



# **Modéliser un tarif avec les Modèles Linéaires Généralisés**

*6 novembre 2017  
Vanessa Désert*

# PLAN

---

- ❖ Introduction : rappels sur la tarification en assurance
- ❖ Les Modèles Linéaires Généralisés
- ❖ Exemple : tarification d'un produit d'assurance automobile
- ❖ Annexe

---

*Modéliser un tarif avec les Modèles Linéaires Généralisés*

# INTRODUCTION

- ⇒ Qu'est-ce qu'un tarif?
- ⇒ Composition d'un tarif
- ⇒ Tarification en assurance Vie VS Non Vie
- ⇒ Enjeux de la tarification
- ⇒ Le cadre usuel de la tarification

# Introduction

---

Qu'est-ce  
qu'un tarif?

Ce qu'il  
couvre?

Comment  
on le  
construit?

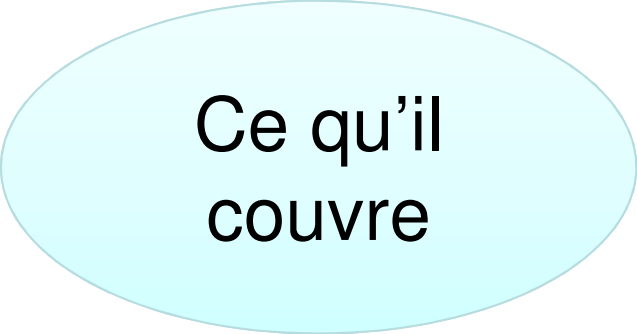
# Qu'est-ce qu'un tarif?

---

❖ La prime demandée à l'assuré permet à l'assureur de payer :

- La charge des sinistres couverts
- Le coût de la réassurance
- Les frais de gestion, d'administration, etc.
- Les taxes

...et dégager un bénéfice.



Ce qu'il  
couvre

❖ Il est construit garantie par garantie

- Risque et segmentation tarifaire différents
- Taxes différentes



Comment  
on le  
construit

❖ Analyse séparée de la composante « sinistres » et des frais.

# Composition d'un tarif (1)

---

❖ On parle de **tarif**, **cotisation** ou **prime** commerciale.

❖ La prime commerciale peut se traduire ainsi :

$$PC = PP + \tau \times PC + \alpha \times PP$$

- PP : **prime pure**, qui est l'espérance de la charge des sinistres.
- $\tau$  : taux de **chargement pour frais de gestion**, de sorte que  $\tau \times PC$  compensent les frais de gestion.
- $\alpha$  : **chargement de sécurité**, proportionnel à la prime pure.

❖ La prime pure est déterminée **garantie par garantie**. C'est la somme des primes pures correspondant aux garanties souscrites qui donnera la prime pure globale.

## Composition d'un tarif (2)

---

### ❖ Prime hors taxe ou TTC?

- Les cotisations d'assurance ne sont pas soumises à la TVA, mais elles intègrent une taxe fiscale dont le taux varie en fonction de la garantie
  - ✓ Exemple : 18% en vol ou dommages auto, 30% pour la garantie incendie en MRH...

### ❖ **Les travaux de tarification sont réalisés Hors Taxes**, c'est seulement au moment du calcul du tarif client que sont prises en comptes les différentes taxes et contributions destinés au Trésor Public.

$$PC_{TTC} = PC \times (1 + \text{taux\_taxe})$$

En annexe : Taxes et contributions sur les cotisations d'assurance

# La tarification en assurance « Vie » vs « Non Vie »

---

## La tarification en assurance Vie

- ❖ Assurance en cas de vie, en cas de décès
- ❖ Importance du temps
  - Quelle est la valeur d'1€ plus tard?
  - Notion d'actualisation (VAN=valeur actuelle nette), choix du taux
- ❖ Evènement aléatoire
  - Utilisation des probabilités
    - ✓ On parle alors de VAP = valeur actuelle probable
- ❖ La tarification repose essentiellement sur:
  - La prévision du taux
  - La prévision de la mortalité
    - ✓ Tables de mortalité normalisées



# Exemple – tarification en assurance Vie

---

- ❖ Produit « Capital différé » : L

- L'engagement de l'assureur : verser à l'assuré C€ dans k années s'il est vivant.

- ❖ L'assureur vend  $N_a$  contrats de ce type au prix PC.

- ❖ Son résultat net en fin de contrat sera donc:

$$R_{N_a} = N_a \times PC \times (1 + i)^k - C \times N_v$$

Taux d'intérêt

Nb d'assurés encore vivants en t=k (aléatoire)

Si on suppose que tous les assurés ont la même probabilité p d'être en vie à t=k, et que ces probabilités sont indépendantes:

- $E(R_{N_a}) = N_a \times PC \times (1 + i)^k - C \times N_a \times p$

- $\sigma(R_{N_a}) = C \times \sigma(N_v) = C \times \sqrt{N_a \times p \times (1 - p)}$

# La tarification en assurance Non Vie

---

## ❖ Différences avec l'assurance vie:

- Échéances beaucoup plus courtes
  - Variance beaucoup plus grande
  - Vitesse de règlement des sinistres plus lente
  - Incertitude sur le nombre et sur le coût des sinistres:
    - ✓ Il peut y avoir plusieurs sinistres sur un même contrat IARD.
    - ✓ Le règlement d'un sinistre repose sur le principe indemnitaire.
- 
- Comptabilité plus compliquée (provisionnement).
  - Importance du chargement de sécurité (implicite/réglementé en vie).
  - Importance des placements sur les marchés financiers (valable également en vie).

# La tarification en assurance Non Vie

---

- ❖ La tarification en assurance est basée sur la logique de mutualisation, légitimée théoriquement par la **loi des grands nombres** (LGN) et le **théorème central limite** (TCL).
- ❖ La LGN est valable pour des risques identiques et indépendants. Elle s'étend aux risques « assez homogènes et indépendants », dont voici des contre exemples:
  - *Maison et usine ne sont pas des risques incendie homogènes.*
  - *50 appartements d'un même immeuble ne sont pas des risques incendie assez indépendants.*
- ❖ Nécessité de suivre les **fréquences** et les **coûts** des sinistres, qui sont deux composantes **aléatoires**.
- ❖ La rentabilité d'un portefeuille est évaluée par le rapport S/P ou le **ratio combiné**.

# Enjeux de la tarification

---

- ❖ **Prévoir** à l'avance la probabilité de survenance et le montant des sinistres à indemniser.
- ❖ S'appuyer sur un **historique de données statistiques** suffisamment long et précis pour fonder les prévisions.
- ❖ Lutter contre l'**asymétrie d'information** (**antisélection** / **aléa moral**).
- ❖ Permettre à l'assureur de **payer les frais** nécessaires à son fonctionnement, notamment la rémunération des apporteurs d'affaires pour les assureurs à intermédiaires.
- ❖ **Limiter le risque de ruine** en prévoyant une marge de sécurité dans le tarif.
- ❖ Pour les compagnies cotées, **rémunérer les actionnaires** pour les fonds propres immobilisés.
- ❖ Tenir compte des **contraintes commerciales** posées par la concurrence afin que les tarifs et leurs variations soient acceptés par les assurés sans bouleverser la base de mutualisation.

# Objectifs de la tarification

---

- ❖ L'élément le plus important dans la constitution d'un tarif est la **prime pure**.
- ❖ Les objectifs de la tarification sont alors:
  - Identifier les « bons » **critères explicatifs** de la prime pure
  - Effectuer des regroupements de risques **homogènes** et « compréhensibles »
  - Se rapprocher de l'équilibre tarifaire, c'est-à-dire proposer le « juste prix » pour chaque groupe.

# Le cadre usuel de tarification

---

❖ On rappelle que :

$S$  : charge totale

$N_a$  : nombre de risques-années

$N$  : nb de sinistres

$$\text{Prime Pure} = \frac{S}{n_a} = \boxed{\frac{N}{n_a}} \times \boxed{\frac{S}{N}}$$

Fréquence      Coût Moyen

Sous réserve de l'indépendance de la fréquence et des coûts, on se ramène ainsi à modéliser l'espérance conditionnelle du nombre de sinistres et l'espérance conditionnelle du coût unitaire.

