# Projet final réalisé dans le cadre du Master 2 ISIFAR

Université Paris Cité

pour l'évaluation du module d'Apprentissage Statistique

autour de la tarification automobile

Rendu le 18/11/2025

**Par :**
Rabarison Junior
Mnamdji Sultana
Daumont Jean-Baptiste

**Titre : Tarification automobile par méthodes d'apprentissage statistique**

# Avant-propos

Notre travail consiste à appliquer des méthodes d'apprentissage statistique sur des données réelles de tarification automobile. Les outils utilisés sont ce qui a été vu en cours, complétés lorsque c'est nécessaire avec des modèles des mêmes familles adaptés aux modèles d'assurance.

L'étude est basée sur un jeu de données précedentes regroupant des informations détaillées sur les conducteurs, les véhicules, l'exposition au risque, le nombre de sinistres et le montant des sinistres observés.Notre but consiste à estimer la prime pure, définie comme l'espérance du coût des sinistres, en modélisant d'une part leur fréquence, et de l'autre leur sévérité.

Le rapport est divisé en plusieurs parties : préparation et structuration des données, modélisation de la fréquence des sinistres, modélisation du coût individuel des sinistres, reconstruction de la prime pure, puis analyse de la cohérence des résultats obtenus. Chaque étape présente le modèle retenu, les motivations de son utilisation, ses limites, et, si c'est le cas, les raisons menant vers un modèle plus adapté.

# Contents

**Avant-propos** **2**

**1. Introduction** **5**

1.1 Tarification automobile et prime pure . . . . . 5

1.2 Problématique, objectifs et démarche . . . . . 5

**2. Présentation des données** **6**

2.1 Description des variables disponibles . . . . . 6

2.2 Préparation et nettoyage des données . . . . . 6

Vérifications des valeurs manquantes et aberrantes . . . . . 6

Contrôles sur les variables explicatives . . . . . 7

Conversion des types . . . . . 7

Fusion des bases pour la sévérité . . . . . 7

Visualisations de contrôle . . . . . 7

2.3 Analyse descriptive . . . . . 7

Statistiques descriptives générales . . . . . 7

Distributions clés . . . . . 8

Analyse univariée de la sinistralité . . . . . 10

**3. Modélisation de la fréquence des sinistres** **13**

3.1 Classification . . . . . 13

Régression logistique . . . . . 13

Modèle de boosting (XGBoost) . . . . . 14

Limites empiriques des modèles de classification . . . . . 15

3.2 Modèles de comptages . . . . . 15

Régression de Poisson . . . . . 15

Régression binomiale négative . . . . . 16

Comparaison des modèles de comptage . . . . . 16

**4. Modélisation du coût des sinistres (sévérité)** **18**

4.1 Point de départ : régression linéaire sur le log-coût . . . . . 18

Résultats globaux . . . . . 19

Analyse graphique . . . . . 19

Conclusion interne . . . . . 22

4.2 Modèle Gamma . . . . . 23

Spécification . . . . . 23

Résultats globaux . . . . . 23

Analyse graphique . . . . . 24
Conclusion interne . . . . . 26
4.3 Modèle Log-normal . . . . . 27
Spécification . . . . . 27
Résultats globaux . . . . . 27
Analyse graphique . . . . . 27
Conclusion interne . . . . . 30
4.5 Conclusion générale sur la modélisation de la sévérité . . . . . 30

**5. Estimation de la prime pure** **31**
5.1 Modèles retenus . . . . . 31
Fréquence Modèle retenu : binomiale négative (offset log-Exposure) . . . . . 31
Sévérité Modèle retenu : Gamma (lien log) . . . . . 31
5.2. Construction de la prime pure prédite . . . . . 31
Prédiction sur l'échantillon test . . . . . 32
5.3. Calibration globale . . . . . 32
5.4. Analyse de cohérence et limites du modèle . . . . . 33
5.5. Conclusion tarifaire . . . . . 33

**6. Conclusion générale** **33**

# 1. Introduction

## 1.1 Tarification automobile et prime pure

La tarification automobile consiste à déterminer le montant que chaque assuré doit payer pour être couvert contre les risques liés à l'usage de son véhicule. Cette tarification suit une règle : le montant de la prime doit être représentatif du niveau de risque que représente l'assuré. Pour y arriver, les compagnies d'assurance utilisent des données qui décrivent les caractéristiques des conducteurs et des véhicules, l'exposition au risque et les sinistres observés au cours des périodes précédentes.

La tarification sépare traditionnellement deux composantes du risque : la probabilité qu'un sinistre se produise et le coût attendu de ce sinistre quand celui-ci se présente. Ces deux éléments, la fréquence et la sévérité, structurent l'ensemble de la démarche actuarielle. Leur combinaison permet d'estimer la prime pure, définie comme le coût moyen attendu des sinistres pour un assuré donné.

Le fait qu'il s'agisse de données individuelles conduit des methodes liées à l'utilisation de modèles d'apprentissage supervisé. Celles-ci permettent de capter des structures plus précises, de comparer différentes familles de modèles et d'améliorer la précision des estimations. Dans ce cadre, l'étude de données automobiles est un environnement pertinent pour appliquer les méthodes étudiées en TP et évaluer leur capacité à prédire le risque assuré.

La prime pure est représentée par l'espérance du coût des sinistres causés par un assuré sur une période donnée. La prime pur ne couvre que le risque technique, sans inclure les frais, les taxes ou les marges.

Elle se formule comme le produit entre deux composantes distinctes :

- la fréquence, qui mesure le nombre d'apparition des sinistres,
- la sévérité, qui mesure le coût moyen d'un sinistre quand il se passe.

Cette grandeur constitue la base du calcul tarifaire : elle représente le niveau de risque que l'assureur doit financer avant toute majoration opérationnelle ou réglementaire.

## 1.2 Problématique, objectifs et démarche

L'enjeu du projet est d'estimer le risque associé à chaque assuré à partir des données choisies, en construisant un modèle capable de prédire les conditions des sinistres et leur coût attendu. La question principale est d'utiliser des méthodes statistiques adaptées à la structure des données pour obtenir une estimation fiable de la prime pure.

La démarche consiste à faire une séparation entre les deux composantes du risque. La fréquence des sinistres est d'abord modélisée avec des méthodes adaptées aux variables de comptage, puis la sévérité est modélisée avec des modèles conçus pour des coûts positifs et asymétriques. Plusieurs approches sont testées, comparées et validées sur un échantillon indépendant qui sera tiré pour mesurer leur convenance.

Les prédictions obtenues des deux parties sont ensuite combinées pour produire l'estimation finale de la prime pure. L'organisation du rapport suit le schéma suivant : préparation des données, modélisation et évaluation de la fréquence, modélisation et évaluation de la sévérité, puis combinaison des deux modèles les plus cohérents et ayant donné les meilleurs résultats vis-à-vis de chaque composante.

# 2. Présentation des données

## 2.1 Description des variables disponibles

Les variables ont été décrites à partir de la documentation fournie avec le jeu de données Kaggle freMTPL, qui précise pour chaque variable son rôle, son unité et son mode de construction. Les tableaux suivants représentent l'ensemble des informations qui seront utiles pour l'analyse de la fréquence et de la sévérité.

#### La fréquence des sinistres

| Variable | Type | Description | Rôle |
|---|---|---|---|
| PolicyID | Entier | Identifiant du contrat, utilisé pour relier fréquences et sévérités. | Clé de liaison |
| ClaimNb | Entier | Nombre de sinistres observés pendant: la période d'exposition. | Cible (fréquence) |
| Exposure | Numérique | Durée d'exposition au risque, en années. | Offset |
| Power | Catégorie | Puissance du véhicule, regroupée en classes ordinales. | Variable explicative |
| CarAge | Entier | Âge du véhicule, en années. | Variable explicative |
| DriverAge | Entier | Âge du conducteur, en années. | Variable explicative |
| Brand | Catégorie | Groupe constructeur du véhicule. | Variable explicative |
| Gas | Catégorie | Type de carburant (diesel ou essence). | Variable explicative |
| Region | Catégorie | Région administrative du contrat (classification France 1970–2015). | Variable explicative |
| Density | Numérique | Densité de population de la commune, en habitants par km². | Variable explicative |

#### La sévérité des sinistres

| Variable | Type | Description | Rôle |
|---|---|---|---|
| PolicyID | Entier | Identifiant du contrat, utilisé pour relier le sinistre aux caractéristiques. | Clé de liaison |
| ClaimAmount | Numérique | Montant du sinistre, évalué à une date récente. | Cible (sévérité) |

## 2.2 Préparation et nettoyage des données

Nous avons importé les deux jeux de données puis les avons soumis à une série de contrôles élémentaires dans le but de vérifier leur cohérence avant leur modélisation.

#### Vérifications des valeurs manquantes et aberrantes

Il n'y a aucune case vide détectée dans les tables freMTPLfreq et freMTPLsev. Les variables quantitatives Exposure, ClaimNb et ClaimAmount respectent leurs domaines :

- Exposure $> 0$ toujours vérifié
- ClaimNb $\geq 0$ et entier
- ClaimAmount $\geq 0$
- Aucun sinistre déclaré lorsque Exposure $= 0$
- Aucun coût de sinistre lorsque ClaimNb $= 0$

Il n'y a aucun doublon dans la colonne PolicyID.

#### Contrôles sur les variables explicatives

- Age conducteur : entre 18 et 99
- Age véhicule : entre 0 et 100, les valeurs supérieures à 50 sont rares mais cohérentes
- Densité de population : distribution fortement asymétrique, concentration dans les zones de faible densité

#### Conversion des types

Les variables qualitatives Power, Brand, Gas et Region sont converties en facteurs pour leur utilisation dans les modèles.

#### Fusion des bases pour la sévérité

La fusion de freMTPLsev avec freMTPLfreq via PolicyID ne met en lumière aucune incohérence entre la présence de sinistre et du coût positif.

#### Visualisations de contrôle

Les visualisations confirment les caractéristiques suivantes :

- ClaimNb : la proportion de zéros est dominante, les valeurs $> 1$ très rares
- ClaimAmount (log-transformé) : la distribution unimodale et fortement asymétrique avant transformation
- Exposure : forte concentration autour d'une année d'exposition

Un échantillon réduit contenant 50 000 observations est extrait pour permettre des tests rapides de modèles.

### 2.3 Analyse descriptive

#### Statistiques descriptives générales

Taille du portefeuille : environ 412 000 contrats en fréquence et 16 000 sinistres en sévérité.
Exposition moyenne proche d'une année, indiquant des contrats annuels standardisés.
Proportion de contrats sinistrés très faible : plus de 96 % des assurés ne déclarent aucun sinistre sur la période observée.

