METHODES DE TARIFICATION EN ASSURANCE

Partie 1: GLM, tarification a priori

Xavier Milhaud

Affilé à l'Institut de Mathématique de Marseille (I2M),
MCF à Aix-Marseille Université

ENSEA Abidjan - Février 2023

Plan du cours

Organisation - Objectifs

Le travail se répartit comme suit:

- 12h de cours sur les GLM;
- 8h de cours sur la théorie de la crédibilité.

L'objectif est **d'avoir une idée des difficultés rencontrées en pratique** et de connaitre certaines **méthodes pour les traiter**.

La mise en pratique sera réalisée sur ordinateur, à l'aide du logiciel R.

1 Introduction et rappels des concepts essentiels

Contrat d'assurance - Tarification

Une police d'assurance est un contrat entre deux parties :

- → l'assuré, détenteur du contrat;
- → l'assureur, pourvoyeur du contrat.

En échange de la couverture d'un risque par l'assureur, l'assuré verse une **prime** d'assurance.

En cas de sinistre, le bénéficiaire du contrat reçoit le montant contractuel prévu en cas de survenance du sinistre.

Ainsi le risque économique initialement supporté par l'assuré est transféré vers l'assureur.

La mutualisation induite par la souscription de nombreux contrats au sein d'une compagnie d'assurance permet l'utilisation grossière de la **loi des grands nombres**.

En effet,

- → un portefeuille d'assurance couvre un risque en particulier: les pertes sont considérées être de même loi de probabilité;
- → les contrats sont a priori indépendants les uns des autres.

Ces propriétés doivent permettre à l'assureur de **prédire avec une précision relative** les pertes encourues pour une période donnée.

Soit un portefeuille d'assurance contenant I polices. Notons la loi du $i^{\text{ème}}$ contrat S_i (perte), et la loi des pertes agrégées S_I.

La LFGN stipule la CV presque sûre de la moyenne empirique de pertes i.i.d., notée $\bar{S}_I = \frac{1}{I}\sum_{i=1}^{I} S_i$, vers l'espérance de la loi:

$$\boxed{\bar{S}_I \xrightarrow[n\to\infty]{p.s.} \mathbb{E}[S_i] = \mu.}$$

Ou encore: $\mathbb{P}\left(\lim_{I\to\infty} \bar{S}_I = \mu\right) = 1.$

Prime technique - prime commerciale

Ce résultat est à l'origine du **principe général de tarification**: la prime vaut au moins μ, aussi appelée prime pure du contrat.

En pratique l'assureur applique des **chargements** à cette prime, car mathématiquement sa ruine est certaine à horizon infini dès lors que la tarification respecte le strict principe d'équivalence.

La prime d'assurance Π_i se décompose donc en $+$sieurs parties:

- $\rightarrow$ la prime pure $\mathbb{E}[S_i]$;
- $\rightarrow$ $+$ les chargements techniques (ou marge de risque MR_i):

$$\Pi_i = \mathbb{E}[S_i] + MR(S_i);$$

- $\rightarrow$ les coûts:

- acquisition,
- administration et gestion du contrat,
- rémunération d'intermédiaires (courtiers, ...).

La stratégie de la compagnie peut également jouer sur la hauteur de ces chargements.

Objectif de l'assureur:

Mettre en place une tarification segmentée tout en conservant le principe de mutualisation.

Cela lui permettra de déterminer

$\rightarrow$ la loi de probabilité de son résultat futur,

$\rightarrow$ sa probabilité de ruine.

Les modélisations concernent la **détermination de la prime pure**.

Contexte d'étude des risques en assurance

Un assureur essaie généralement d'avoir la meilleure connaissance possible de la fréquence et du coût des sinistres.

Les bases de données des assureurs comportent un ensemble d'informations sur les

- caractéristiques de l'assuré: sexe, âge, CSP, adresse...
- options du contrat: franchise, ...
- conditions de marché: indices macroéconomiques, conjoncture, concurrence...

Ces informations **jouent un rôle important** dans la détermination et dans l'estimation des paramètres des modèles mis en place!

Allure d'une base de données (ex: auto)

```
> head(myData, n=16)
```

	PERMIS	ACV	SEX	STATUT	CSP	USAGE	AGECOND	...	GARAGE	CHARGE
1	245	10	F	C	50	2	40	...	3	0
2	348	10	F	A	50	1	63	...	3	0
3	16	10	F	C	26	2	20	...	3	0
4	291	10	F	A	50	1	56	...	3	0
5	123	10	F	A	50	1	29	...	3	0
6	295	10	F	A	37	1	43	...	3	0
7	24	10	F	A	50	2	21	...	3	0
8	181	9	F	A	50	3	35	...	3	0
9	157	10	M	C	55	1	31	...	3	0
10	338	10	M	C	1	2	48	...	2	179
11	20	10	M	C	26	2	19	...	3	0
12	208	10	F	A	50	2	39	...	3	0
13	127	10	F	A	37	1	29	...	1	0
14	93	7	F	C	50	2	39	...	3	0
15	134	10	F	A	50	1	36	...	3	0
16	416	10	F	C	50	1	60	...	3	0

Quelques principes de base en tarification

Soit S_i la somme annuelle des sinistres du contrat i. Le nb N_i de sinistres est une v.a. considérée $\perp\!\!\!\perp$ des coûts Y_{ik}, eux-même i.i.d.:

$$S_i = \begin{cases} 0 & \text{si } N_i = 0 \\ Y_{i1} + \ldots + Y_{in} & \text{si } N_i = n. \end{cases} \quad \Leftrightarrow \quad S_i = \sum_{k=1}^{N_i} Y_{ik}$$

Ainsi, $\mathbb{E}_{\mathbb{P}}[S_i] = \mathbb{E}_{\mathbb{P}}[N_i] \times \mathbb{E}_{\mathbb{P}}[Y_{ik}]$.

