation par commune
— `moyséisme` : Moyenne du coût des sinistres de séisme par commune
— `moymvt` : Moyenne du coût des sinistres de mouvements de terrain par commune
— `sintotal` : Nombre total d'occurrences de sinistres par commune, tout type de sinistre confondu

Variable dépendante :

— `moytotal` : Moyenne totale des coûts des sinistres par commune, tout type de sinistre confondu

Statistiques Descriptives des Variables :

Ce dernier modèle reprend les variables explicatives et dépendantes des modèles précédents, mais avec des données à l'échelle communale plutôt que départementale

Variable	Obs	Moyenne	E.T.	Min	Max
code_insee	0				
sinmvt	34839	0.9804	0.6968	0.00	20.00
sininond	34839	3.9026	2.6750	0.00	42.00
sinsech	34839	0.8327	1.7774	0.00	22.00
sinseisme	34839	0.0181	0.1493	0.00	6.00
sintotal	34839	5.7337	3.8012	0.00	66.00
moymvt	34839	545.7246	3084.524	0.00	25000
moyinond	34839	3040.307	5327.466	0.00	25000
moysech	34839	3732.778	7207.148	0.00	25000
moyseisme	34839	42.2300	583.037	0.00	25000
moytotal	34839	1311.94	2368.144	0.00	15833.33

TABLE 5 – Statistiques descriptives Modèle 2, Observations 1-34839

L'observation des résultats permet ici de voir que le nombre d'observations joue un rôle capital dans l'étude économétrique. Les moyennes ainsi que les écart-types semblent être bien plus plausibles et pertinents. Reprenons le cas de `sinséisme` : une moyenne à *0.018* et un écart-type à *0.149* met en avant des résultats qui semblent bien plus se rapprocher de la réalité.

Test d'Indépendance et multicolinéarité :

La matrice de corrélation est effectuée de la même façon que pour l'ancien modèle, mais avec un nombre de données nettement plus élevé.

	moytotal	sinmvt	sininond	sinsech	sinseisme
moytotal	1.000				
sinmvt	0.1739	1.000			
sininond	0.2800	0.2058	1.000		
sinsech	0.6620	0.1340	0.2624	1.000	
sinseisme	0.0438	-0.0487	0.0620	-0.0125	1.000

Table 6 – Matrice des coefficients de corrélation, modèle 2

Les coefficients de corrélation entre variables semblent bien plus plausible, ne notifiant pas de réel problème dans la multicolinéarité des variables utilisées.

Le Test de **VIF** est aussi effectué, aucun besoin de suppression/transformation de variable.

Variable	VIF	1/VIF
sininond	1.12	0.896094
sinmvt	1.06	0.947269
sinsech	1.08	0.923909
sinseisme	1.01	0.991606
Mean VIF	1.07	

Table 7 – Tableau des VIFs, modèle 2

Aucune valeur supérieure à 10 n'indique la nécessité de supprimer ou de transformer une ou plusieurs variables. Les résultats sont satisfaisants.

Régression GLM :

Variable dépendante : **moytotal**

moytotal	Coefficient	éc. type	z	P. critique
sinmvt	0.2601561	0.0200925	12.95	0.000
sininond	0.1019587	0.0060458	16.86	0.000
sinsech	0.8855518	0.017451	50.75	0.000
sinseisme	0.9496723	0.1068574	8.89	0.000
_cons	5.10792	0.0362218	141.02	0.000
AIC		15.0311	BIC	-315970.6
Déviance		48340.51262	Pearson	274719.6032

Table 8 – Modèle 2 : GLM, 34839 observations

La significativité et la pertinence de tous les coefficients place déjà ce modèle comme étant plus performant que le modèle 1. Aucun coefficient négatif n'est répertorié, ce qui se rapproche bien plus de ce qui est observé empiriquement. Finalement le critère **AIC** à *15.03* renforce l'idée que ce modèle soit nettement supérieur au premier.