Fréquence brute de sinistre obtenue par le ratio

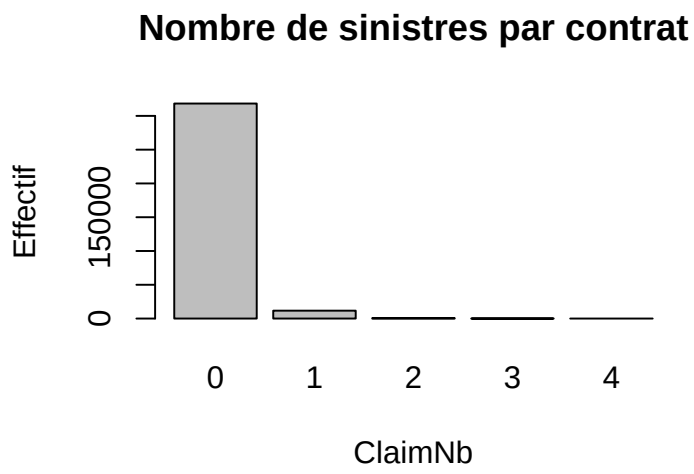
$$\frac{\sum \text{ClaimNb}}{\sum \text{Exposure}},$$

faible et cohérente avec un portefeuille automobile classique.

Montants de sinistres caractérisés par une forte asymétrie : médiane autour du millier d'euros et queue de distribution étendue, indiquant une sévérité lourde.

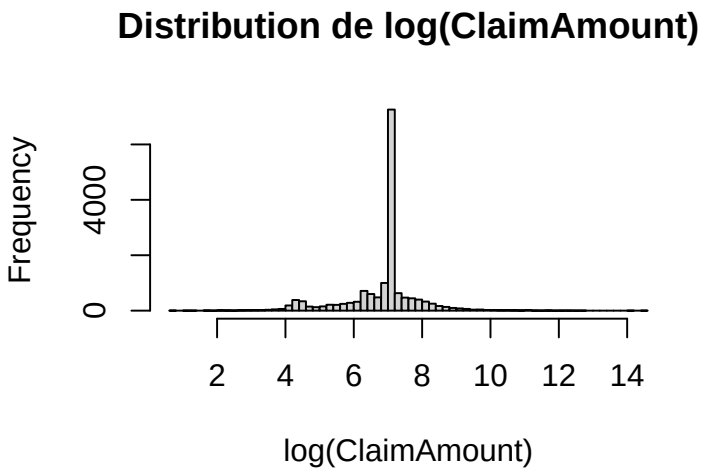
**Distributions clés**

```r
barplot(
  table(df_freq$ClaimNb),
  main = "Nombre de sinistres par contrat",
  xlab = "ClaimNb",
  ylab = "Effectif"
)
```



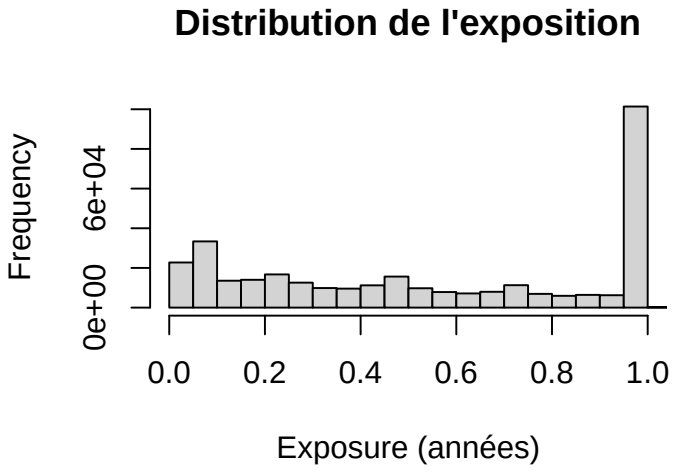

La distribution est dominée par la modalité nulle : la quasi-totalité des contrats ne déclarent aucun sinistre. Ceux declarant pluiseurs sinistres sont très rares. Cette structure renforce l'idée d'un processus de comptage avec forte masse

```r
hist(log(df_sev$ClaimAmount),
     breaks = 50,
     main = "Distribution de log(ClaimAmount)",
     xlab = "log(ClaimAmount)")
```




La distribution du logarithme des montants est unimodale mais très étirée. L'asymétrie marquée et la présence d'une queue lourde témoignent de sinistres de grande ampleur, confirmant la pertinence d'un modèle conçu pour des données positives fortement dispersées

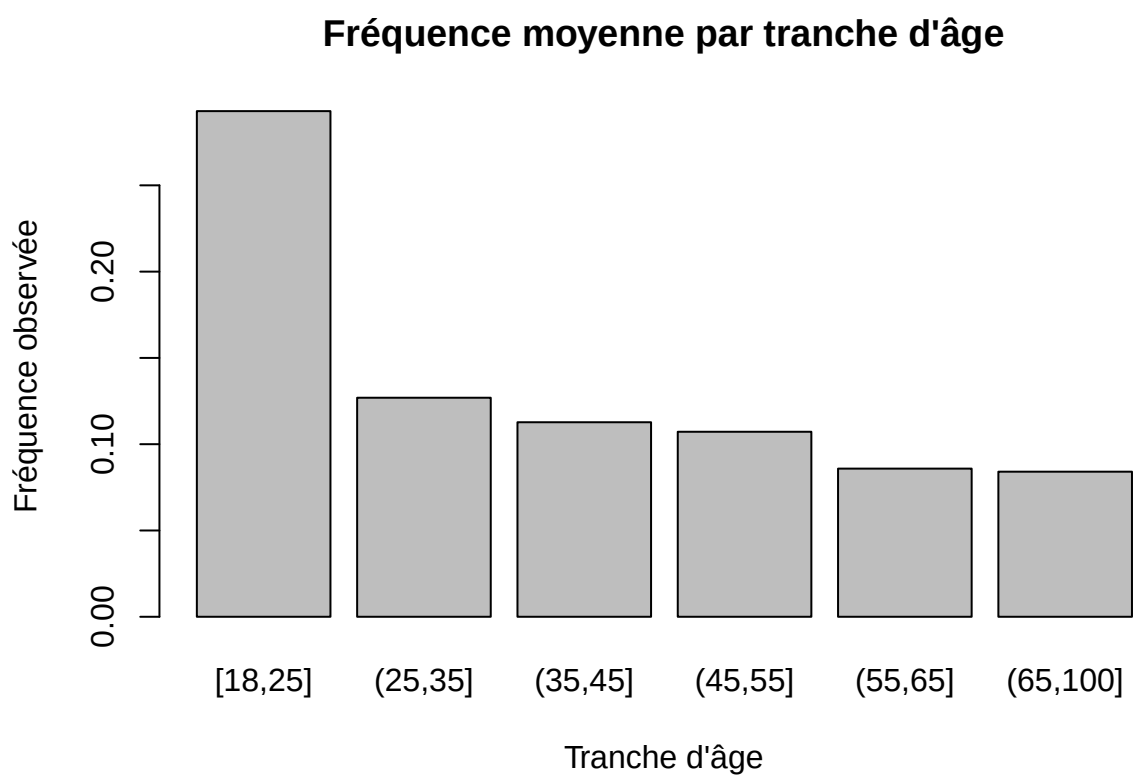
```
hist(df_freq$Exposure,
     breaks = 30, xlim = c(0, 1),
     main = "Distribution de l'exposition",
     xlab = "Exposure (années)")
```




La variable est principalement concentrée autour d'une année d'assurance, avec quelques durées partielles. La disposition standard attendue d'un portefeuille automobile est confirmée.

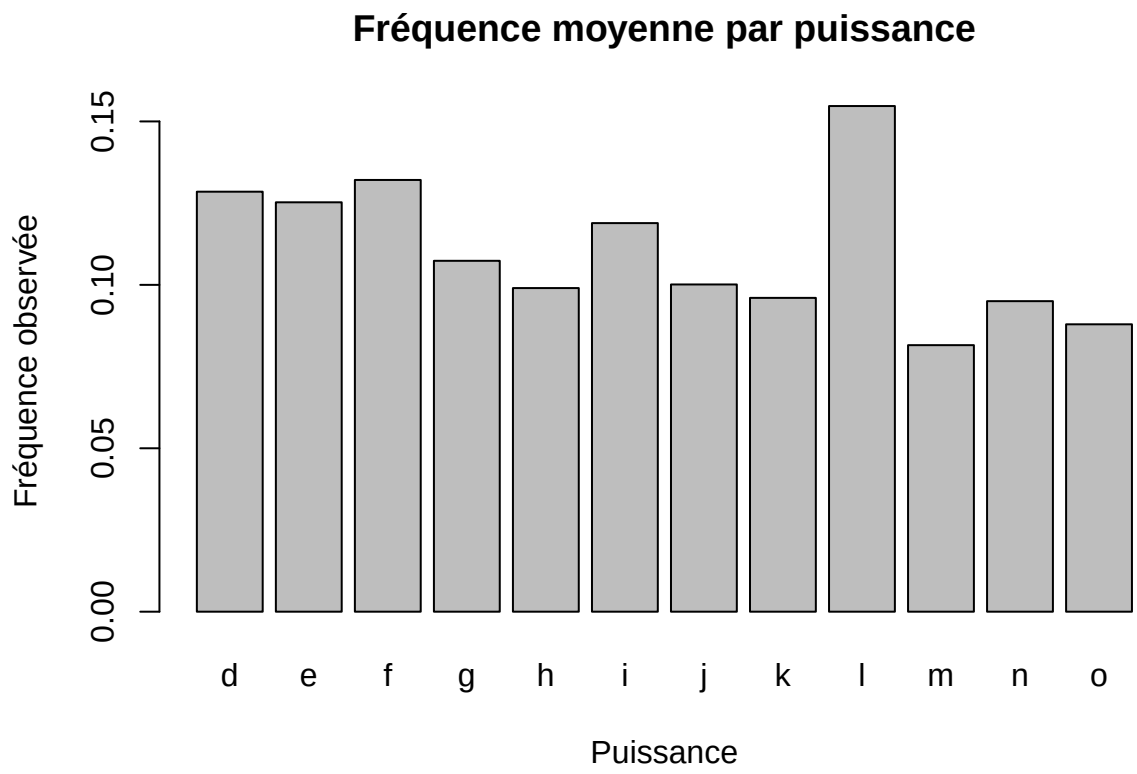
**Analyse univariée de la sinistralité**

```
df_freq$AgeBand <- cut(df_freq$DriverAge,
                       breaks = c(18,25,35,45,55,65,100),
                       include.lowest = TRUE)

ClaimRate_by_Age <- tapply(df_freq$ClaimNb/df_freq$Exposure,
                           df_freq$AgeBand,
                           mean)

barplot(ClaimRate_by_Age,
        main = "Fréquence moyenne par tranche d'âge",
        xlab = "Tranche d'âge",
        ylab = "Fréquence observée")
```