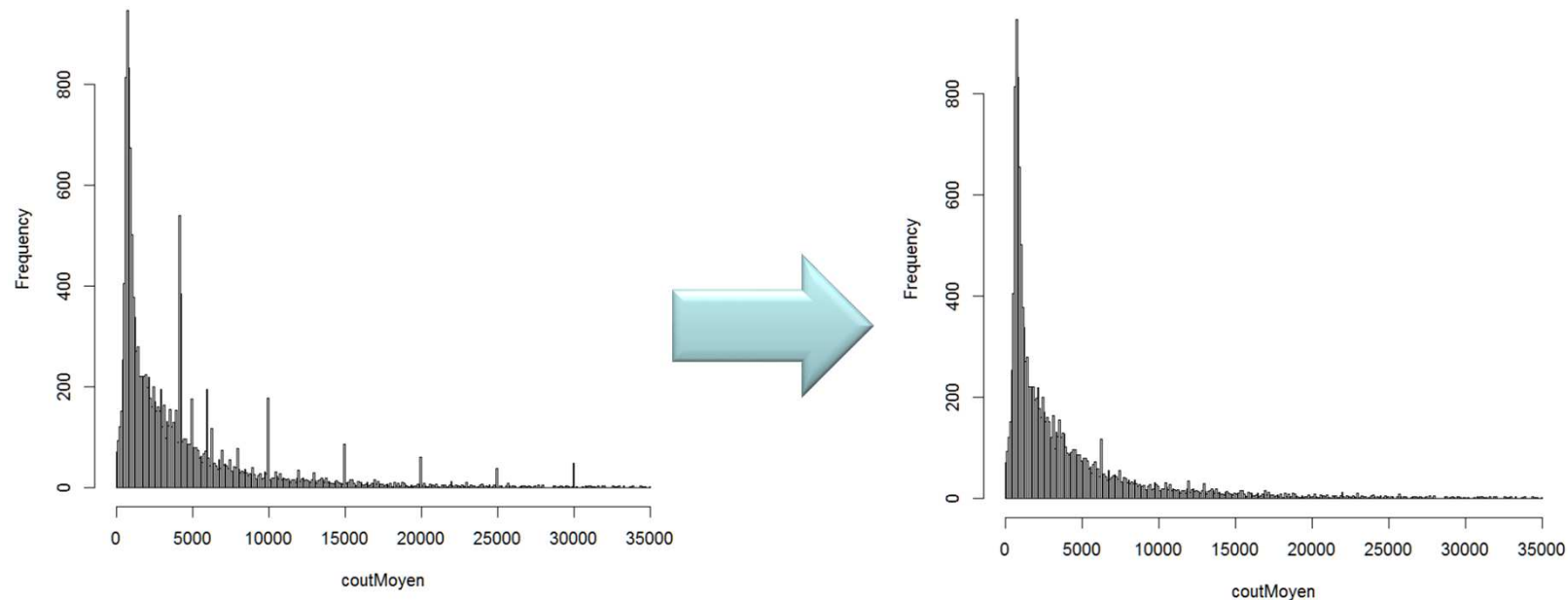
Il s'agit donc de prédire des espérances conditionnelles, ce qui est le cadre général des modèles de régression, notamment les modèles de régression non linéaires (MLG).

# Le cadre usuel de tarification

---

## ❖ Remarque sur le coût moyen

- Dans les branches longues, on peut devoir traiter spécifiquement la prise en compte de forfaits à l'ouverture qui induisent des discontinuités dans la distribution des coûts.
- Des conventions (IRSA, CIDRE) génèrent également sur certaines garanties des discontinuités (règlement d'une somme forfaitaire)



# Les étapes d'une tarification

---

- ❖ La réalisation d'un tarif nécessite plusieurs étapes :
  - La constitution de la base de données fiabilisée
  - La distinction des sinistres attritionnels, graves et sériels
  - Le choix des variables tarifaires
  - La modélisation de l'effet des caractéristiques des individus (représentées par les modalités des variables tarifaires) sur les variables à expliquer (la fréquence et le coût) dans le cadre d'un modèle explicatif de la « charge espérée »
  - Le lissage du tarif brut, qui permet de prendre en compte les contraintes de la politique tarifaire
  - Le passage au tarif commercial.



---

*Modéliser un tarif avec les Modèles Linéaires Généralisés*

# LES MODÈLES LINÉAIRES GÉNÉRALISES

- ⇒ Principes de la régression
- ⇒ Les coefficients du modèle
- ⇒ Loi de Y, Fonction de lien
- ⇒ Qualité d'un MLG, résidus
- ⇒ Notion d'offset
- ⇒ Cas des sinistres graves
- ⇒ Lissage

# Principes de la régression

---

- ❖ Régressions usuelles, selon la nature des données

			Variable(s) explicative(s)		
			Quantitative(s)	Qualitative(s)	Quanti et Quali
Variable réponse	Quantitative	Loi Normale	Régression linéaire simple et multiple	Analyse de la variance	Régression linéaire générale, analyse de la covariance
		Autres lois	Modèle linéaire généralisé		
	Qualitative	Binomiale	Régression logistique		
		Ordinale			
		Nominale	Régression logistique généralisée		

# Principes de la régression

---

## ❖ Equation de régression linéaire :

$$Y = a \times X + b + e$$

Diagram illustrating the components of the linear regression equation:

- $Y$ : Variable à expliquer (Variable to be explained)
- $a$ : Coefficient de  $X$  (Coefficient of  $X$ )
- $X$ : Variable explicative (Explanatory variable)
- $b$ : Constante (Constant)
- $e$ : Résidu (Residual)

## Notation matricielle :

$$Y = XB + e$$

La partie  $a \times X + b$  (resp.  $XB$  en écriture matricielle) est le **prédicteur linéaire**, c'est la **composante déterministe** du modèle. En MLG, on cherchera à se ramener à un cas faisant apparaître dans un membre de l'équation ce prédicteur linéaire.

Les seules **composantes aléatoires** de l'équation sont  $Y$  et  $e$ . Ces deux éléments suivent en toute logique la même loi.

# Principes de la régression

---

- ❖ Modèle linéaire généralisé (MLG)

$$g(E(Y|X)) = XB = \sum \beta_k x_k$$

$g$  est la **fonction de lien**, qui met en conformité la plage de valeurs autorisées pour l'espérance de  $Y$  et les valeurs du prédicteur linéaire.

- ❖ Avantage par rapport aux modèles linéaires classiques: le caractère normal de la variable  $Y$  n'est plus imposé, seule l'appartenance à une famille exponentielle est nécessaire.

# Les MLG – Loi de Y

---

## ❖ Loi de Y

Elle appartient à la **famille exponentielle**.

- C'est-à-dire que la formule qui régit sa densité doit pouvoir se mettre sous la forme

$$f(y|\theta, \varphi) = \exp\left(\frac{y\theta - b(\theta)}{a(\varphi)} + c(y, \varphi)\right)$$

où  $a(\cdot)$ ,  $b(\cdot)$  et  $c(\cdot)$  sont des fonctions, et  $\theta$  est appelé **paramètre naturel**. Le paramètre  $\varphi$  est considéré comme un paramètre de nuisance. Si  $a(\varphi) = \varphi$ , alors  $\varphi$  est appelé paramètre de dispersion. Il vaut 1 pour les lois à un paramètre.

- Exemples : Lois Normale, Gamma, Poisson, Binomiale Négative, Binomiale

# Famille exponentielle

---

❖ Exemple : loi de Poisson  $P(\lambda)$

$$f(y|\lambda) = \exp(-\lambda) \frac{\lambda^y}{y!} = \exp(y \ln(\lambda) - \lambda - \ln(y!)) , y \in \mathbb{N}$$

On retrouve donc :

$$\theta = \ln(\lambda)$$

$$a(\varphi) = \varphi = 1$$

$$b(\theta) = \exp(\theta) = \lambda$$

$$c(y, \varphi) = -\ln(y!)$$

# Famille exponentielle

---

## ❖ Composantes de la famille exponentielle

Distribution	$\theta$	$b(\theta)$	$a(\varphi)$
Normale $N(\mu, \sigma^2)$	$\mu$	$\frac{\theta^2}{2}$	$\sigma^2$
Gamma $G(\mu, \nu)$	$-\frac{1}{\mu}$	$-\ln(-\theta)$	$\frac{1}{\nu}$
Poisson $P(\lambda)$	$\ln(\lambda)$	$e^\theta = e^\lambda$	1
Binomiale $B(n, p)$	$\ln\left(\frac{p}{1-p}\right)$	$n \times \ln(1 + e^\theta)$	1

## ❖ Espérance et variance d'une variable $Y$ appartenant à la famille exponentielle :

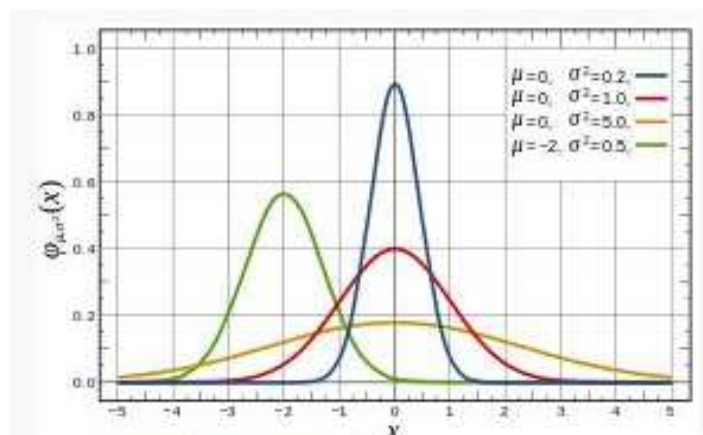
$$E(Y) = b'(\theta) \text{ et } \text{Var}(Y) = b''(\theta) a(\varphi)$$