4.3 Comparaison des modèles :

L'objectif de cette section est de comparer les **modèles 1 et 2**, qui intègrent les variables liées au coût moyen des sinistres. Cette comparaison se fait selon deux critères : la **significativité des variables** et la **qualité de l'ajustement**.

Rappelons que la seule et unique différence entre les deux modèles reposent sur le **nombre d'observations** : *98* pour le modèle 1 (*nombre de départements français*) et *34839* pour le modèle 2 (*nombre de communes françaises*). Il est important de constater pourquoi le modèle 1 a été nettement moins performant que le modèle 2. La taille de l'échantillon a un effet sur l'instabilité des estimations des coefficients, la fiabilité des probabilités, ainsi que sur la puissance d'un test : avec moins d'observations, on risque de ne pas détecter les effets significatifs du modèle.

Le modèle 1 est caractérisé par des coefficients significatifs, mais quelquefois négatifs, ce qui semble assez contre intuitif dans notre cas. Une augmentation d'une unité de la variable `sinmvt` ferait baisser le logarithme de l'espérance de `moytotal` de *-0.0004677*. Une augmentation d'une unité de `sinmvt` entraînerait donc une diminution du coût total, ce qui ne semble pas logique. Le modèle 1 se caractérise aussi par de très faibles coefficients (tous très proches de 0) et par un coefficient de constante très élevé (proche de 10). Un critère **AIC** qui vaut *2787.76* montre une perte d'information trop élevée, renforçant l'idée que le modèle 1 n'est définitivement pas un modèle performant.

Le modèle 2 relève le niveau puisque les résultats semblent bien plus pertinents : des coefficients positifs, qui s'éloignent de 0 (se rapprochant de l'ordre de grandeur du coefficient de la constante). Un critère **AIC** de *15.03*, nettement plus bas, montre qu'il performe bien plus que le modèle 1.

Ce dernier est donc logiquement retenu puisque parmi les modèles testés, c'est celui qui minimise les critères AIC et BIC.

4.4 Tests liés au modèle :

4.4.1 Test d'hétéroscédasticité :

res2	Coef	Ec.type	t	p-value
mu2	-5.86e-23	2.35e-22	-0.25	0.803
_cons	7.885929	0.2408456	32.74	0.000

TABLE 9 – Test d'hétéroscédasticité, modèle 2

Dans un premier temps, un test d'hétéroscédasticité est effectué manuellement puisque la commande du test ARCH n'est pas valide pour un GLM. On retient que la p-value de **mu2** = *0.803* est nettement supérieure au seuil alpha = 5%, on **ne rejette pas l'hypothèse H_0 d'homoscédasticité**. Cette régression auxiliaire met donc en avant l'homoscédasticité du modèle, ce qui est un bon point de départ pour des inférences standards.

4.4.2 Test de Lien :

Un test de lien est effectué afin de comprendre la pertinence de la fonction de lien *log* utilisée, il permet de vérifier la bonne spécification du modèle. Il est testé si une variable quadratique *(_hatsq)* ajoute une information explicative.

moytotal	Coefficient	éc.type	z	P. critique
_hat	2439.969	24.56247	99.34	0.000
_hatsq	-82.96222	1.324835	-65.62	0.000
_cons	-10821.14	102.4506	-105.62	0.000

TABLE 10 – Test de lien *Log*, Modèle 2

Les résultats montrent que le coefficient de *_hatsq* négatif et très significatif (**p = 0.000**), ce qui indique que le modèle est mal spécifié. La significativité de *_hatsq* suggère qu'une relation non linéaire ou une variable manquante est ignorée. D'après le test, la fonction de lien en *log* ne serait pas la plus pertinente à utiliser, c'est malgré tout celle qui sera retenue puisque un changement de fonction de lien entraîne toujours, dans notre cas, une augmentation du critère **AIC**.

Une fois ces tests élémentaires effectués, il est également possible d'effectuer des tests sur les résidus du modèle.

4.4.3 Test de Shapiro-Wilk :

Ce dernier permet de tester si les résidus suivent une distribution *normale* (il est plus intéressant pour les petits échantillons mais on l'utilise ici malgré tout).