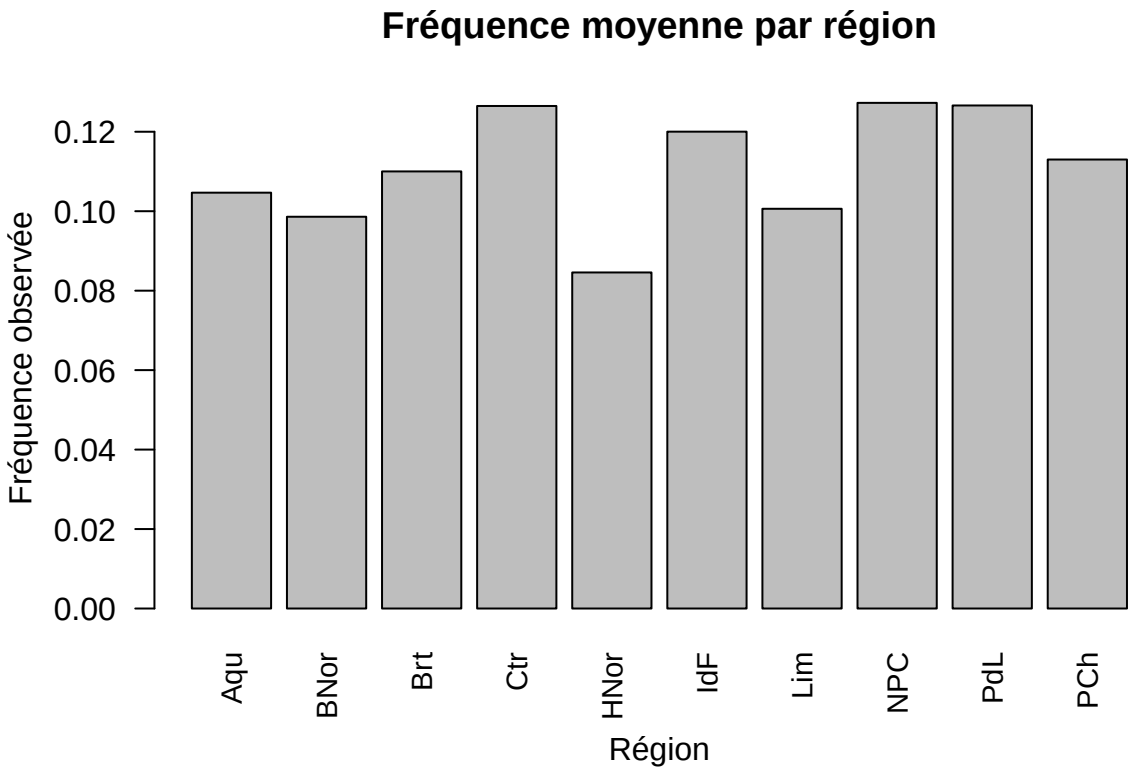

```
ClaimRate_by_Power <- tapply(df_freq$ClaimNb/df_freq$Exposure,
                             df_freq$Power,
                             mean)

barplot(ClaimRate_by_Power,
        main = "Fréquence moyenne par puissance",
```

```
        xlab = "Puissance",
        ylab = "Fréquence observée")
```




```
ClaimRate_by_Region <- tapply(df_freq$ClaimNb/df_freq$Exposure,
                              df_freq$Region,
                              mean)
names(ClaimRate_by_Region) <- c(
  "Aqu",    # Aquitaine
  "BNor",   # Basse-Normandie
  "Brt",    # Bretagne
  "Ctr",    # Centre
  "HNor",   # Haute-Normandie
  "IdF",    # Ile-de-France
  "Lim",    # Limousin
  "NPC",    # Nord-Pas-de-Calais
  "PdL",    # Pays-de-la-Loire
  "PCh"     # Poitou-Charentes
)

barplot(ClaimRate_by_Region,
        main = "Fréquence moyenne par région",
        xlab = "Région",
        ylab = "Fréquence observée",
```

```
        las = 2,
        cex.names = 0.9)
```




La fréquence moyenne des sinistres varie selon plusieurs caractéristiques du portefeuille. Il y a un taux de sinistralité plus élevé chez les plus jeunes, celui-ci qui décroît progressivement avec l'âge avant de se stabiliser. La variable de puissance, bien qu'elle soit codée sous forme de classes anonymisées, révèle également des écarts : certaines catégories présentent des fréquences plus élevées que d'autres, suggérant qu'une partie de l'information liée à la motorisation restent captée malgré l'absence des valeurs réelles des puissances. Enfin, on observe une hétérogénéité géographique avec un niveau de sinistralité plus marqué pour certaines zones densément peuplées. Ces variations motivent l'intégration de l'ensemble de ces variables dans la phase de modélisation.

# 3. Modélisation de la fréquence des sinistres

## 3.1 Classification

La variable ClaimNb est un comptage fortement concentré en zéro, avec une petite proportion de sinistres de valeur 1 et très peu de valeurs supérieures (très rare). La fréquence actuarielle exige un modèle explicitement adapté à ce type de distribution. Cependant avant d'introduire les modèles de comptage, on estime l'occurrence d'au moins un sinistre par des modèles de classification. Cette étape sert uniquement de repère discriminant, la modélisation de la fréquence s'appuie ensuite sur des modèles plus adaptés.

### Régression logistique

On modélise l'occurrence d'un sinistre au moyen d'une régression logistique. La variable cible le est

$$Y_i = \mathbf{1}_{\{\text{ClaimNb}_i > 0\}}, \qquad Y_i \sim \mathcal{B}(p_i),$$

avec lien log :

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + x_i^\top \beta.$$

Le vecteur de valriables explicatives $x_i$ contient notamment :

- $\log(\text{Exposure}_i)$ : un contrat plus long augmente mécaniquement la probabilité d'observer un sinistre,
- les âges du conducteur et du véhicule : $\text{DriverAge}_i,\ \text{CarAge}_i$,
- les caractéristiques techniques du véhicule : $\text{Power}_i,\ \text{Brand}_i,\ \text{Gas}_i$,
- la région et la densité géographique : $\text{Region}_i,\ \log(\text{Density}_i + 1)$.

La forme finale du modèle est donc :

$

$$\begin{aligned}\log\left(\frac{P(Y_i=1)}{1-P(Y_i=1)}\right) &= \beta_0 + \beta_1 \log(\text{Exposure}_i) + \beta_2 \text{DriverAge}_i + \beta_3 \text{CarAge}_i + \beta_4 \text{Power}_i \qquad (1)\\ &\quad + \beta_5 \text{Brand}_i + \beta_6 \text{Gas}_i + \beta_7 \text{Region}_i + \beta_8 \log(\text{Density}_i + 1). \qquad (2)\end{aligned}$$

$ Le jeu de données d'origine est volumineux, alors pour avoir un temps de calcul raisonnable tout en conservant une variabilité répresentative, on utilise un sous-échantillon de 300 000 observations tiré aléatoirement. L'échantillon final utilisé pour l'apprentissage regroupe la variable cible binaire HasClaimFac, l'ensemble des variables explicatives sélectionnées, ainsi que leurs versions transformées (logExposure, logDensity).

Estimation du modèle logistique :

L'ajustement est réalisé sur le sous-échantillon aléatoire de 300 000 observations. L'évaluation s'appuie sur une validation croisée 5-fold avec l'AUC ROC comme critère. L'AUC obtenue est 0.64. Le seuil 0.5 conduit à une sensibilité nulle en raison de la rareté des sinistres.

```
auc_logit
```

```
## [1] 0.641
```

Pour obtenir un compromis exploitable entre erreurs de type I et II, la courbe ROC est construite, puis le critère de Youden est appliqué :

$$J(t) = \text{Sensibilité}(t) + \text{Spécificité}(t) - 1,$$

où $t$ désigne le seuil optimal(seuil de décision.

```
youden_logit
```

```
## threshold      sens      spec
##    0.0358    0.7200    0.4800
```

Le seuil optimal est déterminé par le critère de Youden par la maximisation de la fonction. Il est d'environ 0.04. À ce seuil, la sensibilité atteint 0.72 et la spécificité 0.48. Le modèle capte une part notable des sinistres au prix d'un nombre élevé de faux positifs, ce qui est le comportement typique d'un événement rare.

Ces performances serviront de point de comparaison pour l'évaluation d'un modèle plus flexible présenté dans la section suivante.

### Modèle de boosting (XGBoost)

Un modèle de Boosting par arbres est utilisé pour évaluer la potentielle amélioration en comparant avec la structure linéaire de la régression logistique. XGBoost assemble une suite d'arbres peu profonds, chacun ciblant les erreurs résiduelles des précédents, ce qui permet de capter automatiquement non-linéarités et interactions dans les covariables.

Le modèle est entraîné sur le même sous-échantillon que la régression logistique, sous validation croisée 5-fold, avec l'AUC ROC comme critère unique d'optimisation. L'AUC obtenue est :

```
auc_xgb
```

```
## [1] 0.662
```

Cette valeur dépasse celle du modèle logistique. L'amélioration reste toutefois petite : la rareté des sinistres contraint les deux modèles à travailler dans une zone où les probabilités estimées restent très faibles, ce qui réduit l'effet des relation non-linéaires apprises par le boosting.

Le seuil optimal, déterminé par le critère de Youden :

```
youden_xgb
```

```
##   threshold sensitivity specificity
##      0.0377      0.7060      0.5280
```

reste faible ($\approx$ 0.038). À ce seuil, la sensibilité atteint environ 0.71 et la spécificité environ 0.53. Le compromis s'améliore légèrement, mais le comportement global demeure celui d'un problème d'occurrence rare : la détection des sinistres nécessite d'accepter un volume important de faux positifs.

Enfin, le modèle XGBoost fournit une borne supérieure raisonnable pour les modèles d'occurrence. Le modèle améliore un peu sa capacité à distinguer sinistres et non-sinistres.Mais la structure générale des probabilités prédictives ne change pas, car la variable cible est rare et binaire. Le modèle améliore un peu sa capacité à distinguer sinistres et non-sinistres.Mais la structure générale des probabilités prédictives ne change pas, car la variable cible est rare et binaire.

#### Limites empiriques des modèles de classification

Les modèles d'occurrence offrent une indication sur la probabilité qu'un contrat connaisse au moins un sinistre, mais ils sont fondamentalement mal adaptés à l'objectif actuariel de modélisation de la fréquence.

La transformation binaire de ClaimNb supprime l'information sur le nombre de sinistres : un contrat avec plusieurs sinistres est traité comme un contrat avec un seul. L'exposition n'entre pas comme offset, ce qui empêche de prendre en compte la proportion par rapport à la durée d'observation. Enfin, la forte asymétrie des données impose des seuils de décision très faibles, entraînant une instabilité opérationnelle et limitant la portée pratique des scores produits.