# Loi de Y

---

## ❖ Loi Normale ou de Gauss $N(\mu, \sigma^2)$

- Prend des valeurs réelles, positives ou négatives. Sa dispersion est constante quelle que soit la moyenne.
- C'est la loi que suit la somme d'un grand nombre de variables aléatoires (théorème central limite).
- Espérance :  $\mu$
- Variance :  $\sigma^2$  (indépendante de  $\mu$ , constante pour toutes les valeurs de la moyenne)



Densité de probabilité  
Loi Normale



# Loi de Y

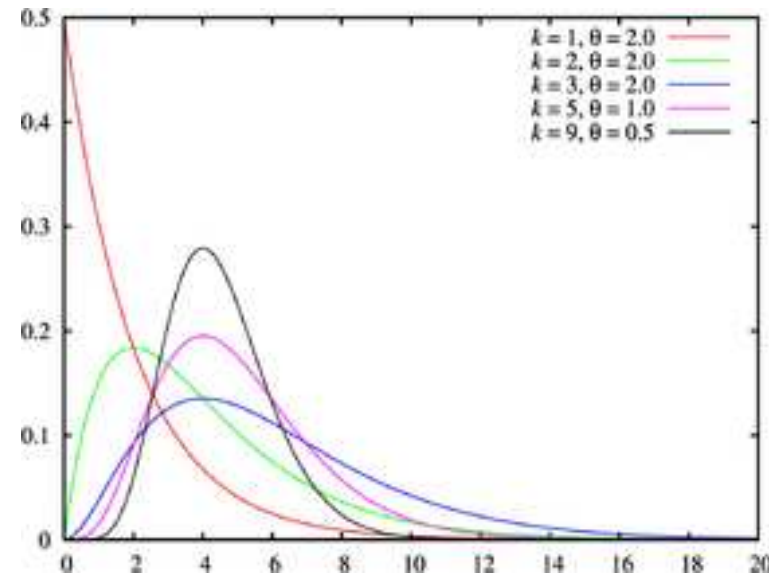
---

## ❖ Loi Gamma $G(\mu, \nu), \mu > 0, \nu > 0$

- Ne prend que des valeurs positives
- Distribution asymétrique.
- Espérance :  $\mu$
- Variance :  $\sigma^2 = \mu^2 / \nu$

Expression de la densité:

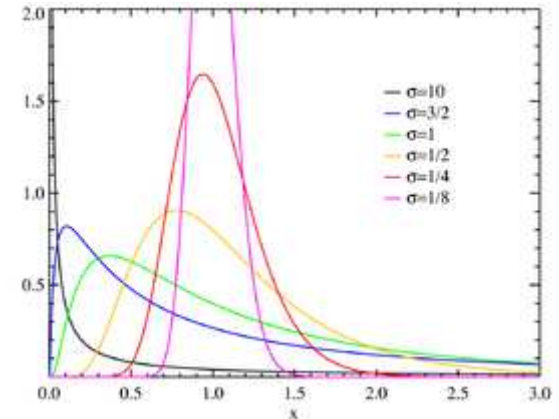
$$f(y|\mu, \nu) = \frac{1}{\Gamma(\nu)} \left(\frac{\nu}{\mu}\right)^\nu y^{\nu-1} \exp\left(-\frac{\nu}{\mu} y\right)$$



Densité de probabilité  
Loi Gamma

# Loi de Y

- ❖ Loi Log normale  $\text{LogN}(\mu ; \sigma^2)$  ,  $\mu \in \mathbb{R}$ ,  $\sigma^2 > 0$ 
  - Ne prend que des valeurs strictement positives
  - Espérance :  $e^{\mu + \sigma^2 / 2}$
  - Variance :  $e^{2\mu + \sigma^2} \times (e^{\sigma^2} - 1)$



Densité de probabilité  
Loi Log Normale,  $\mu=0$

- En tarification IARD, cette loi est plus utilisée que la loi Normale.
- Lorsque la loi Log Normale n'est pas directement utilisable dans le logiciel de modélisation (par exemple dans SAS), on passe par une modélisation utilisant la loi Normale grâce à la propriété suivante:

$$X \sim \text{LogN}(\mu, \sigma^2) \iff \text{LN}(X) \sim N(\mu, \sigma^2)$$

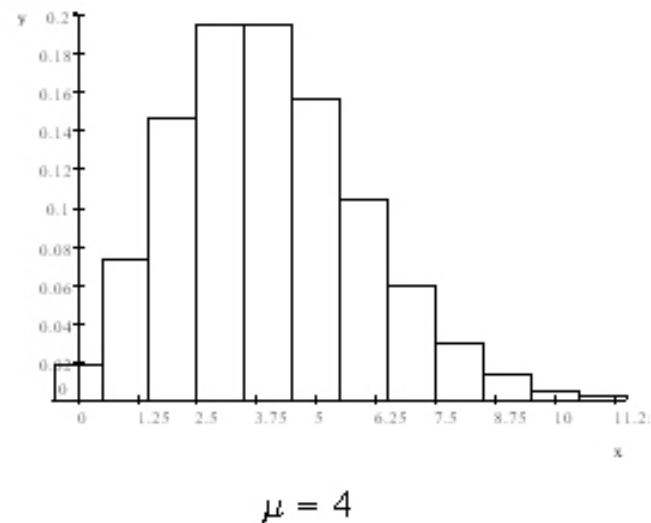
Et on utilise la fonction de lien « Identité ».

# Loi de Y

---

## ❖ Loi de Poisson $P(\lambda)$ , $\lambda > 0$

- Loi discrète. Ne prend pour valeurs que des entiers positifs.
- C'est la « loi des évènements rares ». Elle peut se définir comme le nombre d'évènements survenus au cours d'une période donnée.
- Espérance =  $\lambda$  = Variance

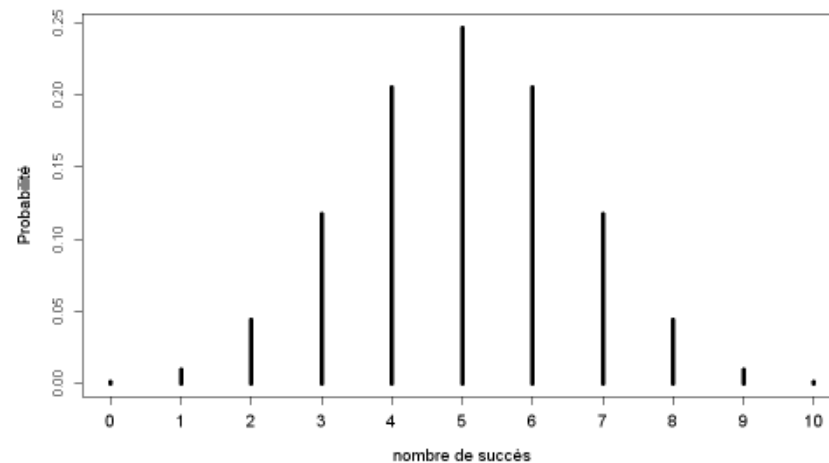


# Loi de Y

---

## ❖ Loi Binomiale $B(n, p)$ , $n > 0$ , $0 < p < 1$

- Loi discrète.
- C'est la somme de  $n$  lois de Bernoulli, qui prennent deux valeurs: 0 et 1.  $p$  est la probabilité de survenance de 1. La loi Binomiale représente le nombre de « 1 » parmi  $n$  tirages.
- Espérance :  $\mu = np$
- Variance :  $\sigma^2 = np(1-p)$



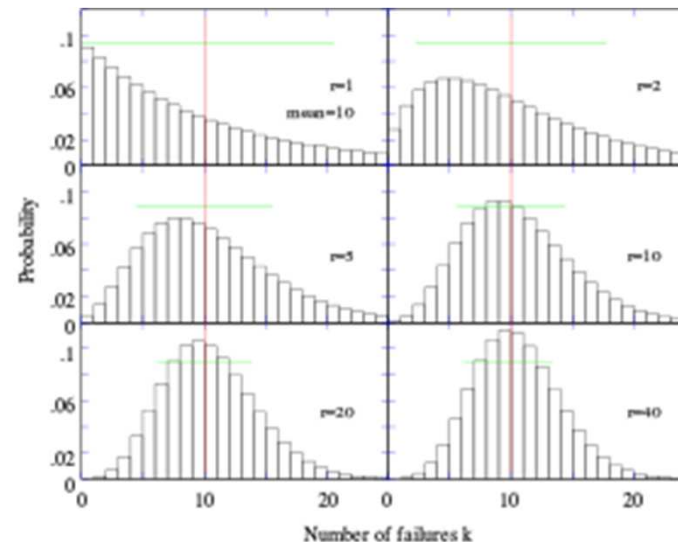
***Densité de probabilité de la loi Bin(10 ; 0,5)***