Variable	Obs	W	V	z	p-value
res	34839	0.60481	5527.093	23.743	0.000

TABLE 11 – Test de Shapiro-Wilk, Modèle 2

On observe une probabilité inférieure à 5%, on rejette donc l'hypothèse de normalité des résidus.

Composantes du test :

Le test de Shapiro-Wilk fournit plusieurs résultats importants pour l'analyse de la normalité des résidus :

- **Obs (Observation)** : Le nombre d'observations utilisées pour effectuer le test. Dans ce cas, $N = 34839$ observations.
- **W** : La statistique du test de Shapiro-Wilk, qui mesure l'adéquation de l'échantillon à une distribution normale. Un W proche de 1 indique que les données suivent une distribution normale, tandis qu'une valeur inférieure à 1 suggère un écart par rapport à la normalité.
- **V** : La statistique de variance utilisée dans le calcul de W.
- **z** : La statistique standardisée associée à la statistique W. Une valeur de z élevée (en termes absolus) indique que l'écart par rapport à la normalité est statistiquement significatif.
- **p-value** : La p-value associée au test. Elle permet de déterminer si l'hypothèse nulle (H_0) peut être rejetée. Si la p-value est inférieure à un seuil donné (généralement 0.05), l'hypothèse nulle est rejetée, indiquant que les résidus ne suivent pas une distribution normale.

Un deuxième test est effectué afin de confirmer les résultats de ce dernier.

4.4.4 Test de Kolmogorov-Smirnov :

Smaller group	D	P-value	Corrected
res	0.5149	0.000	
Cumulative	-0.1587	0.000	
Combined K-S	0.5149	0.000	0.000

TABLE 12 – Test de lien *Log*, Modèle 2

Composantes du test :

Le test de Kolmogorov-Smirnov (K-S) permet de tester l'adéquation d'un échantillon à une distribution théorique (ici la normale). Ce test fournit plusieurs résultats importants :

— **D** : La statistique D est la plus grande différence entre les valeurs observées et la distribution théorique cumulative. Elle mesure l'écart entre les données et la distribution normale. Plus D est grand, plus l'écart entre les données et la distribution théorique est important.

— **P-value** : La p-value associée au test de Kolmogorov-Smirnov. Elle indique la probabilité que l'écart observé entre les données et la distribution théorique soit dû au hasard. Si la p-value est inférieure à un seuil critique (généralement 0.05), l'hypothèse nulle (H_0) est rejetée, ce qui suggère que les données ne suivent pas la distribution théorique (dans ce cas, la normale).

— **Corrected** : La statistique corrigée prend en compte certaines variations dans la méthode de calcul du test, notamment pour des échantillons de grande taille. Elle fournit une version ajustée de la statistique D qui pourrait être plus précise dans certains cas.

— **Cumulative** : Cette statistique représente la différence entre la fonction de répartition empirique cumulée des données et la fonction de répartition cumulative de la distribution théorique.

— **Combined K-S** : La statistique combinée de Kolmogorov-Smirnov, qui résume à la fois la statistique D et la p-value dans un test combiné. C'est une mesure globale de l'adéquation de l'échantillon à la distribution théorique.

Les mêmes résultats que pour le test de **Shapiro-Wilk** sont obtenus, la **non-normalité des résidus est constatée**.

La **P-value** étant inférieure au seuil critique $alpha = 5\%$, on rejette l'hypothèse H_0 de **normalité des résidus**.

5 Discussion : Comparaison avec la littérature :

Afin que nos résultats soient confrontés aux conclusions de la littérature existante sur les coûts moyens des sinistres en France, plusieurs critères d'analyse ont été utilisés. La fréquence des différents types de sinistres et leur impact financier (coût moyen par type) ont été comparés.

5.1 Méthodologie de comparaison :

Pour ce faire, la fréquence relative a été calculée, pour chaque type de sinistre, en fonction du nombre total de sinistres sur la période de *1995* à *2021*. La proportion du coût total attribuée à chaque type de sinistre a également été calculée, afin de mieux comprendre l'impact financier de chaque sinistre par rapport au coût global des sinistres.