Ces limites justifient le passage à des modèles spécifiquement conçus pour les données de comptage, en particulier la régression de Poisson.

## 3.2 Modèles de comptages

Dans la suite de l'analyse, l'objectif n'est plus d'identifier l'arrivée d'un sinistre, mais d'estimer directement le nombre attendu de sinistres par contrat, ce qui correspond à la définition actuarielle de la fréquence. La variable cible retrouve donc sa forme de comptage :

$$Y_i = \text{ClaimNb}_i.$$

#### Régression de Poisson

On modélise le nombre de sinistres par un MLG de Poisson avec lien logarithmique. L'exposition entre comme offset pour assurer la proportionnalité de l'intensité à la durée d'observation :

$$\log(\lambda_i) = \beta_0 + x_i^\top \beta + \log(\text{Exposure}_i), \qquad Y_i \mid X_i \sim \text{Poisson}(\lambda_i).$$

Les covariables couvrent les caractéristiques conducteur, véhicule, région, ainsi que log(Density).

Coefficients :

La structure estimée est stable :
– DriverAge : effet négatif.
– CarAge : effet négatif.
– log(Density) : effet positif.
– Power : plusieurs classes significativement supérieures à la référence.
– Brand (Japanese except Nissan/Korean) : effet négatif.
– Gas = Regular : effet négatif.
– Régions : hétérogénéité significative.

Diagnostics internes :

Null deviance = 105613 : variabilité brute de la fréquence avant modélisation.
Residual deviance = 104315.5 : variabilité restant après ajustement. Diminution mais très limitée, ce qui montre que les covariables n'expliquent qu'une part.
AIC = 135634.6

La surdispersion issue des résidus de Pearson :

$$\hat{\phi} = \frac{1}{n-p} \sum_i \frac{(Y_i - \hat{\mu}_i)^2}{\hat{\mu}_i} \approx 1.76,$$

indique une variance plus grande que celle imposée par la loi de Poisson.

Conclusion interne :

La structure des effets correspond à ce qu'on observe dans la sinistralité, mais la variance dépasse ce que le modèle Poisson suppose. Il faut donc passer à un modèle de fréquence capable de gérer cette surdispersion.

### Régression binomiale négative

On étend le modèle de Poisson en introduisant un paramètre de dispersion $\theta > 0$ pour capturer la variance excédentaire. Le prédicteur linéaire conserve la même structure et le même offset :

$$\log(\mu_i) = \beta_0 + x_i^\top \beta + \log(\text{Exposure}_i), \qquad Y_i \mid X_i \sim \text{NB}(\mu_i, \theta).$$

La variance conditionnelle devient

$$\text{Var}(Y_i \mid X_i) = \mu_i \left(1 + \frac{\mu_i}{\theta}\right),$$

permettant d'élargir la dispersion tout en conservant le lien logarithmique.

Coefficients :

Les effets reproduisent la structure obtenue sous Poisson :
– DriverAge : effet négatif.
– CarAge : effet négatif faible.
– log(Density) : effet positif .
– Power : plusieurs classes significativement positives.
– Brand (Japanese except Nissan/Korean) : effet négatif.
– Gas = Regular : effet négatif.
– Régions : hétérogénéité persistante.

Diagnostics internes :

Null deviance : 90839.5
Residual deviance : 89612.6
AIC : 135240
$\theta \approx 0.82$

La surdispersion issue des résidus de Pearson :

$$\hat{\phi} \approx 1.70,$$

indique une variance élargie mieux captée que sous Poisson.

Conclusion interne :

Le paramètre de dispersion permet de stabiliser l'ajustement et de limiter la surdispersion encore présente dans les résidus. La binomiale négative offre ainsi une meilleure représentation de la fréquence pour un portefeuille où la variance dépasse systématiquement l'espérance conditionnelle.

### Comparaison des modèles de comptage

Les deux modèles sont estimés sur la même base, avec le même ensemble de covariables et le même offset $\log(\text{Exposure}_i)$. Les indicateurs globaux sont :

| **Modèle** | **AIC** | **Deviance résiduelle** | $\hat{\phi}$ | $\theta$ |
|---|---|---|---|---|
| Poisson | 135634.6 | 104315.5 | 1.76 | – |
| Binomiale négative | 135240 | 89612.6 | 1.70 | 0.82 |

Analyse interne :
– L'AIC diminue avec la binomiale négative, montrant un meilleur compromis biais/variance.
– La déviance résiduelle baisse fortement, signe d'un ajustement global plus satisfaisant.
– La surdispersion reste $> 1$ mais est légèrement réduite, et surtout prise en charge explicitement par $\theta$.

**Conclusion** Les deux modèles produisent une structure d'effets similaire, mais la présence du paramètre de dispersion et les indicateurs globaux (AIC, deviance, surdispersion)montrent que la binomiale négative est mieux adaptée.
Le modèle binomial négatif est donc retenu pour la modélisation de la fréquence.

# 4. Modélisation du coût des sinistres (sévérité)

La modélisation de la sévérité vise à estimer le coût moyen conditionnel d'un sinistre positif, c'est-à-dire le montant attendu d'un sinistre dès lors qu'il survient. Cette étape constitue le second pilier de la tarification classique en fréquence–coût : après avoir modélisé le nombre de sinistre, on analyse ici la distribution des montants.

Les données de coût ont généralement des caractéristiques distinctives :
– uniquement des valeurs **strictement positives**.
– une **asymétrie marquée** (par la présence de sinistres lourds).
– une **variance croissante** avec le niveau moyen.
– une dispersion souvent incompatible avec les hypothèses d'un modèle linéaire classique sur le coût brut.

La sévérité est donc traditionnellement modélisée via des transformations (comme le log-coût) ou via des lois spécifiques aux variables positives (Gamma, Log-normal). On procède en plusieurs étapes : extraction des sinistres positifs, analyse du log-coût, diagnostic du modèle linéaire, puis estimation de modèles adaptés.

**Extraction des sinistres positifs**

La sévérité n'est définie que pour les contrats contenant un sinistre. Les contrôles préliminaires confirment que l'ensemble des valeurs de coût est strictement positif, aucune valeur nulle ou négative n'étant présente dans la base. On introduit ensuite la transformation logarithmique :

$$\log(\text{ClaimAmount}_i),$$

afin de préparer l'analyse du modèle linéaire de départ.
Le passage au logarithme n'est pas une hypothèse de modèle mais un outil exploratoire visant à stabiliser la variance et à mettre en évidence la structure des effets.

Les statistiques descriptives illustrent la forte asymétrie du coût brut, atténuée sur l'échelle logarithmique :

- le **ClaimAmount** présente une distribution très étalée, avec des valeurs extrêmes élevées.
- le **logClaimAmount** est recentré, moins asymétrique et davantage compatible avec une approche linéaire.

Cette transformation préparent l'étude du modèle de départ sur le log-coût, qui servira ensuite de point d'appui pour les modèles sur la séverité plus adaptés.

## 4.1 Point de départ : régression linéaire sur le log-coût

L'objectif est d'établir un diagnostic minimal sur le modèle du coût des sinistres s'il y en a. On travaille sur les sinistres positifs et sur la variable transformée :

$$\log(\text{ClaimAmount}_i),$$

dont la distribution est moins étirée que celle des montants bruts mais conserve une asymétrie forte et une queue lourde.

Modèle de référence :

$$\log(\text{ClaimAmount}) = \beta_0 + X^\top \beta + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2).$$

Les variables explicatives sont les mêmes que pour la fréquence : âge conducteur, âge véhicule, puissance, marque, carburant, région et log-densité.

#### Résultats globaux

– Résidu standard $\approx$ **1.11** : dispersion élevée autour du modèle.
– $R^2$ et $R^2_{\text{adj}} < \mathbf{1}$ **%** : pouvoir explicatif quasi nul.
– Coefficients largement non significatifs, instabilité des effets.

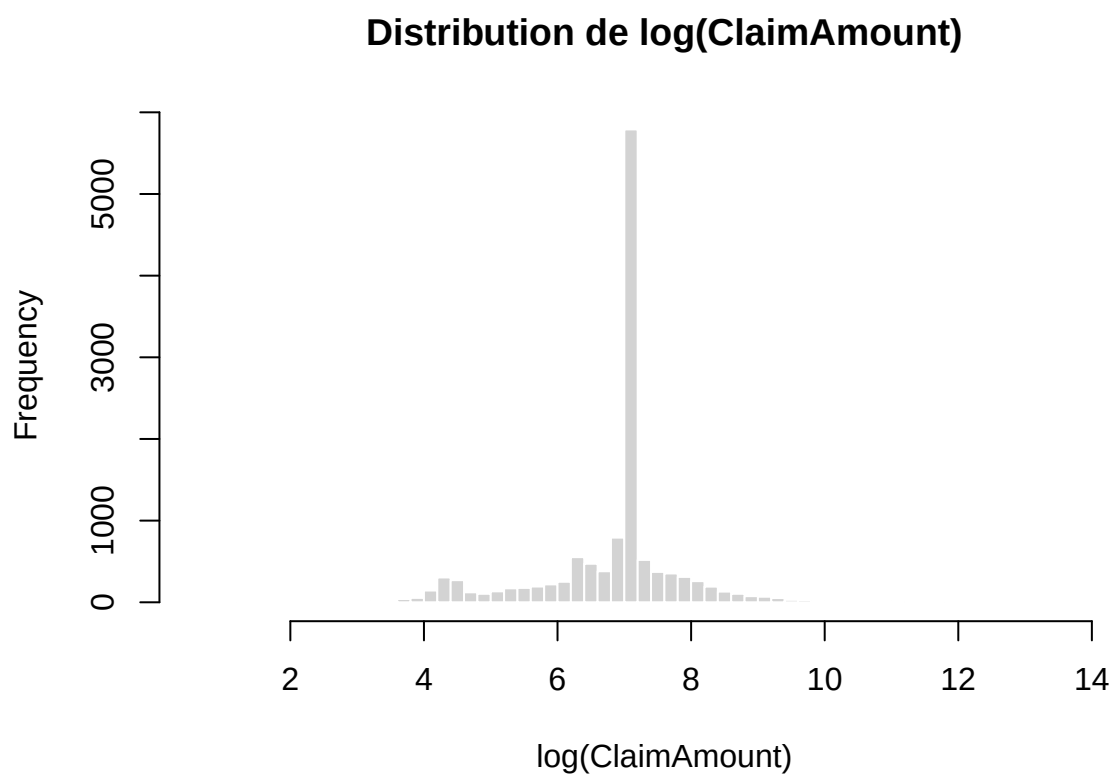
Ces valeurs indiquent que, même après transformation logarithmique, la variabilité du coût individuel n'est presque pas expliquée par les caractéristiques techniques ou géographiques.