# Loi de Y

---

## ❖ Loi Binomiale Négative $BN(r, p)$ , $r > 0$ , $0 < p < 1$

- Loi discrète.
- C'est une généralisation de la loi de Poisson. Au lieu de modéliser une succession d'évènements indépendants d'espérance constante (c'est-à-dire au hasard), elle suppose qu'ils se produisent d'une manière contagieuse : il y a « surdispersion ». C'est la loi du nombre de tentatives effectuées avant d'obtenir le  $n^{\text{ième}}$  succès.
- Espérance :  $\mu = r(1-p)/p$
- Variance :  $\sigma^2 = \mu/p$



# Fonction de lien

---

- ❖ Chacune des lois de la famille exponentielle possède une fonction de lien spécifique, dite **fonction de lien canonique**, permettant de relier l'espérance  $\mu$  au paramètre naturel (ou canonique)  $\theta$ .

- Par exemple, pour la loi de Poisson  $P(\lambda)$ , d'espérance  $\mu = \lambda$  :

$\theta = \ln(\lambda) \Rightarrow$  donc la fonction de lien canonique est « LN »

Loi de Y	Fonction de lien canonique	Formule
Normale	Identité	$g(y) = y$
Gamma	Inverse	$g(y) = 1/y$
Poisson	Logarithme	$g(y) = \text{LOG}(y)$
Binomiale Négative	Logarithme	$g(y) = \text{LOG}(y)$
Binomiale	Logit	$g(y) = \text{LOG}(y/(1-y)) = \text{LOGIT}(y)$

- ❖ L'utilisation du lien canonique est privilégié, mais il est cependant possible de choisir une autre fonction de lien, tant que celle-ci est monotone et définie sur un intervalle compatible avec celui de Y.

# Fonction de lien

---

## ❖ Fonction de lien identité

$$E(Y|X) = Xb$$

- Caractéristiques:

- ✓ L'espérance de Y doit pouvoir prendre **n'importe quelle valeur réelle**.
- ✓ **Lecture** directe des coefficients du modèle, comme dans une régression linéaire.
- ✓ On parle alors de **modèle additif**, car les effets des différentes variables explicatives s'accumulent directement sur la valeur modélisée de la réponse.

# Fonction de lien

---

## ❖ Fonction de lien logarithme

$$\text{LOG}(E(Y|X)) = Xb \Leftrightarrow E(Y|X) = \text{EXP}(Xb)$$

- Caractéristiques:
  - ✓ L'espérance de Y ne prend que des valeurs strictement positives.
  - ✓ **Transformation des coefficients** (fonction exponentielle)
  - ✓ On parle alors de **modèle multiplicatif** : chaque augmentation d'une unité de X (quanti) entraîne la multiplication de l'espérance conditionnelle de Y par EXP(coeff). Si X est qualitative, pour passer de la valeur de référence à la valeur de la modalité analysée, on estime l'effet en multipliant la réponse par l'exponentielle du coefficient associé.



# Estimation des coefficients $\beta$

---

- ❖ Les coefficients de régression  $\beta_0, \beta_1, \dots, \beta_p$  sont estimés, sur base des données, par la méthode du maximum de vraisemblance. Il s'agit donc de maximiser la log-vraisemblance en  $\beta$ .

$$\text{Log } L(\theta_1, \dots, \theta_n, \varphi, y_1, \dots, y_n) = \sum_{i=1}^n \left[ \frac{y_i \theta_i - b(\theta_i)}{a(\varphi)} + c(y_i, \varphi) \right]$$

- ❖ Algorithme couramment utilisé : Newton-Raphson
  - Méthode numérique et itérative
  - Se base sur la matrice d'information de Fisher

# Interprétation des coefficients

---

- ❖ Un coefficient mesure un **impact**
  - Positif ou négatif, selon le signe.
  - Son importance se mesure en regardant sa valeur absolue.
  - Mais ceci est relatif à une référence, qu'il faut fixer. La constante correspond au risque de l'individu de référence.
- ❖ Pour des **variables continues**, le coefficient traduit une augmentation d'une unité de l'échelle dans laquelle la variable est exprimée.
  - Effet linéaire : donc passer d'un âge de 19 à 20 ans a le même effet que le passage de 55 à 56 ans.
  - Pour capter des effets non linéaires, il est alors préférable de passer par des variables discrétisées.
- ❖ Pour des **variables qualitatives ou discrétisées**, on a un coefficient par classe (modalité).

# Les coefficients du modèle

---

- ❖ La régression mène à la production de valeurs pour les coefficients du modèle : ceux des variables explicatives (ou **covariables**) et la constante. Dans les logiciels de modélisation (exemple SAS), les valeurs du vecteur  $b$  sont en général proposées dans un tableau récapitulatif :
  - Pour les variables explicatives **quantitatives**  $\Rightarrow$  **un coefficient par variable**
  - Pour les variables explicatives **qualitatives**  $\Rightarrow$  **un coefficient par valeur (modalité) de la variable**. De plus l'un des coefficient est fixé arbitrairement (en général à la valeur 0), il correspond à la **valeur de référence** de la variable explicative.
- ❖ **P-Value** : pour chaque coefficient non arbitrairement fixé, calcul d'une statistique de test correspondant à l'hypothèse de nullité du coefficient. Si l'hypothèse est confirmée, cela indique :
  - Absence d'influence de la variable explicative si elle est quantitative.
  - Possibilité de regrouper cette modalité de la variable avec la valeur de référence, sans perte de qualité de modélisation, dans le cas d'une variable explicative qualitative.
- **On cherche donc à rejeter l'hypothèse, c'est-à-dire obtenir des p-values faibles.**

# Les coefficients du modèle

❖ Exemple :

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	7.1666	0.01	7.1471	7.1862	516106	<.0001
VAR A	a1	1	0.0726	0.0261	0.0215	0.1237	7.76	0.0053
VAR A	a2	1	0.0591	0.0233	0.0134	0.1048	6.42	0.0113
VAR A	a3	1	0.0297	0.0124	0.0054	0.0541	5.73	0.0167
VAR A	a4	0	0	0	0	0	.	.
VAR A	a5	1	0.0507	0.0081	0.0348	0.0666	39.19	<.0001
VAR A	a6	1	0.0131	0.0066	0.0002	0.026	3.95	0.0468
VAR A	a7	1	0.0051	0.007	-0.0085	0.0188	0.54	0.4626
VAR B	b1	1	-0.0776	0.0201	-0.117	-0.0383	14.95	<.0001
VAR B	b2	1	-0.0727	0.0129	-0.0979	-0.0475	31.9	<.0001
VAR B	b3	1	-0.0532	0.0089	-0.0705	-0.0358	36.06	<.0001
VAR B	b4	1	-0.0227	0.0064	-0.0353	-0.0101	12.49	<.0001
VAR B	b5	0	0	0	0	0	.	.
VAR C	c1	1	0.0458	0.01	0.0262	0.0655	20.98	<.0001
VAR C	c2	0	0	0	0	0	.	.

P-value



Cette modalité pourrait être regroupée avec la modalité de référence (0), « si cela a du sens »



Coefficients  $\beta_k$

Bornes de l'IC à 95%

# Qualité d'un modèle linéaire généralisé

---

## ❖ Vraisemblance (L)

- Minimale pour le modèle ne présentant qu'une constante en guise de prédicteur linéaire, et maximale pour le modèle saturé, qui ajuste exactement toutes la valeurs de la variable à expliquer.

## ❖ Déviance (D)

$$D = -2 \text{ LOG } \left( \frac{L(\text{modèle étudié})}{L(\text{modèle saturé})} \right)$$

- Elle suit une loi du  $\chi^2$  dont le nombre de degrés de liberté est celui qui sépare les deux modèles. Sa valeur « attendue » est approximativement son nombre de DDL puisque la moyenne d'un  $\chi^2$  est ce nombre de DDL.
- On parle de **sur-dispersion** lorsque la déviance est nettement supérieure à son nombre de degrés de liberté. Cette sur-dispersion indique un mauvais ajustement du modèle aux données.
- Inversement, la **sous-dispersion** correspond à une déviance nettement inférieure à son nombre de degré de liberté. Ce phénomène est plus rare.