En réalité, N_i est souvent **conditionnellement** $\perp\!\!\!\perp$ à Y_i, donc

$$\mathbb{E}_{\mathbb{P}}[S_i \,|\, \mathcal{X}_i] = \mathbb{E}_{\mathbb{P}}[N_i \,|\, \mathcal{X}_i] \,.\, \mathbb{E}_{\mathbb{P}}[Y_{ik} \,|\, \mathcal{X}_i],$$

où $\mathcal{X}_i$ est un ensemble d'informations.

Le principe de la tarification est d'approcher $\mathcal{X}$ par un **proxy**.
Ce proxy correspond aux info. indiv. $\rightarrow$ **variables explicatives**:

$$\Rightarrow \text{c'est le contexte des modèles de régression.}$$

Supposons que l'assureur dispose de J facteurs explicatifs du risque, notés $\{X_1, \ldots, X_J\}$, on obtient alors la formule

$$\mathbb{E}_{\mathbb{P}}[S \,|\, X_1, \ldots, X_J] = \mathbb{E}_{\mathbb{P}}[N \,|\, X_1, \ldots, X_J] \,.\, \mathbb{E}_{\mathbb{P}}[Y \,|\, X_1, \ldots, X_J].$$

Le problème est donc d'obtenir

- $\mathbb{E}_{\mathbb{P}}[N \,|\, X_1, \ldots, X_J]$: estimation de la loi de N.
- $\mathbb{E}_{\mathbb{P}}[Y \,|\, X_1, \ldots, X_J]$: idem.

En économétrie, on cherche à estimer $\mathbb{E}_{\mathbb{P}}[Z \,|\, X_1, \ldots, X_J]$ par une fonction des facteurs explicatifs notée $\Phi(X_1, \ldots, X_J)$.

En économétrie **linéaire**, on a coutûme de supposer que

$$Z \,|\, X_1, \ldots, X_J \quad \sim \quad \mathcal{N}(\beta_0 + \beta_1 X_1 + \ldots + \beta_J X_J,\ \sigma^2).$$

En notant $\mathbf{X} = (1, X_1, \ldots, X_J)^T$ le **vecteur des facteurs de risque** et $\beta = (\beta_0, \beta_1, \ldots, \beta_J)^T$ les **coefficients** de régression, on peut simplifier cette écriture sous forme matricielle:

$$Z \,|\, \mathbf{X} \quad \sim \quad \mathcal{N}(\mathbf{X}^T \beta,\ \sigma^2).$$

Problème: le modèle linéaire est rarement adapté en assurance...
Alternative: besoin de supposer relations non-linéaires $\Rightarrow$ GLM.