Fréquence des sinistres :

La **fréquence** de chaque type de sinistre est donnée par **le nombre total de sinistres observés** pour chaque type, exprimé en **pourcentage du nombre total de sinistres**.

Proportion du coût total (ou fréquence des coûts) :

La **proportion** du coût moyen par type de sinistre est obtenue en divisant le **coût moyen total de chaque sinistre** par le **coût global total des sinistres** sur la même période.

Cela permet de calculer la **répartition du coût total** pour chaque type de sinistre et de comparer notre analyse à celle présentée dans la littérature, qui met en lumière des variations notables selon la fréquence et l'intensité des différents types de sinistres.

5.2 Mouvement de terrain :

Le coefficient des mouvements de terrain est de *0,2602*, ce qui représente une **augmentation de *29,7%* du coût moyen par commune lorsqu'un mouvement de terrain se produit**. Bien que les mouvements de terrain soient relativement rares (*17,1%* de la fréquence totale des sinistres), leur impact financier est considérable, représentant *7,41%* du coût total des sinistres.

Ces résultats sont en adéquation avec la littérature qui décrit les mouvements de terrain comme des sinistres rares mais à fort impact. En effet, bien que leur fréquence soit faible, leur coût par sinistre est élevé, car les réparations de fondations et d'infrastructures peuvent s'avérer très coûteuses.

5.3 Inondations :

Le coefficient des inondations est de *0,1019*, soit une **augmentation** de *10,7%* **du coût moyen par commune** lorsque ce type de sinistre survient. Les inondations représentent *68,1%* de la fréquence totale des sinistres, ce qui en fait le sinistre le plus fréquent. En termes de coût, elles représentent *41,3%* du coût total des sinistres.

Cela est cohérent avec la littérature qui mentionne que les inondations, bien qu'elles soient fréquentes, peuvent entraîner des coûts élevés localement, surtout dans les zones urbaines densément peuplées. Le coût moyen par sinistre est moins élevé que celui des mouvements de terrain ou de la sécheresse, mais le grand nombre de sinistres entraîne un coût total important.

5.4 Sécheresse :

Le coefficient de la sécheresse est de *0,8856* ; soit une **augmentation** de *142,4%* **du coût moyen par commune**. Bien que la sécheresse soit moins fréquente (*14,5%* de la fréquence totale des sinistres), elle génère des **coûts très élevés**, représentant *50,7%* du coût total des sinistres.

Cela correspond bien à la littérature, qui décrit la sécheresse comme un sinistre moins fréquent mais avec un impact structurel très lourd, notamment sur les fondations et réseaux souterrains, ce qui en fait un sinistre particulièrement coûteux à réparer.

5.5 Séisme :

Le coefficient des séismes est de *0,9497*, soit une **augmentation** de *158,5%* du **coût moyen par commune.** Bien que les séismes soient extrêmement rares (représentant *0,0032 %* de la fréquence totale des sinistres), leur impact financier par événement est élevé. Cependant, en termes de coût total, les séismes représentent seulement *0,57%* du coût global des sinistres.

Cela peut sembler légèrement en décalage par rapport à la littérature, qui décrit les séismes comme des événements rares mais avec un **fort potentiel de dégâts locaux.** La différence réside probablement dans le fait que, dans notre base de données, **le nombre de séismes enregistrés est relativement faible** (630 sinistres seulement), ce qui **limite l'échantillon et donc la précision de l'estimation du coût total associé.** En outre, la répartition des coûts peut être influencée par l'absence de sinistres importants dans certaines zones géographiques ou par des événements de moindre magnitude qui n'ont pas généré de coûts extrêmement élevés. La littérature, quant à elle, repose sur des bases de données plus complètes ou des événements plus marquants, ce qui peut expliquer les différences observées.

Les résultats confirment ainsi les modèles économiques et actuariels utilisés dans la littérature pour évaluer les coûts des sinistres en France et soulignent la nécessité d'adopter des modèles différenciés en fonction des types de sinistres pour mieux appréhender leur fréquence et leur coût.