# Qualité d'un modèle linéaire généralisé

---

## ❖ Critère d'Akaike AIC et BIC (Bayesian Informative Criterion)

- Plus le nombre de variables du modèle augmente, plus la déviance diminue (plus la vraisemblance augmente), même si la variable ajoutée n'est pas pertinente. L'AIC et le BIC permettent de contrebalancer la réduction de la déviance avec une quantité traduisant la complexité du modèle.
- Ils se calculent à partir de la log-vraisemblance, mais tiennent également compte du nombre d'observations et de paramètres du modèle (=nb de variables+1).

$$AIC = -2 \log(L) + 2p$$

$$BIC = -2 \log(L) + p \log(n)$$

Où  $L$  est la vraisemblance maximisée,  $p$  est le nombre de paramètres et  $n$  le nombre d'observations

- On cherche à minimiser ces quantités.

# Surdispersion

---

- ❖ La **déviante** suit une loi du Khi-deux dont le nombre de degrés de liberté est celui qui sépare les deux modèles.
- ❖ On parle de **sur-dispersion** lorsque la déviante est nettement supérieure à son nombre de degrés de liberté. Elle indique un mauvais ajustement du modèle aux données.
- ❖ Dans SAS, on peut corriger la sur-dispersion avec l'option DSCALE.
- ❖ Notons que ce cas ne s'applique qu'aux lois n'intégrant pas de paramètre d'échelle, comme la loi de Poisson.

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	223	685,4341	3,0737
Scaled Deviance	223	685,4341	3,0737
Pearson Chi-Square	223	1136,9209	5,0983
Scaled Pearson X2	223	1136,9209	5,0983
Log Likelihood		1121.8627	

*On repère ici de la sur-dispersion. La déviante se comportant comme un  $\text{Khi}^2$ , on devrait avoir une valeur proche du nombre de degrés de liberté.*

# Qualité d'un modèle linéaire généralisé

---

## ❖ Analyse de la déviance

- Pour mesurer l'apport d'une variable dans un modèle, on peut étudier la variation de déviance entre un modèle avec et un modèle sans cette variable  $X_i$ . On s'interroge sur la significativité de la perte de qualité du modèle sans la variable  $X_i$ .
- Dans SAS : **statistiques de type 1 et de type 3**
  - ✓ Le test de Type 1 est un test du rapport des vraisemblances. Il permet d'effectuer une analyse séquentielle. Il s'agit d'ajuster les modèles, en commençant par le modèle nul (ne contenant qu'une constante) et en ajoutant au fur et à mesure les variables précisées dans le modèle. On regarde alors la significativité de chaque variable ajoutée.

*Ce test est très dépendant de l'ordre d'énumération des variables, ce qui limite son utilisation.*

- ✓ Le test de Type 3 est également un test du rapport des vraisemblances. Il consiste à considérer le modèle avec toutes les variables énumérées, et à tester l'influence du retrait de chacune d'entre elles sur la qualité du modèle. Il est indépendant de l'ordre des variables.



# Importance des résidus

---

- ❖ Dans tous les cas de régression, la quantité modélisée est l'espérance de  $Y$  sachant les valeurs prises par les variables explicatives  $X$ . Les résidus « complètent » l'écart entre valeurs observées de la variable à expliquer et cette espérance.
- ❖ L'analyse des résidu permet de :
  - Apprécier la qualité du modèle : les résidus doivent être décorrélées de la variable réponse  $Y$ , sans quoi cela signifierait qu'une variable explicative importante a été omise.
  - Repérer les observations atypiques, qui produisent des valeurs « hors normes » du résidu. Ces dernières sont d'autant plus faciles à repérer que la forme attendue de la distribution des erreurs est connue et correspond à la loi de la variable  $Y$ .

# Analyse des résidus

## ❖ Résidu « simple »

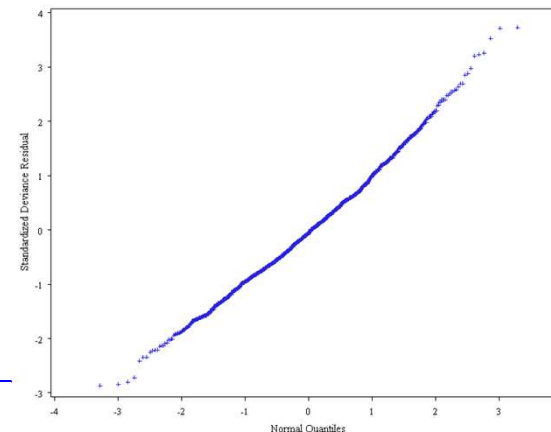
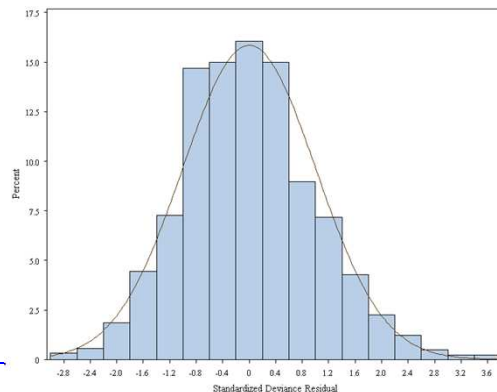
- Il se définit comme la part non expliquée de la variable Y, car non explicable linéairement par les variables explicatives utilisées.

$$r = Y - E(Y|X)$$

- Ce résidu n'est pas exploitable dans le cadre d'un MLG.

## ❖ Résidus de la déviance

- Quantités dérivées du résidu simple, de sorte que leur somme est égale à la déviance du modèle. Ces quantités sont standardisées, c'est à dire divisées par leur variance.
- Les résidus de la déviance suivent une loi quasiment analogue à une **loi normale**, quelle que soit la loi de la variable réponse Y. On peut donc faire des tests d'adéquation à la loi normale et des QQ-plots des résidus de la déviance, afin de repérer les valeurs anormales.



# En pratique : modéliser un tarif avec les MLG

---

- ❖ Contraintes à l'utilisation d'un modèle linéaire généralisé
  - **Colinéarité** : comme tous les modèles de régression, les variables explicatives ne doivent pas être corrélées entre elles.
  - **Fiabilité des données** : on ne réalise pas un modèle de tarification sans s'être assuré au préalable de la qualité des données utilisées!
    - ✓ Comme dans toute étude statistique, la phase de préparation des données est ce qui demande le plus de temps.
- ❖ Choix du modèle :
  - Minimiser la Déviance
- ❖ Adéquation du modèle
  - Déviance normalisée (scaled deviance)
  - Statistique du Khi-deux de Pearson

# En pratique : modéliser un tarif avec les MLG


---

- ❖ Objectif de la modélisation : expliquer la PRIME PURE
  - Par **GARANTIE**
  - En fonction de **CRITERES EXPLICATIFS** (⇒ segmentation)
  
- ❖ Problème : difficultés à estimer directement la prime pure
  - Loi de probabilité?
  - On perd en finesse d'analyse : certains critères influent plutôt sur la fréquence de sinistres, d'autres sur le coût.
  - Le modèle est construit en premier lieu pour produire un tarif, mais l'analyse des résultats peut être utile à d'autres fins (découverte de « niches » sur lesquelles axer le développement, études spécifique sur la fréquence ou le coût des sinistres, surveillance de portefeuille...)

# En pratique : modéliser un tarif avec les MLG

---

$$\textit{Prime Pure} = \textit{Fréquence} \times \textit{Coût Moyen}$$


$$F = \frac{\textit{Nb sinistres}}{\textit{Nb Risques Années}}$$


$$CM = \frac{\sum \textit{coûts}}{\textit{Nb sinistres}}$$

- ❖ En pratique : modélisations séparées de la fréquence et du coût des sinistres
  - Des lois usuelles qui sont généralement bien adaptées
    - ✓ Fréquence : lois discrètes, souvent Poisson ou Binomiale Négative
    - ✓ Coût : Lois continues, souvent Gamma ou LogNormale
  - Des critères de segmentation qui peuvent être différents.
  
- ❖ **Hypothèse fondamentale** : indépendance entre la fréquence et le coût des sinistres.

# Notion d'offset

---

- ❖ Exemple, avec fonction de lien Log:

$$\text{LOG} \left( \frac{\text{Nb sinistres}}{\text{Nb RA}} \right) = Xb$$

Le nombre de risques-années est très variable selon les contrats et le mode de constitution de la base de données



$$\text{LOG}(\text{Nb sinistres}) = Xb + \underbrace{\text{LOG}(\text{Nb RA})}_{\text{OFFSET}}$$

En pratique, cela revient à intégrer l'offset comme une variable explicative, mais en forçant son coefficient à 1.

Dans la proc GENMOD de SAS (ainsi que dans la fonction glm de R), l'instruction OFFSET permet de définir ce type de variable.