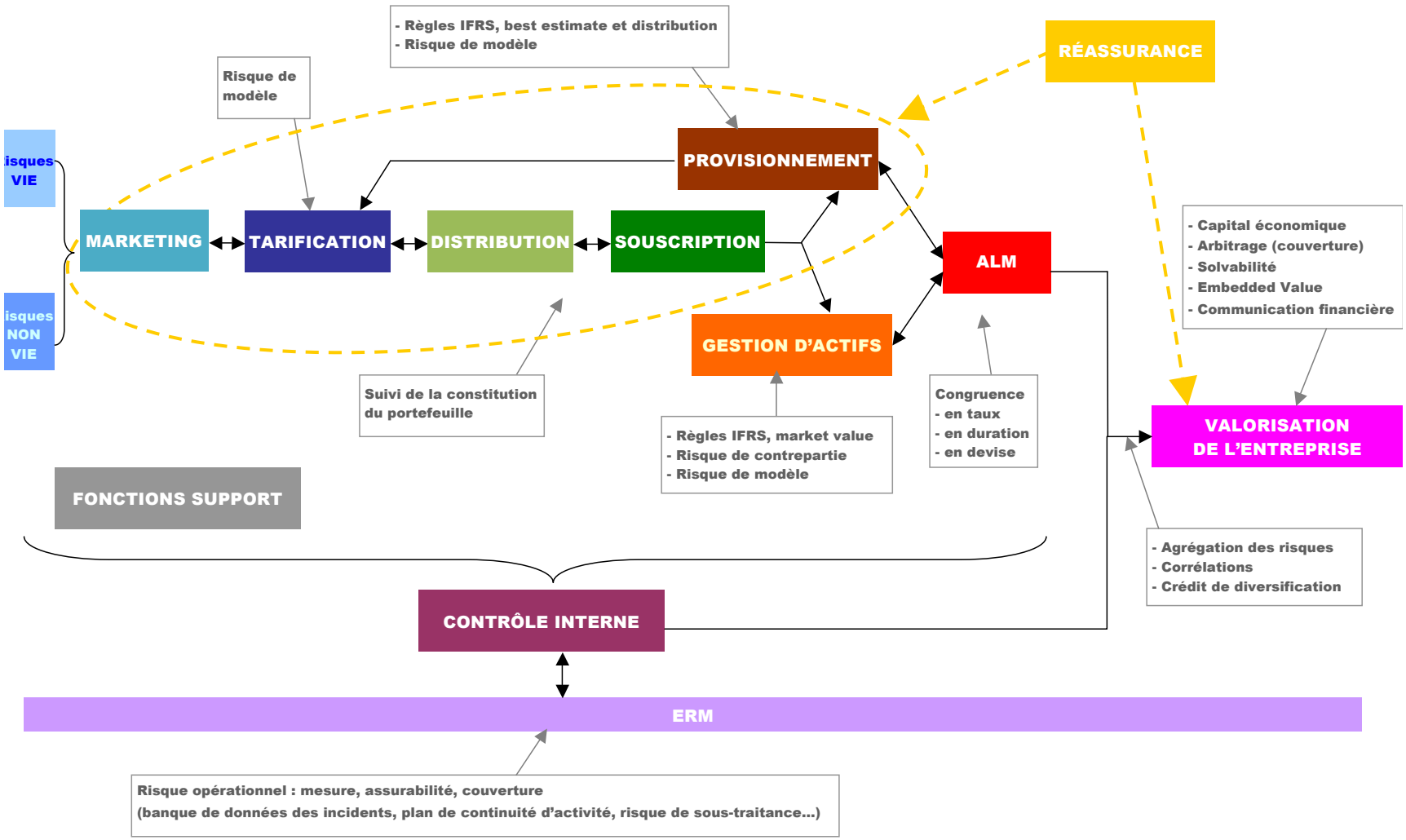
Dangers d'une mauvaise tarification

Se tromper dans la tarification d'un produit peut avoir plusieurs conséquences dommageables:

- comme cela est souvent lié à la segmentation, il y a un risque de composition du portefeuille (bons et mauvais risques);
- investir dans 1 politique de vente (marketing, ...) mal adaptée;
- impact néfaste sur la concurrence, déficit d'image;
- mauvaise évaluation de la marge de risque, et donc in fine du provisionnement: (pour rappel, $S_I = \sum_i S_i$)

$$VaR_\alpha(S_I) = \inf\{s \in \mathbb{R}^+ : \ \mathbb{P}(S_I > s) \leq (1-\alpha)\}$$

La chaîne de gestion des risques dans l'assurance



Des difficultés liées à la réglementation

La législation a également un impact en termes de segmentation et de tarification. L'exemple récent le plus célèbre :

"Les compagnies d'assurances ne pourront plus, à partir du 21 décembre 2012, prendre en considération le critère du sexe pour calculer les primes et prestations d'assurances dans leurs contrats." a jugé la Cour de justice de l'UE.

Source: `http://www.lemonde.fr/economie/article/2011/03/02/les-assureurs-ne-pourront-plus-appliquer-des-tarifs-differents-selon-le-sexe_1487077_3234.html`

Remarque: ce n'est pas le cas pour le provisionnement...

Etapes statistiques dans la tarification

1. Modélisation de la fréquence par un GLM adapté (choix d'une loi pour la réponse, intégration des covariables), cela donne

$$\mathbb{E}[N \mid \mathbf{X}] = f_1(\mathbf{X}\beta)$$

2. Modélisation du coût par un autre GLM adapté, on obtient

$$\mathbb{E}[Y \mid \mathbf{X}^{'}] = f_2(\mathbf{X}^{'}\beta)$$

3. Synthèse pour en déduire la prime (pure):

$$\mathbb{E}[S_i \mid \mathbf{X}, \mathbf{X}^{'}] = E[N \mid \mathbf{X}] \times E[Y \mid \mathbf{X}^{'}]$$

La potentielle propagation des erreurs

En construisant deux modèles (1 pour la fréquence et 1 pour la sévérité), on prend le risque de propager des erreurs...

Parfois il vaut mieux essayer de construire un unique modèle qui rende compte à la fois de la fréquence et de la sévérité: cela **dépend de la qualité d'adéquation de la loi de fréquence notamment**.

En réalité dans cette ultime approche, on perd l'info sur le nb de sinistres et on s'intéresse à la charge totale par contrat. La masse en 0 (contrats non-sinistrés) induit des difficultés de calibration.

Gestion / utilisation des données

La sinistralité se décompose généralement en trois typologies de sinistre:

- attritionnels: haute fréquence, petite sévérité;
- graves: basse fréquence, grande sévérité;
- CAT: très basse fréquence, sévérité extrême.

Nécessité de séparer ces données en amont car les **GLM ne fonctionnent que sur les sinistres attritionnels (voire graves)** à cause des queues des distributions des lois utilisées.

2 Applications classiques des GLM en assurance

- Assurance non Vie
- Assurance Vie

Quelques applications en assurance IARD

L'usage des GLM est ancré depuis longtemps dans les moeurs.
On peut citer parmi les domaines concernés:

- **assurance santé**: remboursements soins, frais d'hospitalisation;
- **assurance auto** / **moto**: dommages matériels, vol, ...;
- **assurance Multi-Risques Habitation** (MRH): incendie, vol, dégâts des eaux, ...
- **assurance Responsabilité Civile** (RC): dommages à autrui.