#### Analyse graphique

Les graphiques de diagnostic montrent que le modèle linéaire ne fonctionne pas correctement et qu'aucune condition nécessaire n'est respectée.

Distribution du log-coût :

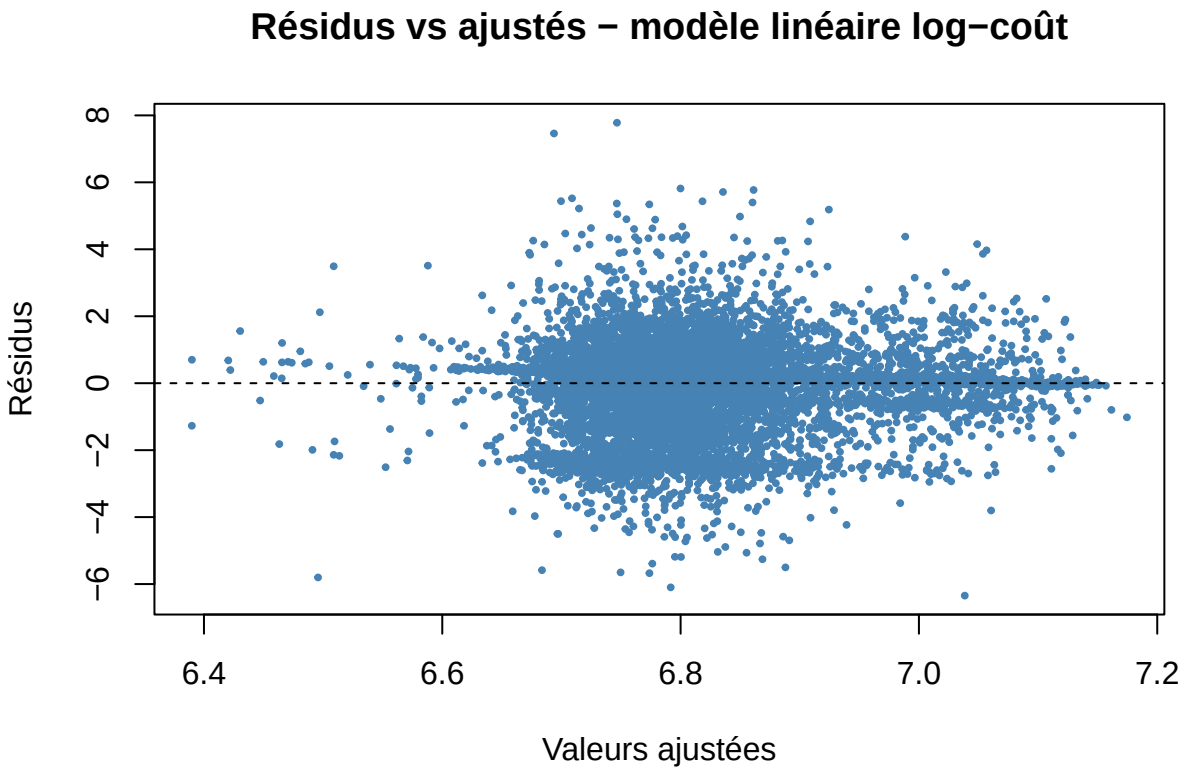
```r
hist(df_sev_pos$logClaimAmount,
     breaks = 50,
     main = "Distribution de log(ClaimAmount)",
     xlab = "log(ClaimAmount)",
     col = "lightgray",
     border = "white")
```




La transformation log réduit l'échelle mais la distribution reste asymétriquement marquée et une queue lourde persistent : la distribution reste clairement non-gaussienne.

Résidus vs valeurs ajustées

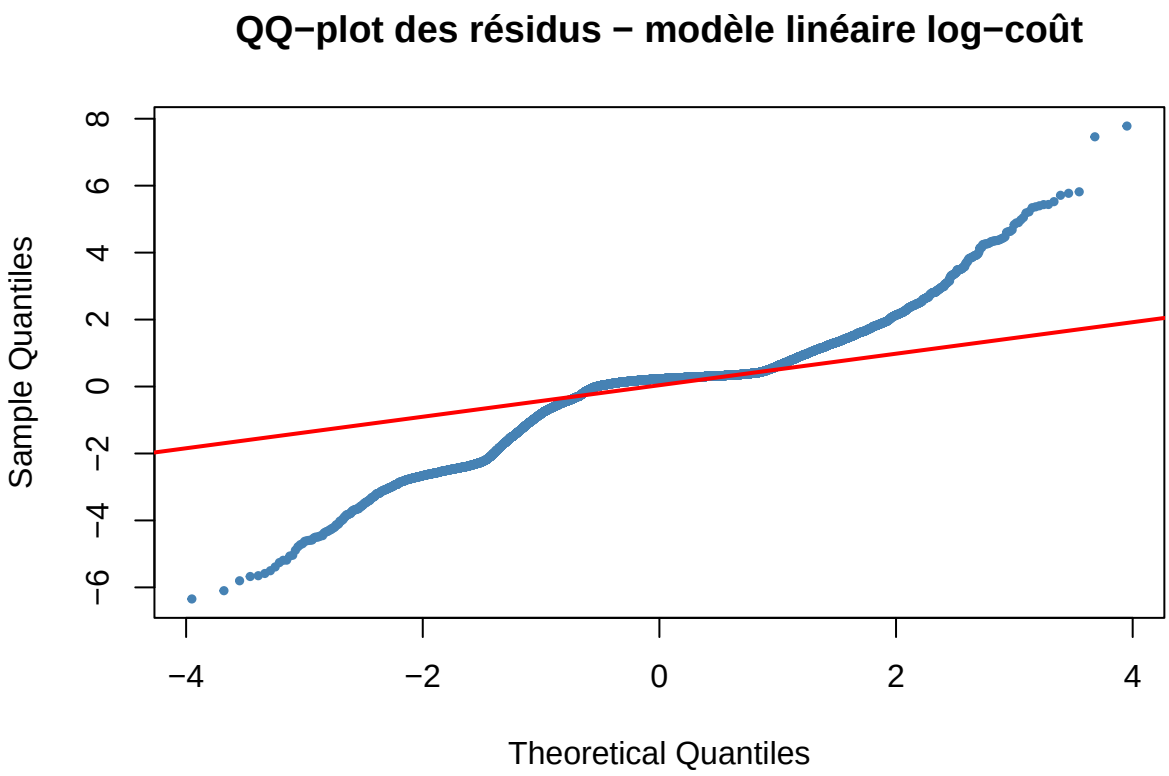
```r
plot(
  fitted(lm_sev), resid(lm_sev),
  xlab = "Valeurs ajustées",
  ylab = "Résidus",
  main = "Résidus vs ajustés - modèle linéaire log-coût",
  pch = 19, cex = 0.4, col = "steelblue"
)
abline(h = 0, lty = 2)
```




Hétéroscédasticité marquée : l'amplitude des résidus varie clairement selon les valeurs ajustées, signalant une rupture stable de l'hypothèse de variance constante.

QQ-plot des résidus

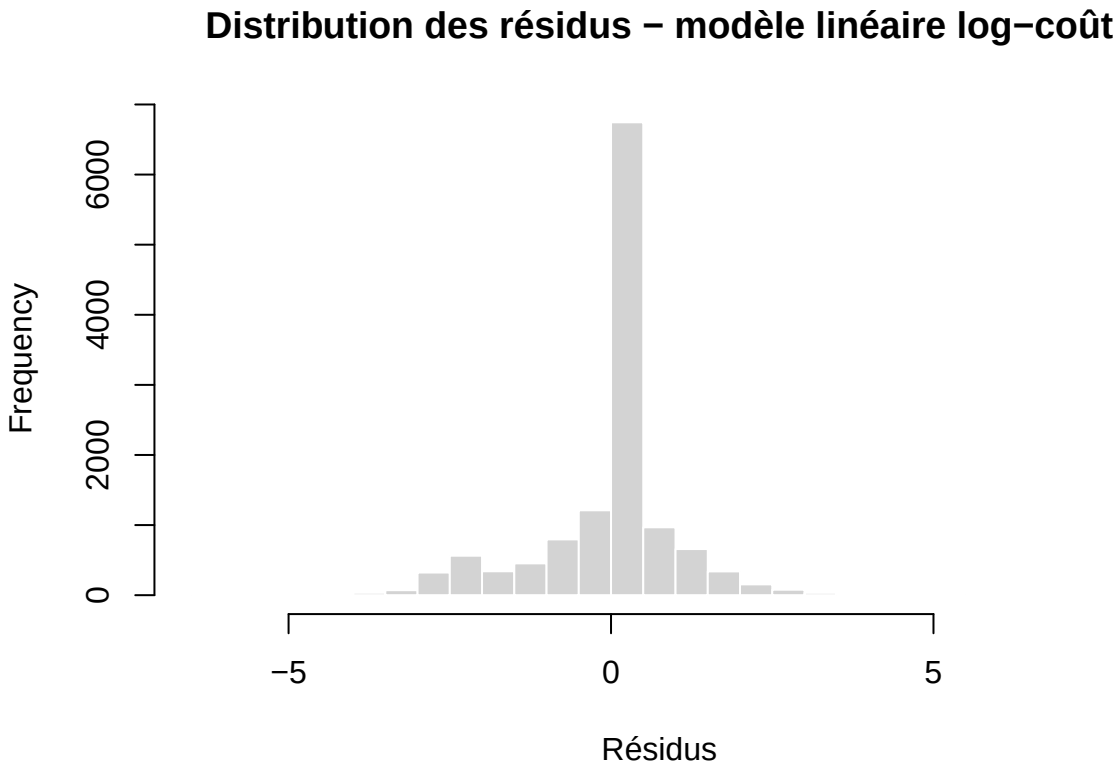
```
qqnorm(resid(lm_sev),
       main = "QQ-plot des résidus - modèle linéaire log-coût",
       pch = 19, cex = 0.5, col = "steelblue")
qqline(resid(lm_sev), col = "red", lwd = 2)
```




Les résidus s'écartent fortement de la droite de référence dans les deux queues. Le comportement se rapproche d'une loi à queue épaisse, incompatible avec l'hypothèse de normalité.

Distribution des résidus

```r
hist(resid(lm_sev),
     breaks = 50,
     main = "Distribution des résidus - modèle linéaire log-coût",
     xlab = "Résidus",
     col = "lightgray",
     border = "white")
```




Pic central très étroit et queues étirées : absence de symétrie, absence de normalité, dispersion conditionnelle élevée.

**Conclusion interne**

**Le modèle linéaire sur le log-coût échoue sur tous les axes structurants** variance non constante, résidus non gaussiens, queue lourde persistante, pouvoir explicatif quasi nul.
Il ne peut donc servir que de diagnostic exploratoire.