# Modèle multiplicatif ou additif?

---

## ❖ Dépend du lien canonique

- Identité,  $1/y \Rightarrow$  modèle additif
- LOG  $\Rightarrow$  modèle multiplicatif

*Point de vigilance:*

- *Domaine de définition de la variable cible*
- *Interprétation des coefficient (transformation,...)*

## ❖ Exemple : cas de la régression Gamma

- Fonction de lien canonique =  $1/y$
- Problèmes :
  - ✓ La positivité des valeurs prédites n'est pas garantie, or la loi Gamma ne prend que des valeurs strictement supérieures à 0.
  - ✓ Les coefficients des covariables sont à interpréter à l'inverse de la logique habituelle.
- Solution : utiliser la fonction de lien Logarithme
  - ✓ Correction des 2 problèmes ci-dessus.
  - ✓ Baisse de qualité peu notable.

*Pour vérifier que la fonction de lien est adéquate : analyse de la déviance (comparaison avec la fonction de lien canonique)*

# Cas des sinistres graves

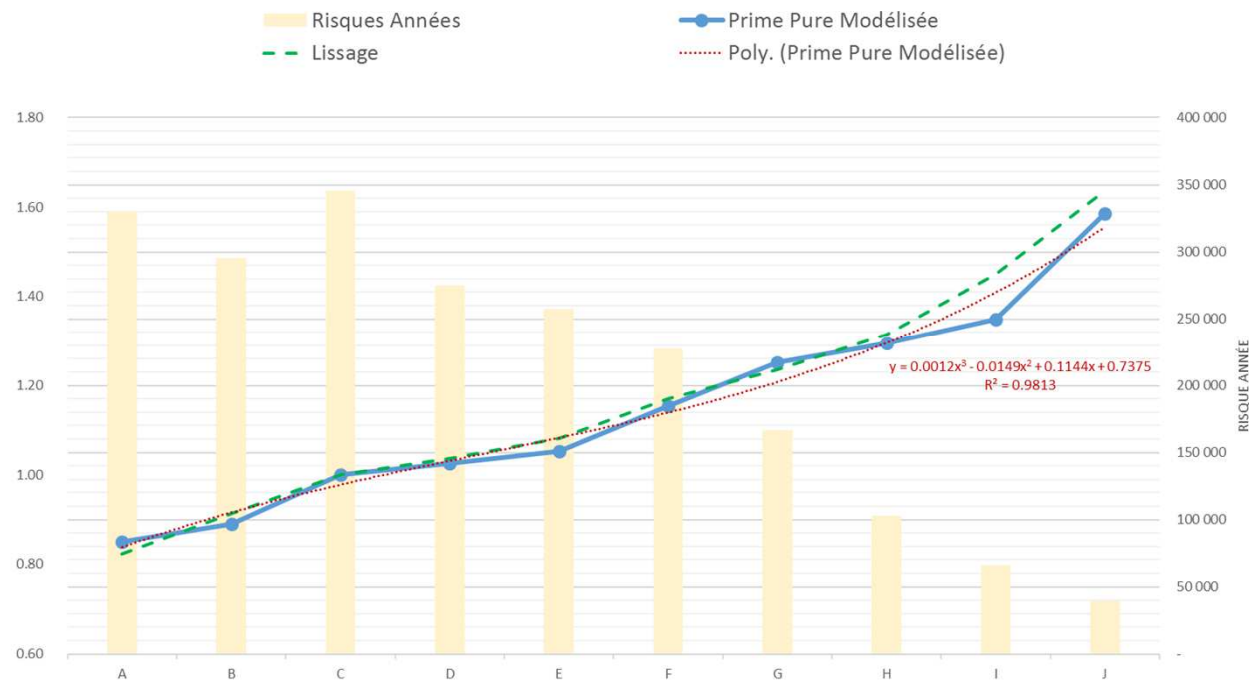
---

- ❖ Les MLG sont adaptés pour modéliser une sinistralité « standard » (« attritionnelle »).
- ❖ Les sinistres « graves », rares, sont plus difficiles à ajuster.
  
- ❖ Ce que l'on peut faire, en pratique :
  - Modélisation de l'ensemble des sinistres, mais avec écrêtement de la charge.
  - Puis application d'une charge « sur-crête » moyenne, ou bien répartition de la sur-crête selon la probabilité de survenance d'un grave selon le profil du contrat (méthodes de scoring).
  - On peut aussi choisir de construire deux MLG: l'un pour la sinistralité attritionnelle, l'autre pour la sinistralité grave.



# Lissage des coefficients

- ❖ Les résultats des MLG sur la fréquence et le coût des sinistres permettent de constituer une prime pure théorique propre à chaque individu.
- ❖ Dans la pratique, les coefficients des modèles ne sont généralement pas utilisés tels quels, mais peuvent être soumis à un lissage permettant par exemple de prendre en compte les contraintes apportées par la politique tarifaire.



---

*Modéliser un tarif avec les Modèles Linéaires Généralisés*

## **EXEMPLE : TARIFICATION D'UN PRODUIT D'ASSURANCE AUTOMOBILE**

- ⇒ Introduction
- ⇒ Modélisation de la fréquence
- ⇒ Modélisation du coût
- ⇒ Modélisation de la prime pure

## Exemple : tarification d'un produit d'assurance automobile

- ❖ Quelles sont les informations nécessaires à la constitution d'un tarif auto?

### CONDUCTEUR

Age	Ancienneté de permis
Ancienneté d'assurance	Conducteur complémentaire

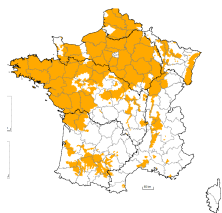


### VEHICULE

Groupe/classe	Cout réparation
ancienneté	garage



### ZONE GEOGRAPHIQUE



Risque circulation
Risque vol
Risque bris de glace
Lieu de stationnement

### CONTRAT

Sinistralité antérieure	
Franchise	CRM (bonus)
Garanties	



Etc....

## Exemple : tarification d'un produit d'assurance automobile

---

### ❖ Quelles garanties peuvent être proposées?

- Responsabilité civile : c'est la seule qui est obligatoire
- Individuelle du conducteur
- Vol
- Bris de glace
- Dommages
- Assistance au véhicule et aux personnes
- Objets et marchandises transportés

# Modélisation de la Fréquence

- ❖ On modélise la garantie « G ».
- ❖ On cherche à modéliser la variable « nb\_sinistre » en fonction de critères segmentant le portefeuille.
- ❖ On opte pour un modèle multiplicatif, avec fonction de lien LOG et loi de Poisson
- ❖ On met en « offset » la variable « LOG(nb risque années) », qui permet de rapporter les nombres de sinistres à la durée d'observation

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	3.80E+06	481060	0.1282
Scaled Deviance	3.80E+06	3753709	1
Pearson Chi-Square	3.80E+06	4192989	1.117
Scaled Pearson X2	3.80E+06	3.3E+07	8.7161
Log Likelihood		-2475794	

*Correction de la surdispersion avec l'option DSCALE*

LR Statistics For Type 3 Analysis			
Source	Num DF	Chi-Square	Pr > ChiSq
age_cond	5	9755.66	<.0001
zone_circu	6	2918.71	<.0001
SGT	4	2236.04	<.0001
anveh	8	1680.55	<.0001
.....			

*Toutes ces covariables sont significativement influentes.*

# Modélisation de la Fréquence

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	<b>-2.7823</b>	0.0107	-2.8032	-2.7614	125776	<.0001
age_cond	a.18-20	1	<b>-0.2176</b>	0.0527	-0.3209	-0.1142	17.03	<.0001
age_cond	b.21-25	1	<b>-0.282</b>	0.0097	-0.3011	-0.2629	836.94	<.0001
age_cond	c.26-35	1	<b>-0.2015</b>	0.0054	-0.2121	-0.1909	1389.28	<.0001
age_cond	d.36-50	0	<b>0</b>	0	0	0		
age_cond	e.51-69	1	<b>0.1059</b>	0.0042	0.0976	0.1142	623.17	<.0001
age_cond	f.>=70	1	<b>0.5377</b>	0.0059	0.5261	0.5494	8186.35	<.0001
anveh	a.0-1	1	<b>-0.0167</b>	0.0078	-0.032	-0.0015	4.65	0.031
anveh	b.2-5	1	<b>0.066</b>	0.0077	0.0509	0.081	73.83	<.0001
anveh	c.6-9	1	<b>0.1387</b>	0.0072	0.1245	0.1528	368.54	<.0001
anveh	d.10-15	1	<b>0.1511</b>	0.0067	0.138	0.1641	514.69	<.0001
anveh	e.>=16	0	<b>0</b>	0	0	0		
SGT	A	1	<b>-0.082</b>	0.0088	-0.0992	-0.0648	87.41	<.0001
SGT	B	1	<b>-0.0619</b>	0.0043	-0.0703	-0.0534	207.48	<.0001
SGT	H	1	<b>0.2071</b>	0.0057	0.1959	0.2184	1302.18	<.0001
SGT	M	0	<b>0</b>	0	0	0		
SGT	autre	1	<b>0.1286</b>	0.0074	0.114	0.1431	300.5	<.0001
zone_circu	1	1	<b>-0.0727</b>	0.0056	-0.0836	-0.0618	171.14	<.0001
zone_circu	2	0	<b>0</b>	0	0	0		
zone_circu	3	1	<b>0.032</b>	0.0054	0.0214	0.0426	35.09	<.0001
zone_circu	4	1	<b>0.0585</b>	0.0055	0.0477	0.0693	112.69	<.0001
zone_circu	5	1	<b>0.1155</b>	0.0057	0.1043	0.1267	410.69	<.0001
zone_circu	6	1	<b>0.2025</b>	0.0063	0.1902	0.2147	1048.04	<.0001
zone_circu	>=7	1	<b>0.2358</b>	0.0072	0.2217	0.2499	1069.92	<.0001
....								