Les cas de la RC, de l'assurance CATNAT et de la réass. IARD sont un peu $\neq$ car font intervenir des montants CAT en général.

Les applications en VIE

On se sert aussi des GLM en Vie, notamment en

- **épargne**: essentiellement du risque comportemental sur les produits en taux garantis (euro) ou non (UC);
- **prévoyance**: DC, LTC (Long-Term Care: dépendance), CI (Critical Illness: maladies redoutées), incap/inval. ;
- **réassurance vie**: même remarque qu'en non vie.

Remarque: de par la nature des contrats, il y a souvent une dimension temporelle dans la modélisation en Vie qui $\nexists$ en non-vie $\rightarrow$ modèles de durée.

Exemple en risque décès (DC): Lee Carter

Lee and Carter (1992)

C'est le **modèle le plus utilisé en mortalité** (longévité):

$$log(\mu_x(t)) = \alpha_x + \beta_x \kappa(t) + \epsilon_x(t)$$

- x est l'âge, t l'année;
- $\mu_x(t)$ est le taux de mortalité instantané l'année t à l'âge x;
- α_x: structure de la mortalité en fonction de l'âge;
- $\kappa(t)$: vitesse d'amélioration de la mortalité (série temp.);
- β_x: la vitesse d'amélioration a des impacts $\neq$ selon l'âge;
- les résidus $\epsilon_x(t) \sim \mathcal{N}(0, \sigma^2)$.

Exemple 2: modèle de Brass

Brass (1964), Brass and Macrae (1984)

C'est un **modèle relationnel** basé sur la régression logistique:

$$ln\left(\frac{q^{exp}(x,t)}{1-q^{exp}(x,t)}\right) = a + b \times ln\left(\frac{q^{ref}(x,t)}{1-q^{ref}(x,t)}\right)$$

où

- x est l'âge de la personne, t est le facteur temporel,
- q^{ref} est une table de mortalité de référence,
- q^{exp} est la table de mortalité d'expérience.

Calibre les coef. (a, b) pour établir le passage d'1 table à l'autre, par ex. d'une population nationale à une population d'assurés.

3 Les Modèles Linéaires Généralisés (GLM)

- Les GLM: brefs rappels
- Caractérisation et formalisation
- Validation
- Implémentation
- Lecture des résultats de la calibration
- Sélection de modèle et de variables

Intérêt des GLM

Les GLM permettent de

- modéliser des réponses diverses $\in \mathbb{R}, \mathbb{R}^+, \mathbb{N}, [0,1], ...$;
- intégrer toute type d'**information exogène** susceptible d'influer sur la variable dépendante (réponse Y),
- **quantifier l'impact** des facteurs de risque $\mathbf{X}$ sur N et Y (sens/intensité).

Ils nécessitent d'introduire deux hypothèses fondamentales:

- les individus Y_i sont $\perp\!\!\!\perp$ entre eux (rq: si les indiv. étaient corrélés, cela résulterait aussi à avoir – d'indiv., donc $n \searrow$);
- les variables explicatives $\mathbf{X}$ sont $\perp\!\!\!\perp$ deux à deux.

Attention à la notion de corrélation entre variables

$\exists$ plusieurs mesures de dépendance, e.g. corrélation de rang (Kendall, Spearman). La $+$ répandu est Pearson,

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y},$$

où $\mu_X = E[X]$ et σ_X est l'écart-type de X.

Mesure la corrél. **linéaire**. En effet, considérons la v.a. X telle que $X \sim \mathcal{N}(0,1)$. Ainsi $\mu_X = 0$, et $\mu_{X^3} = 0$. Notons $Y = X^2$, on a

$$\rho_{X,Y} = \frac{E[(X-\mu_X)(X^2-\mu_{X^2})]}{\sigma_X \sigma_{X^2}} = \frac{\mu_{X^3} - \mu_X \mu_{X^2}}{\sigma_X \sigma_{X^2}} = 0.$$

Corrélation nulle alors que X et X^2 parfaitement corrélées!

3 Les Modèles Linéaires Généralisés (GLM)

- Les GLM: brefs rappels
- Caractérisation et formalisation
- Validation
- Implémentation
- Lecture des résultats de la calibration
- Sélection de modèle et de variables

Composants d'un GLM (i^e individu)

McCullagh and Nelder (1989)

1. La **loi de la réponse aléatoire** Y_i: par hyp. elle $\in$ à une *distribution de la famille exponentielle.*

2. Le **prédicteur** $\eta_i = \sum_{j=1}^{J} \beta_j X_{ij}$, linéaire et déterministe: *les facteurs de risque explicatifs le constituent.*

3. La **fonction de lien** g: monotone, dérivable, inversible t.q.

$$g(\mathbb{E}[Y_i]) = \eta_i.$$

Ex. du modèle linéaire: $g = Id \quad \eta_i = \sum_{j=1}^{J} \beta_j X_{ij} \quad Y_i \sim \mathcal{N}(\eta_i, \sigma^2)$.

Effets additifs VS effets multiplicatifs

- Régression linéaire standard:
  - $\mathbb{E}[Y_i] = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p$;
  - l'influence des facteurs de risque (variables explicatives) a un effet additif sur la réponse Y_i.
  - Y_i est un réel, et peut notamment donc être négatif.