Les pathologies observées justifient l'usage de modèles explicitement conçus pour des montants positifs, asymétriques et hétéroscédastiques, en particulier :
– Gamma avec lien log pour variance proportionnelle au carré de la moyenne
– Log-normal pour des effets multiplicatifs et une queue plus lourde.

Ces modèles constituent la base de la modélisation de la sévérité en assurance IARD.

## 4.2 Modèle Gamma

L'objectif est d'abandonner l'approximation linéaire sur le log-coût au profit d'un modèle adapté aux données strictement positives et fortement asymétriques. La loi Gamma avec lien logarithmique impose une variance proportionnelle au carré de la moyenne, ce qui correspond au comportement empirique du coût des sinistres.

### Spécification

$$\mathbb{E}[Y_i \mid X_i] = \mu_i, \qquad \log(\mu_i) = \beta_0 + x_i^\top \beta, \qquad Y_i \sim \Gamma(\mu_i, \phi).$$

Toutes les covariables explicatives utilisées précédemment sont conservées : caractéristiques conducteur, véhicule, marque, carburant, région, et densité.

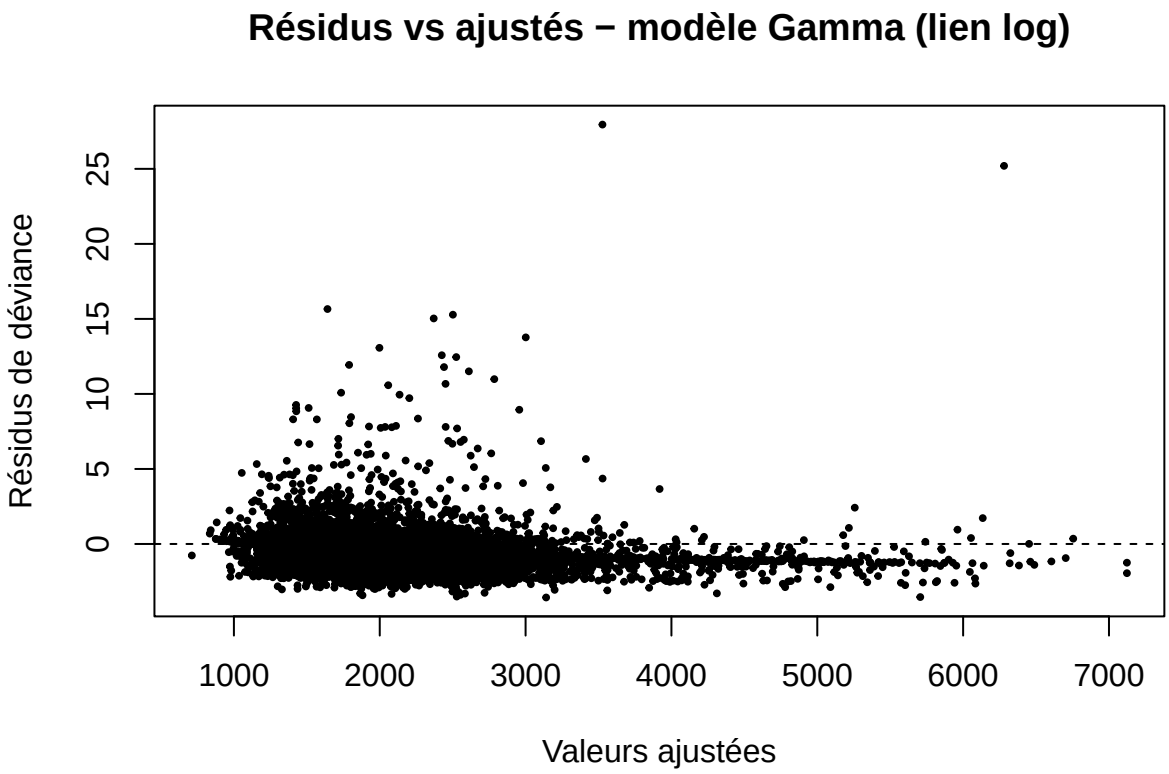
### Résultats globaux

– Null deviance = 27064 : Variabilité brute du coût avant modélisation.
– Residual deviance = 25379 : Variabilité restant après ajustement. Diminution mais limitée, ce qui montre que les covariables n'expliquent qu'une part.
– AIC = 278214
- Dispersion estimée $\phi \approx 31.43$ : Une valeur nettement supérieure à 1 révèle une variance empirique très élevée, non captée par la structure Gamma-log.

**Analyse graphique**

Résidus de déviance vs valeurs ajustées :

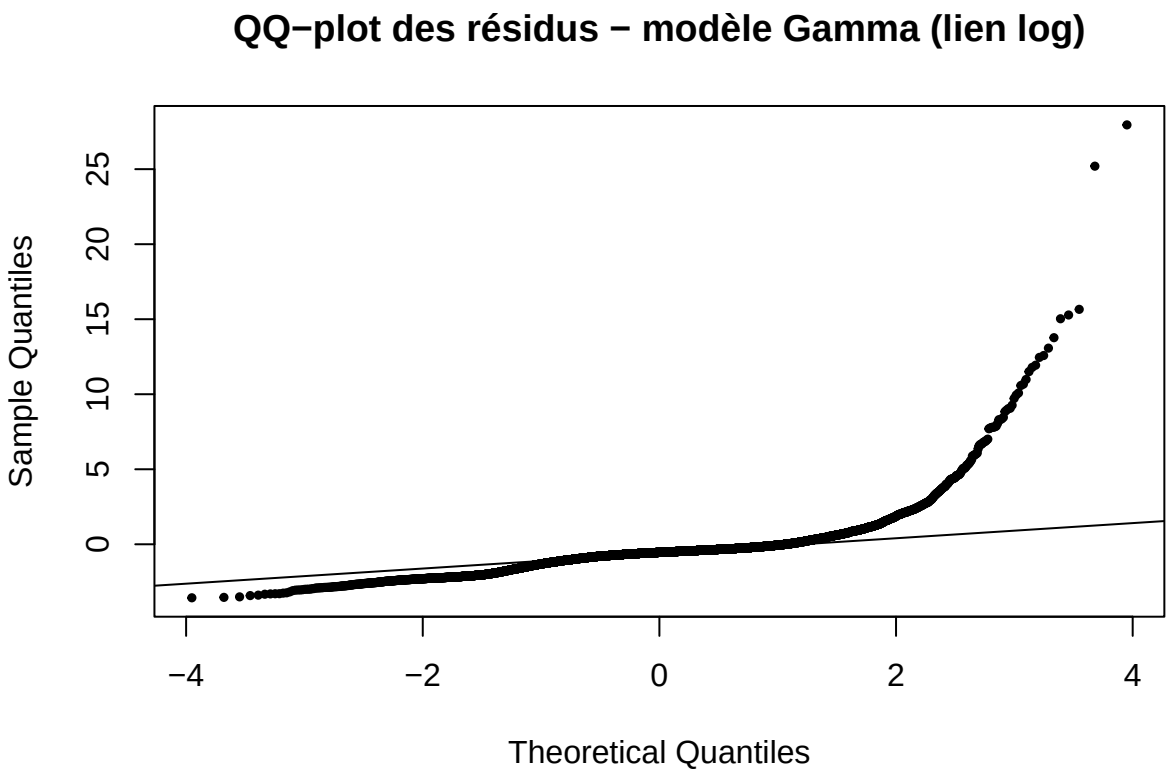
```
plot(
  fitted(gamma_log_fit),
  residuals(gamma_log_fit, type = "deviance"),
  xlab = "Valeurs ajustées",
  ylab = "Résidus de déviance",
  main = "Résidus vs ajustés - modèle Gamma (lien log)",
  pch  = 19, cex = 0.4
)
abline(h = 0, lty = 2)
```




Le modèle donne une variance proportionnelle au carré de la moyenne. Ici, cette relation n'est pas observée visuellement. Les faibles valeurs ont une variable anormalement comprimée, et les valeurs intermédiaires produisent au contraire une dispersion trop importante. Cela montre que le modèle ne capture pas correctement la variabilité des sinistres avec la loi de Gamma sous lien log : la variance empirique croît trop vite dans la queue.

QQ-plot des résidus de déviance :

```
qqnorm(residuals(gamma_log_fit, type = "deviance"),
       main = "QQ-plot des résidus - modèle Gamma (lien log)",
       pch = 19, cex = 0.5)
qqline(residuals(gamma_log_fit, type = "deviance"))
```

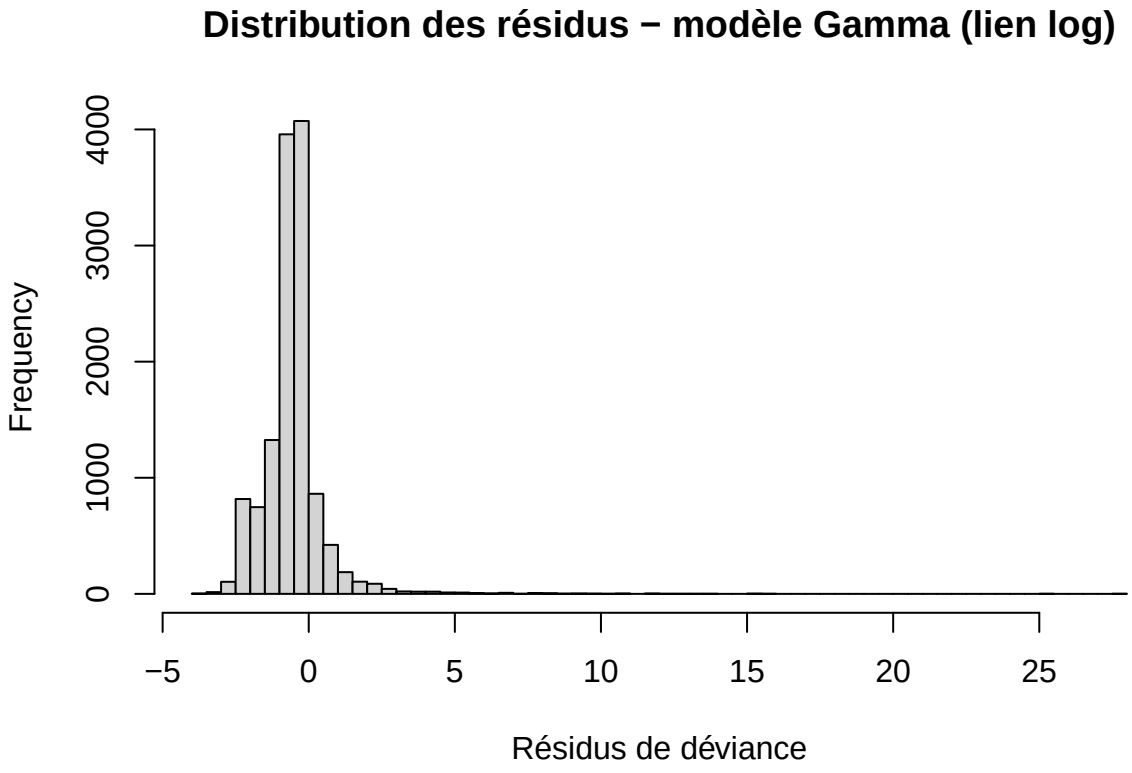



Les résidus suivent une distribution plus lourde que celle générée par une loi Gamma. La queue droite est trop étendue et signifie des coûts extrêmes que le modèle sous-estime fortement. La queue gauche, plus comprimée, signale une tendance à sur-ajuste

r les petits sinistres.