En cas de modalités non significativement différentes de la valeur 0, on procède généralement à des agrégations de modalités.

*Toutes les modalités des covariables ont un effet significativement différent de la modalité de référence (valeur 0).*

# Modélisation de la Fréquence

---

1. Quelle est la fréquence de sinistre estimée pour un conducteur âgé de 30 ans, dont le véhicule a été immatriculé en 2010, correspond au segment M et est garé habituellement à l'adresse du conducteur, en zone 3?
2. Quel est le profil le plus risqué? Quelle est sa fréquence de sinistre annuelle théorique?
3. Quel est le profil le moins risqué? Quelle est sa fréquence de sinistre annuelle théorique?
4. Tracer la pente de l'effet fréquence pour la variable AGE\_COND

# Modélisation de la Fréquence

---

## Réponse

1. Rappelons que la fonction de lien est LOG, donc on doit utiliser l'exponentielle des coefficients.

Ainsi,  $F1 = \exp (-2,7823-0,2015+0,1387+0+0,032) = 5,54\%$

2. Le plus risqué : conducteur de plus de 70 ans, conduisant un véhicule de 10 à 15 ans, de segment H et circulant en zone à risque  $\geq 7$

$F = 19,2\%$

3. Le moins risqué : conducteur de 21 à 25 ans, conduisant un véhicule de moins de 2 ans, de segment A et circulant en zone à risque de niveau 1

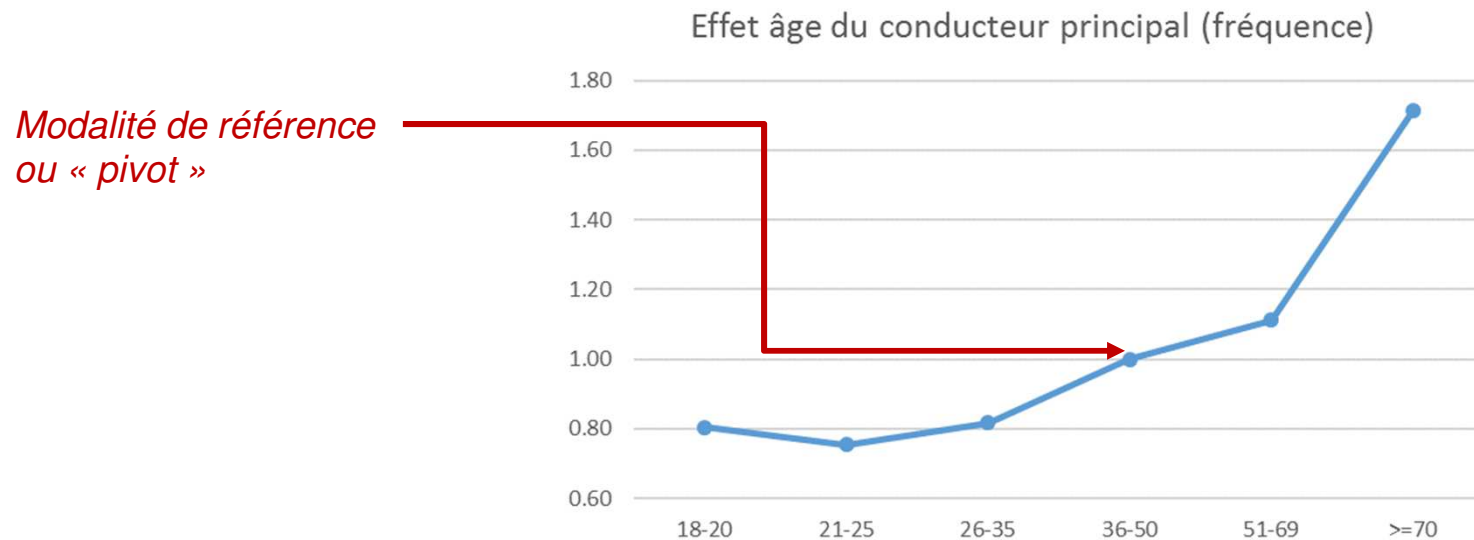
$F = 3,9\%$



# Modélisation de la Fréquence

## 4. Effet âge du conducteur principal:

variable	modalités	$\beta_k$	$\exp(\beta_k)$
age_cond	a.18-20	-0.2176	0.8044
	b.21-25	-0.2820	0.7543
	c.26-35	-0.2015	0.8175
	d.36-50	0.0000	1.0000
	e.51-69	0.1059	1.1117
	f.>=70	0.5377	1.7121



# Modélisation du coût

- ❖ On modélise la même garantie « G ».
- ❖ On cherche à présent à modéliser la variable « cout\_sinistre » en fonction de critères segmentant le portefeuille.
- ❖ On opte pour un modèle multiplicatif, avec fonction de lien LOG et loi Gamma
- ❖ Les montants sont écrêtés.

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	7.1666	0.01	7.1471	7.1862	516106	<.0001
age_cond2	a.18-25	1	0.0401	0.0109	0.0188	0.0614	13.57	0.0002
age_cond2	c.26-50	0	0	0	0	0	.	.
age_cond2	d.>50	1	-0.0293	0.0056	-0.0404	-0.0183	27.28	<.0001
anveh2	a.0-3	0	0	0	0	0	.	.
anveh2	b.4-7	1	0.0179	0.0071	0.0041	0.0317	6.43	0.0112
anveh2	c.8-9	1	0.0426	0.0077	0.0274	0.0578	30.19	<.0001
anveh2	d.10-15	1	0.0297	0.0081	0.0139	0.0455	13.59	0.0002
anveh2	e.>=16	1	0.0338	0.0084	0.0174	0.0503	16.22	<.0001
quartier	A	1	-0.0532	0.0089	-0.0705	-0.0358	36.06	<.0001
quartier	B	1	-0.0727	0.0129	-0.0979	-0.0475	31.9	<.0001
quartier	G	1	-0.0776	0.0201	-0.117	-0.0383	14.95	0.0001
quartier	I-C-E-F	0	0	0	0	0	.	.
quartier	J-K	1	-0.0227	0.0064	-0.0353	-0.0101	12.49	0.0004
....								

*Coefficients estimés par le modèle.  
Toutes les modalités ont un effet significativement différent de la modalité de référence (valeur 0).*

# Modélisation du coût

---

1. Quel est le profil le plus risqué? Quel est son coût moyen théorique?
2. Quel est le profil le moins risqué? Quel est son coût moyen théorique?
  
1. Quelle est le coût moyen estimé d'un sinistre pour un conducteur âgé de 30 ans, dont le véhicule a été immatriculé en 2010, correspond au segment M et est garé habituellement à l'adresse du conducteur, en zone 3, dans un quartier de type C?

Quelle est sa prime pure?