- Régression log-poisson:
  - $\log(\mathbb{E}[Y_i]) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p$;
  - d'où $\mathbb{E}[Y_i] = \exp(\beta_0)\ \exp(\beta_1 X_1) \times ... \times \exp(\beta_p X_p)$;
  - les effets sont multiplicatifs sur la réponse;
  - la réponse ne peut être que positive!

Famille exponentielle

→ La représentation exponentielle facilite la dérivation de résultats.

Les GLM sont issus de la famille exponentielle, dont la densité est couramment exprimée par

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

où $a(.)$, $b(.)$ et $c(.)$ sont des fonctions spécifiques suivant le modèle considéré, et θ et ϕ sont les paramètres.

La fonction $a(\phi)$ est de la forme $\frac{\phi}{\omega}$, où

- ω correspond à un poids (une “exposition” dans le jargon),
- très souvent constant égal à 1 (cas individuel).

Vocabulaire

- lien canonique: permet de vérifier $\theta_i = \mu_i$ (où $\mu_i = \mathbb{E}[Y_i]$)
- paramètre de tendance: le paramètre θ_i;
- paramètre de dispersion: le paramètre ϕ_i.

On peut facilement exprimer les quantités clefs pour l'inférence:

- Log-vraisemblance pour une observation y_i:

$$\log L(\theta, \phi; y_i) = \log f_Y(y_i; \theta, \phi) = \frac{y_i\theta - b(\theta)}{a(\phi)} + c(y_i, \phi).$$

- Espérance de la réponse: $\mathbb{E}[Y_i] = \mu_i = b'(\theta_i)$;
- Variance: $\mathbb{V}ar[Y_i] = a(\phi)\, b''(\theta_i) = a(\phi) \underbrace{V(\mu_i)}_{\text{fn. variance}}$.

Loi de Y:	Normale $\mathcal{N}(\mu, \sigma^2)$	Binomiale $B(n, \mu)$	Poisson $\mathcal{P}(\mu)$	Gamma $\mathcal{G}(\mu, \nu)$	Inverse Gaussienne $\mathcal{IN}(\mu, \sigma^2)$
Supports	$y \in \mathbb{R}$ $\mu \in \mathbb{R}$ $\sigma^2 \in \mathbb{R}^{+*}$	$y \in [\![0, n]\!]$ $n \in \mathbb{N}^*$ $\mu \in [0, 1]$	$y \in \mathbb{N}$ $\mu \in \mathbb{R}^+$	$y \in \mathbb{R}^+$ $\mu \in \mathbb{R}^{+*}$ $\nu \in \mathbb{R}^{+*}$	$y \in \mathbb{R}^+$ $\mu \in \mathbb{R}^{+*}$ $\sigma^2 \in \mathbb{R}^{+*}$
Tendance $\theta(\mu)$	μ	$\log[\mu/(1-\mu)]$	$\log \mu$	$-\mu^{-1}$	$-(2\mu^2)^{-1}$
Support de θ	$\theta \in \mathbb{R}$	$\theta \in \mathbb{R}$	$\theta \in \mathbb{R}$	$\theta \in \mathbb{R}^{-*}$	$\theta \in \mathbb{R}^{-*}$
Dispersion ϕ	σ^2	1	1	ν^{-1}	$1/\sigma^2$
Support de ϕ	$\phi \in \mathbb{R}^{+*}$			$\phi \in \mathbb{R}^{+*}$	$\phi \in \mathbb{R}^{+*}$
Fonction $b(\theta)$	$\theta^2/2$	$\log(1+e^\theta)$	e^θ	$-\log(-\theta)$	$-(-2\theta)^{1/2}$
Fonction $c(y, \Phi)$	$-\frac{1}{2}\left(\frac{y^2}{\Phi} + \log(2\pi\phi)\right)$	$\log(C_n^{ny})$	$-\log(y!)$		$-\frac{1}{2}\left\{\log(2\pi\Phi y^3) + \frac{1}{\Phi y}\right\}$
$\mu(\theta) = \mathbb{E}[Y; \theta]$	θ	$e^\theta/(1+e^\theta)$	e^θ	$-1/\theta$	$(-2\theta)^{-1/2}$

Law	Distribution	θ	ϕ	$a(x)$	$b(x)$
$\mathcal{N}(\mu,\sigma^2)$	$\frac{1}{\sqrt{2\pi\sigma}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2	x	$\frac{x^2}{2}$
$\mathcal{G}(\alpha,\beta)$	$\frac{\beta^\alpha x^{\alpha-1}}{\Gamma(\alpha)}e^{-\beta x}$	$-\frac{\beta}{\alpha}=\frac{1}{\mu}$	$\frac{1}{\alpha}$		
$\mathcal{IN}(\mu,\lambda)$	$\sqrt{\frac{\lambda}{2\pi x^3}}e^{-\frac{\lambda(x-\mu)^2}{2\mu^2 x}}$	$-\frac{1}{2\mu^2}$	$\frac{1}{\lambda}$		
$\mathcal{B}(\mu)$	$\mu^x(1-\mu)^{1-x}$	$\log(\frac{\mu}{1-\mu})$	1		
$\mathcal{P}(\mu)$	$\frac{\mu^x}{x!}e^{-\mu}$	$\log(\mu)$	1	1	e^x
$\mathcal{OP}(\phi,\mu)$	$\frac{\mu^{\frac{x}{\phi}}}{\frac{x}{\phi}!}e^{-\mu}$	$\log(\mu)$	ϕ		