Histogramme des résidus de déviance :

```r
hist(
  residuals(gamma_log_fit, type = "deviance"),
  breaks = 50,
  main   = "Distribution des résidus - modèle Gamma (lien log)",
  xlab   = "Résidus de déviance"
)
```




Un modèle Gamma bien ajusté produit des résidus centrés avec une dispersion modérée et une queue relativement contrôlée. Ici, le pic central indique une sous-dispersion sur les petits sinistres, tandis que la traîne droite reflète une sur-dispersion extrême non absorbée par le paramètre $\phi$.

#### Conclusion interne

Les trois diagnostics convergent : – variance mal capturée, – queue droite trop lourde, – résidus structurellement incohérents avec la théorie Gamma.

Le modèle n'explose pas, mais son insuffisance structurelle est manifeste. Le Gamma-log ne reproduit pas la dynamique réelle des sinistres lourds, même avec une dispersion élevée.

## 4.3 Modèle Log-normal

L'idée est de modéliser le coût des sinistres en supposant que son logarithme est lié de manière linéaire aux variables explicatives. Cette transformation permet de réduire la forte asymétrie de la distribution des coûts et de la ramener sur une échelle où la variance est plus stable et les résidus deviennent à peu près gaussiens.

### Spécification

On pose :

$$Z_i = \log(Y_i), \qquad Z_i = \beta_0 + x_i^\top \beta + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2).$$

Sur l'échelle originale :

$$\mathbb{E}[Y_i \mid X_i] = \exp(\mu_i + \tfrac{1}{2}\sigma^2), \qquad \mu_i = \beta_0 + x_i^\top \beta.$$

Les mêmes covariables que dans les modèles précédents sont conservées :
âge conducteur, âge véhicule, puissance, marque, carburant, région, densité.

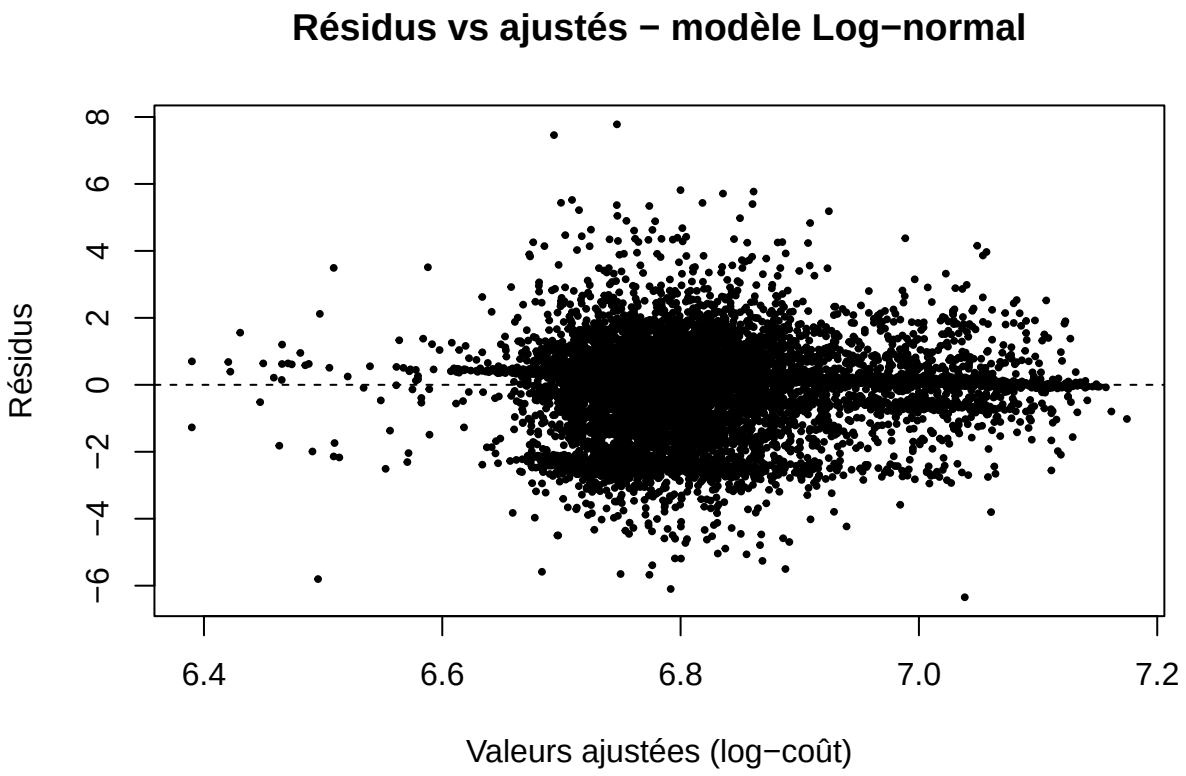
### Résultats globaux

– $R^2 = 0.0068$ : Les covariables n'expliquent qu'une fraction extrêmement faible de la variabilité du log-coût.
– $R^2$ ajusté $= 0.0050$ : Corrigé pour la dimension du modèle, l'explication reste très faible.
– Erreur standard résiduelle $\sigma \approx 1.112$ : Dispersion importante des résidus sur l'échelle du log.
– Log-vraisemblance $= -135145.9$.
– AIC $= 270355.9$ : Meilleur que celui du modèle Gamma, indiquant un ajustement numérique plus favorable sur la vraisemblance.

### Analyse graphique

Résidus de déviance vs valeurs ajustées :

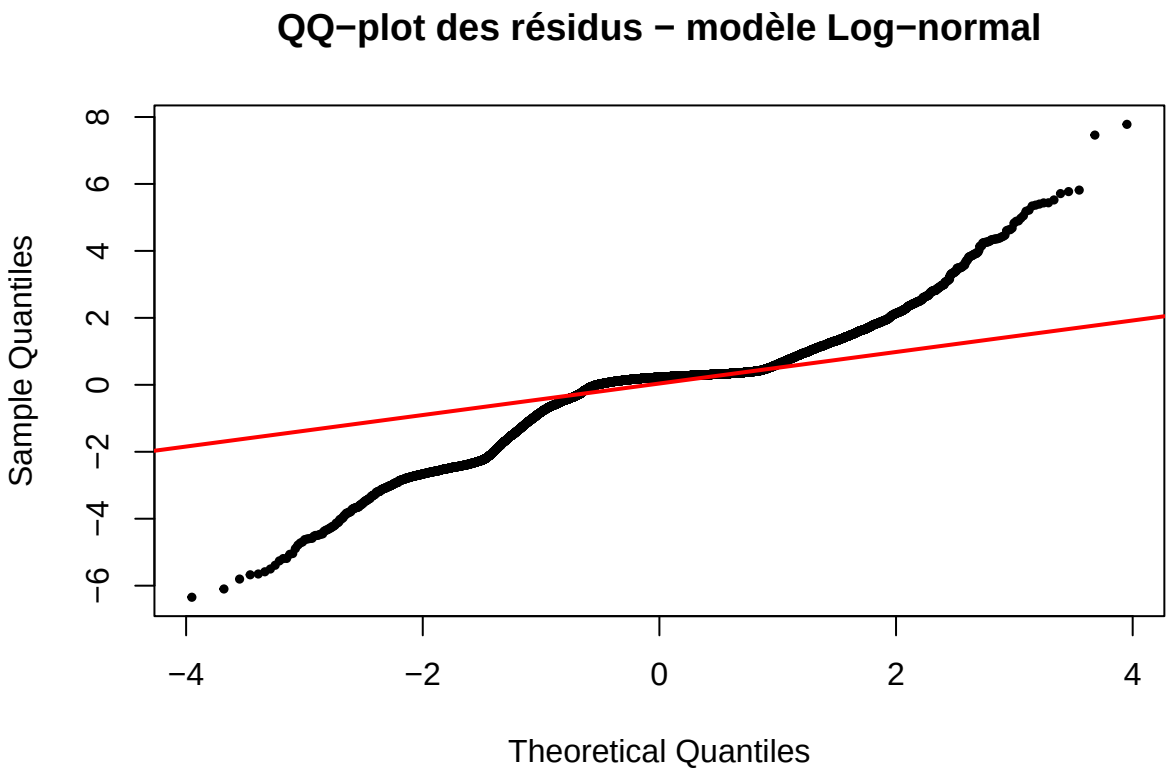
```r
plot(
  fit_ln, resid_ln,
  xlab = "Valeurs ajustées (log-coût)",
  ylab = "Résidus",
  main = "Résidus vs ajustés - modèle Log-normal",
  pch  = 19, cex = 0.4
)
abline(h = 0, lty = 2)
```




Le modèle log-normal exige une variance constante sur l'échelle du log. Le graphique ne montre pas d'accroissement évidente de la variance avec les valeurs ajustées, ce qui est cohérent avec un modèle homoscédastique. Cependant, la dispersion importante montre que la variance supposée constante reste trop faible pour absorber les extrêmes.