6 Limites de l'étude :

Utilisation des coûts moyen par commune :

Dans cette étude, les coûts par commune étaient initialement fournis sous forme d'intervalles. Il a été fait le choix d'utiliser la moyenne de chaque intervalle pour chaque commune afin d'avoir seulement une valeur par commune. Toutefois, cela a eu pour conséquence de lisser les coûts, ce qui a donc réduit la précision des données. Ce traitement de données a empêché une prise en compte exacte des variations spécifiques à chaque sinistre. La variable dépendante aurait pu être plus précise si l'accès aux coûts exacts avait été possible.

Période d'étude (1995-2023) :

L'étude couvre la période de 1995 à 2023, cependant, l'agrégation des données sur cette période à eu pour conséquence de lisser les variations annuelles. Une analyse par année aurait permis de mieux comparer les coûts entre les différentes années, d'observer des tendances spécifiques et de voir comment les événements exceptionnels ou les évolutions climatiques ont influencé les coûts au fil du temps.

De plus, cette étude analyse portait sur les coûts moyens des sinistres sur une période de 28 ans, mais il est possible que des événements extrêmes aient eu un impact disproportionné sur les résultats. Ces événements ont potentiellement pu fausser les résultats en raison de leur forte intensité et de leur caractère exceptionnel.

Exclusion d'autres types de sinistres :

Le format des données de certains types de sinistres a limité l'inclusion de données supplémentaires dans notre étude. Par exemple, les TGN (Troubles de Grande Nature) étaient disponibles uniquement au niveau départemental, et non au niveau communal. Cette limitation des données a restreint l'analyse des sinistres dont les informations étaient disponibles à l'échelle communale, excluant ainsi un éventail plus large de sinistres qui auraient pu enrichir l'étude.

Manque de données sur les déterminants des coûts :

Une autre limitation majeure de cette étude a été le manque de données suffisantes pour analyser les déterminants spécifiques des coûts des sinistres. En raison de problèmes de disponibilité et de format des données, il a été nécessaire de se concentrer sur l'analyse des coûts moyens par sinistre. Cette absence de données détaillées a limité la portée de l'analyse.

Choix du modèle et du lien logarithmique :

Le modèle utilisé repose sur un lien logarithmique, qui, au regard des données disponibles, est celui qui offre les meilleurs résultats en minimisant les critères statistiques tels que l'**AIC**, le **BIC** et la **déviance**. Cependant, il est important de rester humble dans le choix attendu, car le lien logarithmique n'est pas parfaitement adapté à toutes les situations. Bien qu'il soit le plus pertinent parmi les options disponibles, il peut y avoir des modèles alternatifs qui conviendraient mieux si davantage de données étaient

accessibles. Toutefois, au regard des données dont nous disposions, ce modèle reste celui qui a fourni les meilleures performances.

Impact de la variabilité régionale :

Bien que l'analyse se soit concentrée sur les données communales, il est possible que les caractéristiques régionales (telles que le niveau de développement économique ou encore les infrastructures de protection contre les catastrophes naturelles) aient eu un effet significatif sur les coûts des sinistres. Ces facteurs régionaux ont été insuffisamment explorés et pourraient expliquer des variations significatives qui n'ont pas été prises en compte dans le modèle.

7 Conclusion :

Ce projet avait pour objectif de mieux comprendre les mécanismes sous-jacent à l'augmentation générale des primes d'assurance dans un contexte où le changement climatique remet en question les cadres d'analyse traditionnels du risque. L'intensification attendue des événements extrêmes complexifie la tâche des assureurs, des chercheurs et des collectivités. Les zones, jusqu'à présent modérément exposées deviennent progressivement vulnérables tandis que la sinistralité s'amplifie.

En mobilisant des données issues du régime CatNat, nous avons mis en évidence des différences marquées entre les types de sinistres, tant en termes de fréquence que de gravité unitaire. L'utilisation de modèles économétriques, notamment les GLM avec fonction de lien logarithmique et distribution Gamma, s'est révélée pertinente pour traiter des montants de sinistres positifs et fortement dispersés. Toutefois, ces modèles ont étés construits à partir de moyennes communales, ce qui a nécessité un lissage des coûts et limité la prise en compte de spécificité locales. De plus, en l'absence de données précises concernant certains types de sinistres, certaines variables n'ont pu être intégrées.