Law	$c(x,\theta)$	Expectation	Var. function	Support
$\mathcal{N}(\mu,\sigma^2)$	$-\frac{1}{2}(\frac{x^2}{\theta}+\log(2\pi\theta))$	$\mu=\theta$	1	$\mathbb{R}$
$\mathcal{G}(\alpha,\beta)$		$\mu=-\frac{1}{\theta}$	μ^2	$\mathbb{R}_+$
$\mathcal{IN}(\mu,\lambda)$		$\mu=(-2\theta)^{-\frac{1}{2}}$	μ^3	$\mathbb{R}_+$
$\mathcal{B}(\mu)$		$\mu=\frac{e^\theta}{1+e^\theta}$	$\mu(1-\mu)$	$\{0,1\}$
$\mathcal{P}(\mu)$	$-\log(x!)$	$\mu=e^\theta$	μ	$\mathbb{N}$
$\mathcal{OP}(\phi,\mu)$		$\phi\mu$	$\mu(1+\phi\mu)$	$\mathbb{N}$

Inférence

En annulant la dérivée de la log-vraisemblance $L\left(\theta, \phi; (x_{i,j})_{i+j\leq n}\right)$, on retombe sur le système (S) des équations de Wedderburn (où ϕ ne figure pas):

$$(S) \begin{cases} \sum_{i=1}^{n} \omega_i \dfrac{(y_i - \mu_i)}{V(\mu_i)} \dfrac{\partial \mu_i}{\partial \eta_i} b_i^{(k)} = 0, & k = 1, ..., p, \end{cases}$$

avec $b_i^{(k)}$ est la dérivée partielle de η_i par rapport au k^{eme} élément de la suite $(\beta_j)_{j=0,...,p}$.

On résoud ce système par l'algorithme de Newton-Raphson, ce qui nous donne l'E.M.V. $\hat{\xi} = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p)$ de $\xi = (\beta_0, \beta_1, ..., \beta_p)$.

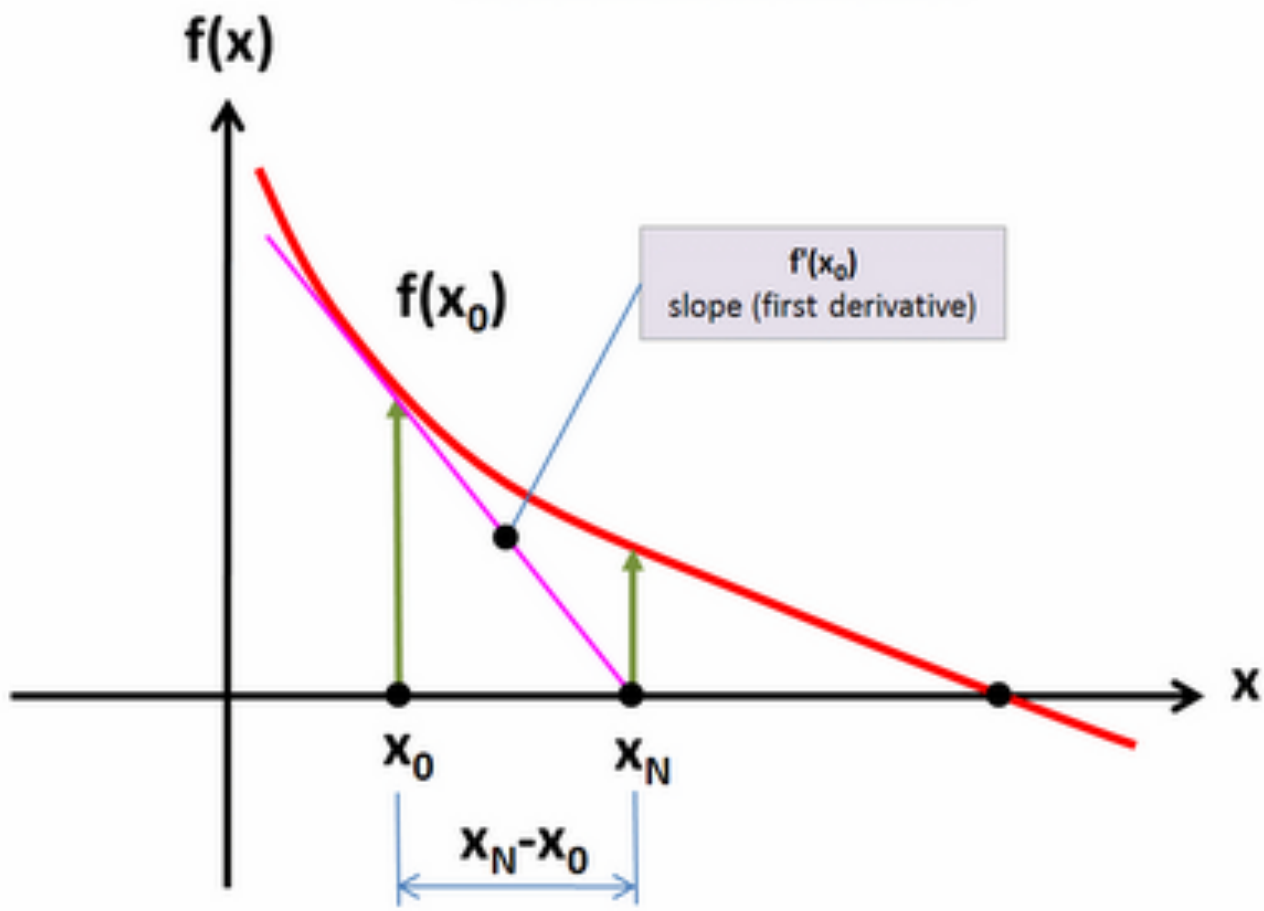
La matrice d'information de Fisher est obtenu via l'expression