2. Tracer la pente de l'effet coût pour la variable AGE\_COND, puis la pente de la prime pure

# Modélisation du coût

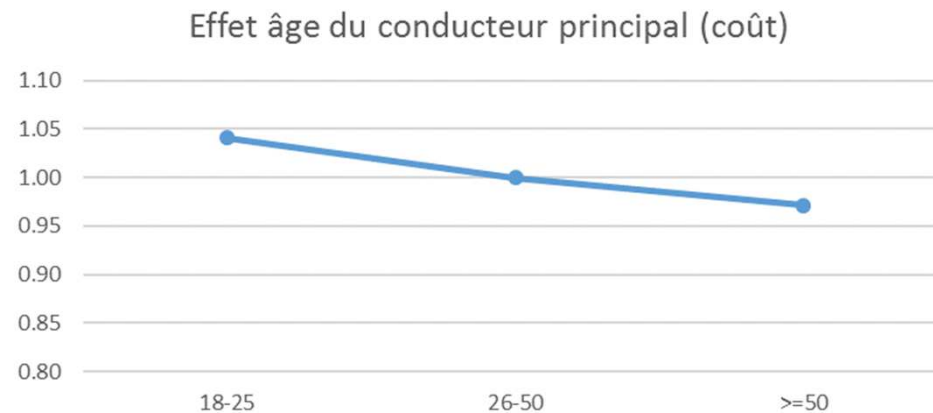
---

## Réponses

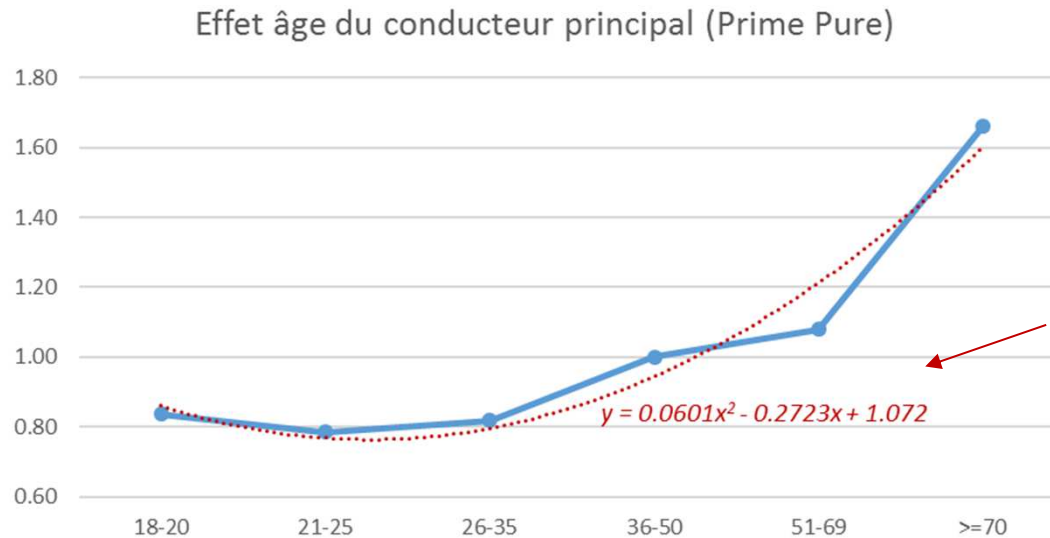
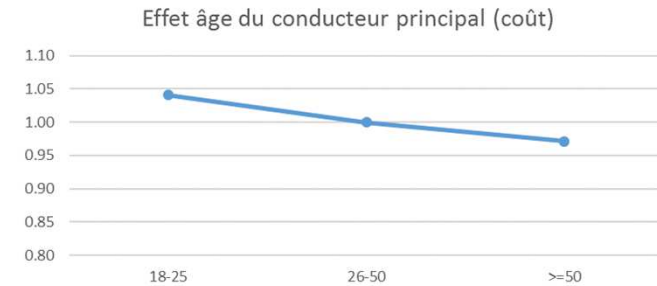
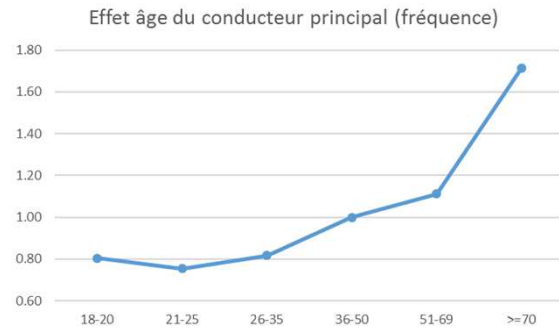
1. Le plus risqué :  $CM = \exp(7,1666+0,0401+0,0426+0)=1407\text{€}$
2. Le moins risqué :  $CM= 1164\text{€}$
3.  $CM= \exp(7,1666+0+0,0179+0)=1319\text{€}$

Prime Pure =  $F \cdot CM = 5,54\% \cdot 1319 = 73\text{€}$

4. Zoom sur la variable « Age du Conducteur principal » :



# Modélisation de la prime pure



*Lissage polynomial*

---

*Modéliser un tarif avec les Modèles Linéaires Généralisés*

# **ANNEXE**

# Taxes et contributions sur les cotisations d'assurance

---

- ❖ Les cotisations d'assurance ne sont pas soumises au régime de la TVA, mais à la taxe d'assurance (articles 991 et suivants du Code général des impôts).
- ❖ Le montant de la taxe est calculé en fonction d'un taux qui varie selon les catégories d'assurances ou de souscripteurs.

- Exemple: RC automobile – véhicule personnel de moins de 3.5t

Taxe fiscale	18 %
Contribution prévue par le Code de la Sécurité sociale	15 %
Contribution au Fonds de garantie des assurances obligatoires de dommages	1,2 %
Contribution supplémentaire au Fonds de garantie des assurances obligatoires	0,8%

soit au total, pour la garantie RC automobile : 35% de taxe

## ❖ Contributions:

- Fond de garantie des assurances obligatoires de dommages(FGAO)
  - ✓ Créée en 1951 pour l'assurance automobile: indemniser les victimes d'accidents de la circulation dont les auteurs ne sont pas assurés ou pas identifiés
  - ✓ Élargi et devenu FGAO en 2003
- Fonds de garantie contre les actes de terrorisme et autres infractions(FGTI)
  - ✓ Créé en 1986 pour indemniser les victimes de terrorisme, étendu en1990 à l'indemnisation des victimes d'infractions de droit commun et, en 2008, à l'aide au recouvrement des dommages et intérêts obtenus par une décision de justice



# Taxes et contributions sur les cotisations d'assurance

---

## ❖ Autres exemples :

- Autres garanties relatives aux véhicules (dommages, assistance aux véhicules ...)

Taxe fiscale	18 %
Contribution au Fonds de garantie contre les actes de terrorisme et autres infractions	3,30 euros par contrat

- Risques des particuliers

Taxe fiscale	30 %
Contribution au Fonds de garantie contre les actes de terrorisme et autres infractions	3,30 euros par contrat

- Autres assurance ...

Taxe fiscale	9 %
Contribution au Fonds de garantie contre les actes de terrorisme et autres infractions (pour les contrats d'assurance de biens)	3,30 euros par contrat

- ...

- ## ❖ Toutefois de nombreux cas particuliers existent, certaines assurances bénéficient d'exonération (crédit à l'export, dépendance, ...) d'autres de taux spécifiques (assurance maladie, risques agricoles, ...)



# Quelques syntaxes SAS

---

## ❖ MLG avec la proc GENMOD

```
PROC GENMOD DATA = Base ;  
CLASS Variables Qualitatives / param=ref;  
MODEL Variable Réponse = Variables explicatives  
/
```

```
OFFSET = Variable offset
```

```
DIST = Loi de la variable réponse
```

```
LINK = Fonction de lien
```

```
DSCALE
```

```
TYPE3 ;
```

```
OUTPUT OUT= NomTable
```


```
P=pred
```

```
stdxbeta=std /*ecart type*/
```

```
resdev=residus STDRESDEV=nom_res /*résidu de deviance et son écart type*/
```

```
RUN ;
```

Récupérer dans une table SAS  
certaines informations relatives aux  
individus (valeur prédite de la moyenne,  
valeur du prédicteur linéaire, résidu de la  
déviante)



# Quelques syntaxes SAS

---

- ❖ Pour tester l'adéquation à une distribution de probabilité : PROC UNIVARIATE

Par exemple, tester la normalité des résidus de la déviance:

```
proc univariate data=NomTable plots normal;  
var nom_res;  
histogram / normal;  
qqplot;  
run;
```

```
/* Graphique des résidus de la déviance */  
PROC GPLOT data=NomTable;  
plot residus*pred;  
run;
```

# Les MLG avec le logiciel R

---

## ❖ Fonction glm()

```
> glm( var à expliquer ~ var explicative1 + var  
explicative2 + ... , table , type de loi (fonction de  
lien), ... )
```

## ➤ Exemple:

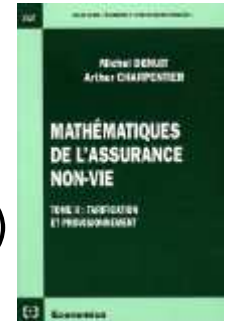
```
> modele <- glm(Y ~ X1 +X2, data=table,  
family=Gamma("log"))
```

# Quelques sources et références...

---

## ❖ Ouvrages et site internet de A. Charpentier

- <http://freakonometrics.hypotheses.org/>
- « Mathématiques de l'assurance non vie" (Economica,2005)



## ❖ Site internet d'Olivier Decourt

- <http://www.od-datamining.com/>

## ❖ Support SAS

- <http://support.sas.com/>

## ❖ Documentation R

---