Comprendre l'évolution des primes d'assurance dans ce contexte incertain, c'est aussi poser une question plus large : comment faire coexister un modèle économique viable pour les assureurs avec une protection raisonnable pour les assurés ? Ce projet s'est inscrit dans cette réflexion en partant d'une hypothèse simple : si certains types de sinistres pèsent plus lourdement sur le régime assurantiel, alors une analyse fine de leur coût pourrait permettre de repenser les mécanismes de tarification. Plutôt qu'une augmentation généralisée des primes, il pourrait être envisageable d'adopter une approche plus ciblée, ou même d'imaginer de nouveaux modèles d'assurance, plus adaptés à la nature et à la répartition des risques.

Ce défi dépasse les frontières d'une seule discipline : entre climat, économie, actuariat et politique publique ; il est nécessaire d'intégrer dès maintenant l'incertitude comme une dimension structurelle du risque, et non une anomalie temporaire.

8 Références et Bibliographie

8.1 Sources de données

— Installations industrielles rejetant des polluants — Géorisques

— Inventaire des cavités souterraines — Géorisques

— Retrait-gonflement des argiles — Géorisques

— Mouvements de terrain — Géorisques

— https://www.georisques.gouv.fr/articles-risques/onrn/acceder-aux-indicateurs-sinistralite

8.2 Littérature

— Thèse : HERANVAL, M. (2022). Contributions des données de l'assurance à l'étude des risques naturels: application de méthodes d'apprentissage statistique pour l'évaluation de la nature et du coût des dommages assurés liés aux événements naturels en France. [Sorbonne Université] HAL Thèses.

— Mémoire : EL HASSANI, M. (2017). Modélisation stochastique des inondations en France et application en réassurance. [ENSAE PARISTECH] Institut des Actuaires.

— Thèse : ENJOLRAS, M. (2008). De l'assurabilité des catastrophes naturelles: modélisation et application a l'assurance récolte. Sciences de l'Homme et Société. [Université Montpellier 1] HAL Thèses

— Thèse : BOURGUIGNON, M.(2014). Evénements et territoires - le coût des inondations en France : analyses spatio-temporelles des dommages assurés. [Université Paul Valery - Montpellier III] Economie EauFrance

— Thèse : MAO, Mme. (2019). Estimation des coûts économiques des inondations par des approches de type physique sur exposition. [Université de Lyon] HAL Thèses.

— Mémoire: JERRARI, M.(2017). Modélisation dynamique du coût des inondations historiques en France. [Université Lyon 1] . Institut des Actuaires.

8.3 Logiciels utilisés

— **STATA** : Logiciel avancé pour les analyses statistiques.

— **Python** : Utilisé pour certaines analyses de données et visualisations.

— **Excel** : Utilisé pour toutes la partie traitement de données.

8.4 Références bibliographiques

1 Crawley, M. J. (2015). *Statistics : An Introduction using R.* Wiley.

2 McCullagh, P., Nelder, J. A. (1989). *Generalized Linear Models* (2nd ed.). CRC Press.

3 Fox, J. (2016). *Applied Regression Analysis and Generalized Linear Models* (3rd ed.). SAGE Publications.

4 Agresti, A. (2018). *Statistical Methods for the Social Sciences* (5th ed.). Pearson.

5 Faraway, J. J. (2016). *Extending the Linear Model with R : Generalized Linear, Mixed Effects and Nonparametric Regression Models*. CRC Press.