$$I(\xi) = \frac{1}{\phi} M^T W M,$$

où M sont les régresseurs, et W diagonale d'éléments $w_i = \frac{\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2}{V(\mu_i)}$.

Donc le paramètre de dispersion ne joue aucun rôle ds l'estimation de $\hat{\xi}$, mais a une influence sur la dispersion de $\hat{\xi}$!

$\rightarrow$ Faire un exemple concret (ex: le modèle logistique).



Taylor-Lagrange avec

$$f = L^{'} \text{ et } f(x_{k+1}) = 0 \qquad \Rightarrow \qquad x_{k+1} = x_k - \frac{f(x_k)}{f^{'}(x_k)}$$

3 Les Modèles Linéaires Généralisés (GLM)

- Les GLM: brefs rappels
- Caractérisation et formalisation
- Validation
- Implémentation
- Lecture des résultats de la calibration
- Sélection de modèle et de variables

Résidus et déviance

En régression linéaire, on dispose du R^2 comme indicateur de la qualité de la modélisation.

Avec les GLM, les mesures de la qualité d'ajustement proviennent de la déviance, et du Chi-deux de Pearson.

Dû au fait que les observations ne sont pas supposées suivre une loi normale. Cependant, l'analyse des résidus reste indispensable:

- résidus de Pearson: pour $i + j \leq n$,

$$r_i^{(P)} = \frac{Y_i - \hat{\mu}_i}{\sqrt{\hat{Var}(Y_i)}} \simeq \frac{Y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}.$$

On déduit la statistique de Pearson: $\chi^2 = \sum_i \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$.

$\chi^{2*} = \chi^2/\phi$ est le χ^{2*} de Pearson standardisé de $Var(Y_i) = \phi V(\mu_i)$.

- la déviance: elle compare deux vraisemblances:
  - le modèle saturé $\mathbb{E}[Y_i] = Y_i$ (autant de paramètres que d'observations, donc erreur nulle);
  - et le modèle calibré $\mathbb{E}[Y_i] = \hat{\mu}_i$ avec $\hat{\mu}_i = g^{-1}(\eta_i)$:

$$-2 \ln \frac{L(\hat{\theta}, \phi; y)}{L(\tilde{\theta}, \phi; y)} = \frac{2}{\phi} \sum_i \left\{ y_i(\tilde{\theta}_i - \hat{\theta}_i) - [b(\tilde{\theta}_i) - b(\hat{\theta}_i)] \right\}$$

La déviance est alors définie par

$$D = 2 \sum_{i=1}^{n} \left\{ y_i(\tilde{\theta}_i - \hat{\theta}_i) - [b(\tilde{\theta}_i) - b(\hat{\theta}_i)] \right\},$$

Par analogie, la déviance standardisée : $D^* = D/\phi$.
Les résidus de déviance sont définis par

$$r_i^{(D)} = \text{signe}(Y_i - \hat{\mu}_i)\sqrt{d_i},$$

où $d_i = 2\left\{y_i(\tilde{\theta}_i - \hat{\theta}_i) - [b(\tilde{\theta}_i) - b(\hat{\theta}_i)]\right\}$.

On remarque ainsi que

$$\chi^2 = \sum_{i=1}^{n}\left[r_i^{(P)}\right]^2 \qquad D = \sum_{i=1}^{n}\left[r_i^{(D)}\right]^2$$

Modèle bien ajusté $\Leftrightarrow \chi^2$ et D prennent de faibles valeurs.

Ces résidus ne doivent faire apparaitre aucune structure non-aléatoire: on effectuera un Q-Q Plot de normalité.

Mise en oeuvre en R

Nous travaillons sur le jeu de données esoph de la librairie datasets:

```
> library(datasets) ; data(esoph)
> dim(esoph)
[1] 88  5
> head(esoph, n=4)
  agegp     alcgp    tobgp ncases ncontrols
1 25-34 0-39g/day 0-9g/day      0        40
2 25-34 0-39g/day    10-19      0        10
3 25-34 0-39g/day    20-29      0         6
4 25-34 0-39g/day      30+      0         5
```

On a le nb de personnes ayant un cancer de l'oesophage pour une tranche d'âge donnée, consommation d'alcool et de tabac, ainsi que l'exposition.

Pour les variables explicatives catégorielles, chaque modalité est codée en R comme une indicatrice pour le calcul des coef. de régression. La matrice des régresseurs est appelée matrice de schéma (ou design).