QQ-plot des résidus de déviance :

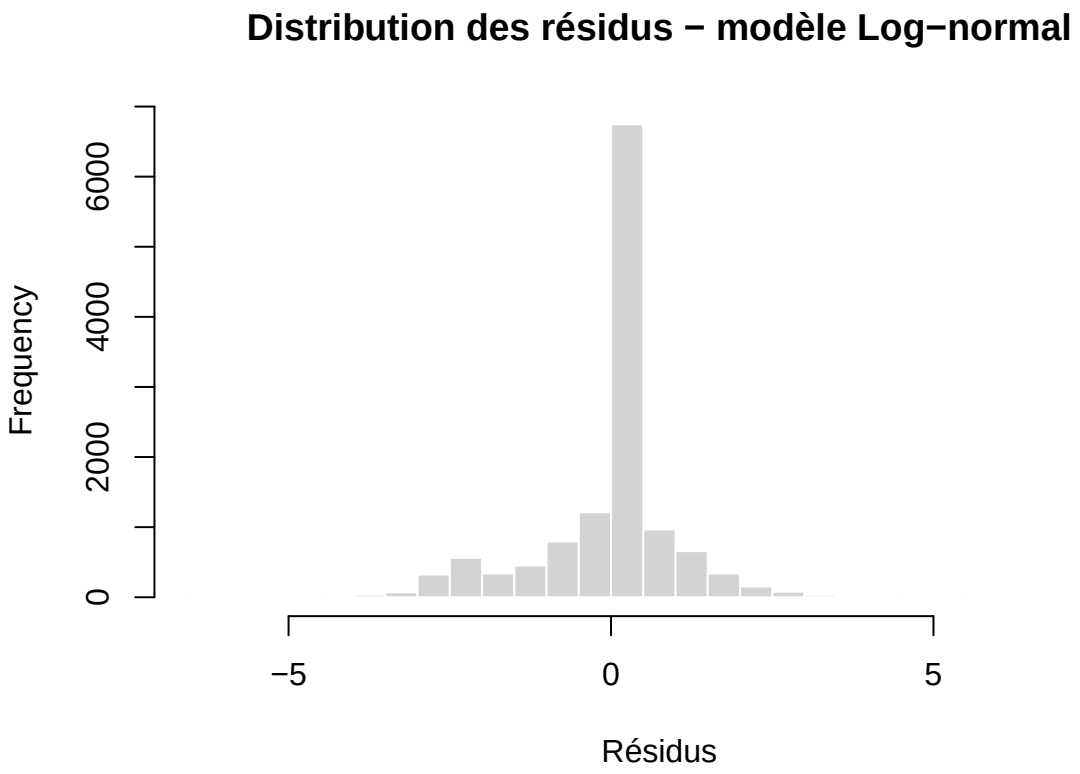
```
qqnorm(
  resid_ln,
  main = "QQ-plot des résidus - modèle Log-normal",
  pch = 19, cex = 0.5
)
qqline(resid_ln, col = "red", lwd = 2)
```




L'hypothèse clé du modèle de la normalité des résidus, est contredite. La distribution du log-coût possède des queues beaucoup plus lourdes que la normale. Le modèle capte l'ordre de grandeur central mais échoue à représenter la variabilité extrême, ce qui compromet l'interprétation statistique des coefficients et la validité des intervalles.

Histogramme des résidus

```
hist(
  resid_ln,
  breaks = 50,
  main   = "Distribution des résidus - modèle Log-normal",
  xlab   = "Résidus",
  col    = "lightgray",
  border = "white"
)
```




La forme attendue sous un modèle log-normal est qu'il soit approximativement symétrique en log. Ici, l'asymétrie résiduelle persiste, ce qui montre que la standardisation de la variance n'a pas fonctionné. Les extrêmes révèlent une sous-estimation continue des sinistres importants.

**Conclusion interne**

Les diagnostics révèlent : – une meilleure cohérence numérique (AIC inférieur), – une incapacité structurelle à respecter la normalité, – des résidus lourds dans les deux queues.

Le modèle s'ajuste mieux par vraisemblance mais viole les hypothèses fondamentales. Son pouvoir explicatif reste limité.

## 4.5 Conclusion générale sur la modélisation de la sévérité

Les deux modèles présentent des limites importantes, mais de nature différente. Le modèle Gamma souffre d'une inadéquation marquée entre la variance empirique et la structure de variance imposée par la loi Gamma : les résidus montrent une dispersion très élevée, en particulier dans la queue droite, et les diagnostics confirment une incapacité structurelle à reproduire la variabilité des sinistres lourds. Malgré cela, la forme générale du modèle reste compatible avec les propriétés attendues d'une variable strictement positive et asymétrique.

Le modèle log-normal optimise mieux l'AIC mais viole ses hypothèses fondamentales. Les résidus s'écartent fortement de la normalité, avec sous-dispersion dans la queue gauche et sur-dispersion massive dans la queue droite. La variance, supposée stabilisée sur l'échelle logarithmique, reste nettement insuffisante pour absorber la dispersion observée. Le modèle montre une stabilité apparente mais ne respecte ni la normalité ni l'homoscédasticité, ce qui limite son interprétation et fragilise sa validité statistique.

Dans un cadre actuariel où la cohérence structurelle prime sur l'optimisation brute de la vraisemblance, le modèle Gamma demeure le choix le plus approprié. Il reflète correctement la nature positive et asymétrique des coûts de sinistres et fournit une base théorique plus solide, même si sa capacité à capturer la sévérité extrême des montants reste imparfaite.

# 5. Estimation de la prime pure

## 5.1 Modèles retenus

La prime pure repose sur la décomposition :

$$\pi_i = \mathbb{E}[\text{ClaimNb}_i \mid X_i] \times \mathbb{E}[\text{ClaimAmount}_i \mid X_i,\ \text{ClaimAmount} > 0].$$

Les deux composantes sont modélisées séparément. Les diagnostics présentés dans les sections précédentes déterminent les choix finaux.

**Fréquence Modèle retenu : binomiale négative (offset log-Exposure)**

Le modèle de Poisson a été écarté en raison d'une surdispersion marquée.
La binomiale négative est retenue car elle stabilise la variance résiduelle et fournit la structure la plus cohérente pour un comptage rare et dispersé.

**Sévérité Modèle retenu : Gamma (lien log)**

Le modèle linéaire sur log-coût et le modèle log-normal présentent des violations d'hypothèses incompatibles avec un usage actuariel.
Le Gamma-log est retenu car il est structurellement adapté aux montants positifs asymétriques et constitue la spécification la plus cohérente compte tenu des diagnostics.

## 5.2. Construction de la prime pure prédite

La prime pure est obtenue en combinant les prédictions des deux modèles retenus.
La construction suit directement la définition actuarielle de la prime technique.

Pour chaque contrat $i$ :

$$\hat{\pi}_i \ = \ \hat{\mu}_i^{(\text{freq})} \times \hat{\mu}_i^{(\text{sev})}.$$

- $\hat{\mu}_i^{(\text{freq})}$ : nombre attendu de sinistres issu du modèle binomial négatif.
- $\hat{\mu}_i^{(\text{sev})}$ : coût moyen attendu du sinistre issu du modèle Gamma (lien log).

Cette construction assure la séparation stricte fréquence–sévérité qui définit la prime pure.

**Prédiction sur l'échantillon test**

Les modèles sont ajustés sur l'échantillon d'apprentissage.
Les prédictions de prime pure sont produites exclusivement sur l'échantillon test :

$$\hat{\mu}_{i,\text{test}}^{(\text{freq})} = \text{predict}(\text{NB},\ \text{newdata} = \text{test},\ \text{type} = \text{"}response\text{"}),$$
$$\hat{\mu}_{i,\text{test}}^{(\text{sev})} = \text{predict}(\text{Gamma},\ \text{newdata} = \text{test},\ \text{type} = \text{"}response\text{"}).$$

La prime pure prédite est alors :

$$\hat{\pi}_{i,\text{test}} = \hat{\mu}_{i,\text{test}}^{(\text{freq})} \times \hat{\mu}_{i,\text{test}}^{(\text{sev})}.$$

La quantité $\hat{\pi}_i$ représente le coût moyen annuel attendu des sinistres pour le profil $X_i$.

Elle intègre à la fois la fréquence estimée d'occurrence et le coût moyen conditionnel d'un sinistre.

Ce calcul constitue la base de l'évaluation de la prime pure et servira pour les diagnostics de calibration dans les sections suivantes.

## 5.3. Calibration globale

La calibration globale évalue la capacité de la prime pure prédite à reproduire le coût réel des sinistres sur un échantillon indépendant. Les modèles de fréquence et de sévérité étant ajustés sur l'échantillon d'apprentissage, les quantités présentées ici sont calculées exclusivement sur l'échantillon test.

Les deux grandeurs comparées sont :

$$\text{Coût total observé} = \sum_i \text{total_claim}_{i,\text{test}}, \qquad \text{Coût total prédit} = \sum_i \hat{\pi}_{i,\text{test}}.$$

Sur l'échantillon test :

- Coût total observé : **6 674 509**
- Coût total prédit : **6 859 136**
- Ratio de calibration : **1.0277**

Le ratio est proche de l'unité, avec une surévaluation globale d'environ 2.8 %. À l'échelle du contrat, les moyennes sont également proches :

- Coût moyen observé : **80.77**
- Coût moyen prédit : **83.01**

L'erreur quadratique moyenne (RMSE $\approx 1744$) reste élevée du fait de la forte volatilité intrinsèque des coûts individuels, liée à la rareté des sinistres et à la présence de montants extrêmes.

Cette calibration globale indique que, sur un échantillon réellement indépendant, le modèle restitue correctement la charge totale attendue. Elle ne contredit pas les limites structurelles identifiées auparavant (faible pouvoir explicatif, variance mal capturée, sévérité très dispersée) : la prime pure est une espérance, et une calibration agrégée correcte peut être obtenue même lorsque les prédictions individuelles demeurent très bruitées.

### 5.4. Analyse de cohérence et limites du modèle

La calibration test indique une restitution correcte de la charge moyenne (ratio $\approx 1.028$), ce qui confirme que le modèle fournit une estimation cohérente à l'échelle du portefeuille. Cette cohérence globale doit toutefois être mise en regard des diagnostics internes. La fréquence reste marquée par une surdispersion importante et la sévérité présente une variabilité extrême, peu captée par les distributions paramétriques retenues. Les covariables expliquent une part réduite de l'hétérogénéité individuelle et les résidus montrent une inadéquation des formes imposées, en particulier dans la queue droite. L'ensemble révèle une structure prédictive fragile, typique du risque automobile : les espérances sont stables en agrégé mais restent dominées par le bruit au niveau individuel.

### 5.5. Conclusion tarifaire

Le modèle fréquence–sévérité donne une estimation acceptable de la prime pure moyenne et une calibration globale satisfaisante sur un échantillon indépendant. Cette cohérence valide la démarche pour quantifier le coût attendu du portefeuille. Elle ne garantit pas une segmentation fine : la dispersion de la sévérité et la faible influence des covariables limitent la capacité du modèle à discriminer précisément les profils de risque. Le résultat obtenu doit donc être interprété comme une estimation agrégée plutôt que comme une base tarifaire opérationnelle. Des extensions possibles incluent des distributions plus flexibles pour la sévérité, des structures non linéaires pour la fréquence ou une segmentation en groupes homogènes de risque.

## 6. Conclusion générale

Le travail réalisé montre qu'une estimation cohérente de la prime pure peut être obtenue malgré des modèles individuels aux performances limitées, phénomène caractéristique de la tarification automobile. La stabilité en moyenne contraste avec la faible précision micro, ce qui justifie l'usage d'approches complémentaires pour renforcer la segmentation. Des pistes d'amélioration incluent l'adoption de modèles mieux adaptés aux queues lourdes, l'exploration de structures non linéaires et la construction de classes de risque homogènes, permettant d'accroître la robustesse tarifaire au-delà de la seule espérance moyenne.