9 Annexe :

9.1 Modèle 1 :

Regression GLM :

```
. glm moymoy sinmvt sininond sinsech sinseisme, family(poisson) link(log)
note: moymoy has noninteger values

Iteration 0:   log likelihood = -135566.11
Iteration 1:   log likelihood = -135201.41
Iteration 2:   log likelihood = -135201.35
Iteration 3:   log likelihood = -135201.35

Generalized linear models                         No. of obs      =         97
Optimization     : ML                             Residual df     =         92
                                                  Scale parameter =          1
Deviance         =  269236.3766                   (1/df) Deviance =  2926.482
Pearson          =  268659.4062                   (1/df) Pearson  =  2920.211

Variance function: V(u) = u                       [Poisson]
Link function    : g(u) = ln(u)                   [Log]

                                                  AIC             =   2787.76
Log likelihood   = -135201.3452                   BIC             =  268815.5
```

moymoy	OIM Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
sinmvt	-.0004677	2.77e-06	-168.65	0.000	-.0004731	-.0004622
sininond	.0000955	8.08e-07	118.27	0.000	.000094	.0000971
sinsech	.0000346	1.59e-06	21.76	0.000	.0000315	.0000378
sinseisme	-.0001359	.0000242	-5.60	0.000	-.0001834	-.0000883
_cons	10.25764	.0012372	8291.09	0.000	10.25522	10.26007

Statistiques Descriptives :

```
. summarize
```

Variable	Obs	Mean	Std. Dev.	Min	Max
Dpartement	0				
sinmvt	98	697.5408	3424.093	1	34156
moymvt	95	8110.705	4687.385	1250	26458.33
sininond	98	2791.643	13617.81	16	135962
moyinond	98	6241.352	1881.304	3500	17011.23
sinsech	98	592.0306	2927.979	0	29009
moysech	95	12941.02	4156.927	3750	25000
sinseisme	98	12.85714	67.88559	0	630
moyseisme	21	3790.431	2717.139	1250	11875
sintotal	97	2677.297	1622.149	18	7813.287
moymoy	97	28094.75	9516.901	9128.836	68469.56

Matrice de corrélation :

```
. correlate moymoy sinmvt sininond sinsech sinseisme
(obs=97)
```

	moymoy	sinmvt	sininond	sinsech	sinsei~e
moymoy	1.0000				
sinmvt	-0.2181	1.0000			
sininond	0.0902	0.5196	1.0000		
sinsech	0.0432	0.2225	0.3251	1.0000	
sinseisme	0.0504	-0.1266	0.0599	-0.0905	1.0000

Test de VIF :

```
. vif
```

Variable	VIF	1/VIF
sininond	1.51	0.664026
sinmvt	1.42	0.703593
sinsech	1.14	0.880314
sinseisme	1.05	0.951452
Mean VIF	1.28	

9.2 Modèle 2 :

Regression GLM :

```
. glm moymoy sinmvt sininond sinsech sinseisme, family (gamma) link(log)

Iteration 0:   log likelihood = -266384.65
Iteration 1:   log likelihood = -261891.36
Iteration 2:   log likelihood =  -261829.7
Iteration 3:   log likelihood = -261829.49
Iteration 4:   log likelihood = -261829.49

Generalized linear models                         No. of obs      =     34839
Optimization     : ML                             Residual df     =     34834
                                                  Scale parameter =  7.886536
Deviance         =  48340.51262                   (1/df) Deviance =  1.387739
Pearson          =  274719.6032                   (1/df) Pearson  =  7.886536

Variance function: V(u) = u^2                     [Gamma]
Link function    : g(u) = ln(u)                   [Log]

                                                  AIC             =  15.03111
Log likelihood   = -261829.4912                   BIC             = -315970.6
```

moymoy	OIM Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
sinmvt	.2601561	.0200925	12.95	0.000	.2207755	.2995366
sininond	.1019587	.0060458	16.86	0.000	.0901092	.1138081
sinsech	.8855518	.017451	50.75	0.000	.8513484	.9197551
sinseisme	.9496723	.1068574	8.89	0.000	.7402357	1.159109
_cons	5.10792	.0362218	141.02	0.000	5.036927	5.178914

Statistiques Descriptives :

```
. summarize
```

Variable	Obs	Mean	Std. Dev.	Min	Max
sinmvt	34839	.9803955	.6968146	0	20
moymvt	34839	545.7246	3084.524	0	25000
sininond	34839	3.90258	2.674955	0	42
moyinond	34839	3040.307	5327.466	0	25000
sinsech	34839	.8326588	1.777383	0	22
moysech	34839	3732.778	7207.148	0	25000
sinseisme	34839	.0180832	.149305	0	6
moyseisme	34839	42.22997	583.037	0	25000
sintotal	34839	5.733718	3.801208	0	66
commune	0				
code_insee	0				
coutmoyeni~d	34839	2907.697	4816.956	0	20000
coutmoyenmvt	34839	485.7344	2622.314	0	20000
coutmoyens~h	34839	3408.572	6298.305	0	20000
coutmoyens~s	34839	41.51239	554.7071	0	20000
moymoy	34839	1311.94	2368.144	0	15833.33

```
.
```

Matrice de corrélation :

```
. correlate moymoy sinmvt sininond sinseisme sinsech
(obs=34839)
```

	moymoy	sinmvt	sininond	sinsei~e	sinsech
moymoy	1.0000				
sinmvt	0.1739	1.0000			
sininond	0.2800	0.2058	1.0000		
sinseisme	0.0438	-0.0487	0.0620	1.0000	
sinsech	0.6620	0.1340	0.2624	-0.0125	1.0000

Test de VIF :

```
. vif
```

Variable	VIF	1/VIF
sininond	1.12	0.896094
sinsech	1.08	0.923909
sinmvt	1.06	0.947269
sinseisme	1.01	0.991606
Mean VIF	1.07	

9.3 Tests Modèle 2 :

Test d'hétéroscédasticité :

```
. predict mu, mu
(210 missing values generated)

. predict res, r
res already defined
r(110);

. gen res2 = res^2
(210 missing values generated)

. gen mu2 = mu^2
(210 missing values generated)

. reg res2 mu2
```

Source	SS	df	MS	Number of obs =	34839
Model	125.885919	1	125.885919	F(1, 34837) =	0.06
Residual	70396465.6	34837	2020.73846	Prob > F =	0.8029
				R-squared =	0.0000
				Adj R-squared =	-0.0000
Total	70396591.5	34838	2020.68407	Root MSE =	44.953

res2	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
mu2	-5.86e-23	2.35e-22	-0.25	0.803	-5.19e-22	4.02e-22
_cons	7.885929	.2408456	32.74	0.000	7.413864	8.357994

Test de Shapiro-Wilk :

```
. swilk res

                    Shapiro-Wilk W test for normal data

    Variable |     Obs       W           V         z       Prob>z
-------------+-------------------------------------------------------
         res |   34839     0.60481    5527.093    23.743    0.00000
```

Test de Kolmogorov-Smirnov :

```
. ksmirnov res = normal(res)

One-sample Kolmogorov-Smirnov test against theoretical distribution
           normal(res)

 Smaller group       D       P-value  Corrected
 ---------------------------------------------
 res:                 0.5149    0.000
 Cumulative:         -0.1587    0.000
 Combined K-S:        0.5149    0.000      0.000

Note: ties exist in dataset;
      there are 2608 unique values out of 34839 observations.
```

Test de lien :

```
.
. linktest

Iteration 0:   log likelihood = -307626.94

Generalized linear models                         No. of obs      =      34839
Optimization     : ML                             Residual df     =      34836
                                                  Scale parameter =    2736327
Deviance         =  9.53227e+10                   (1/df) Deviance =    2736327
Pearson          =  9.53227e+10                   (1/df) Pearson  =    2736327

Variance function: V(u) = 1                       [Gaussian]
Link function    : g(u) = u                       [Identity]

                                                  AIC             =   17.66009
Log likelihood   = -307626.9394                   BIC             =   9.53e+10
```

moymoy	OIM Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_hat	2439.969	24.56247	99.34	0.000	2391.828	2488.111
_hatsq	-82.96222	1.324835	-62.62	0.000	-85.55885	-80.3656
_cons	-10821.14	102.4506	-105.62	0.000	-11021.94	-10620.34