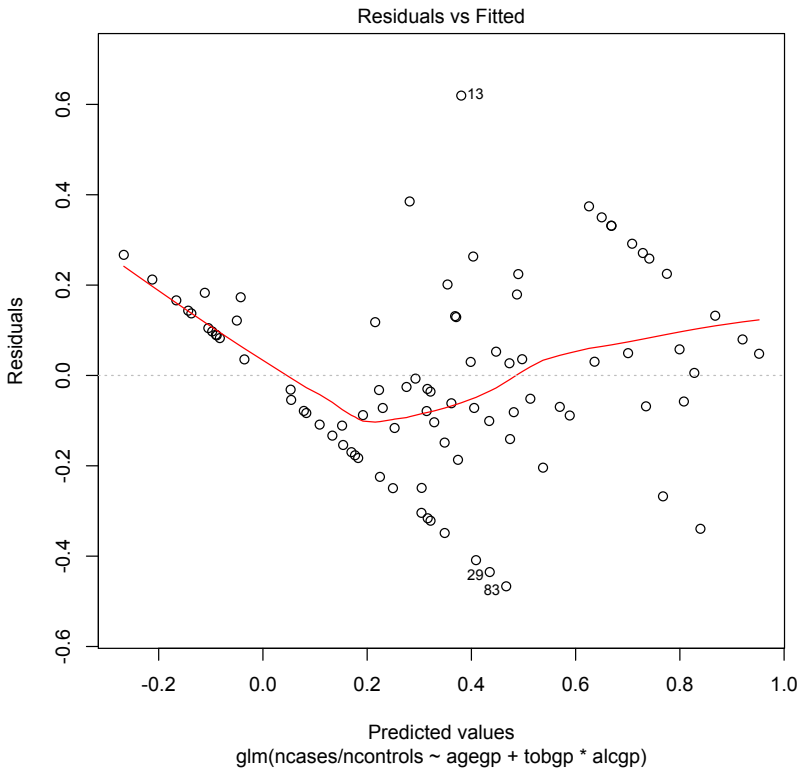
Q: prédire le taux d'atteinte en fonction des facteurs de risque $\rightarrow$ lien logit

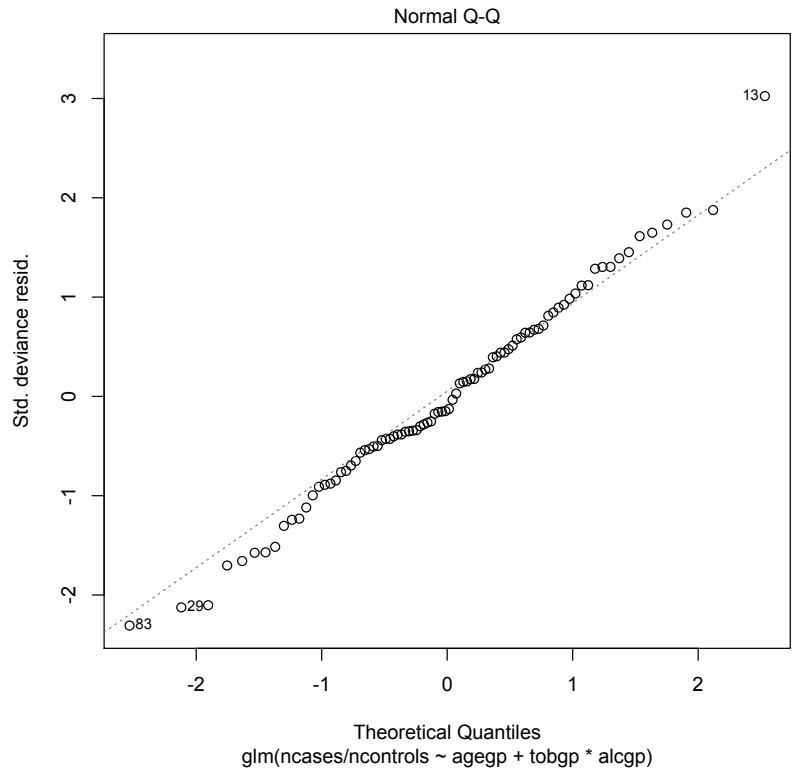
Que donne la régression linéaire classique?

```
>  esoph.lm <- glm(ncases/ncontrols ~ agegp + tobgp * alcgp, family=gaussian, da
>  summary(esoph.lm)

Call:
glm(formula = ncases/ncontrols ~ agegp + tobgp * alcgp, family = gaussian,
    data = esoph)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.46676  -0.11238  -0.02766   0.13151   0.61950
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.3683021  0.0253322  14.539  < 2e-16 ***
agegp.L         0.4982970  0.0643470   7.744 6.97e-11 ***
agegp.Q        -0.0667736  0.0636627  -1.049   0.2980
agegp.C        -0.0273635  0.0619999  -0.441   0.6604
agegp^4         0.0961215  0.0602277   1.596   0.1152
agegp^5        -0.0291224  0.0586807  -0.496   0.6213
tobgp.L         0.1026511  0.0504525   2.035   0.0459 *
...
tobgp.C:alcgp.C  0.0001316  0.1006860   0.001   0.9990
```

Un oeil sur les résidus

```
> plot(esoph.lm)
```





Comparaison des prévisions

On peut comparer les prévisions associées aux deux modèles:

- modèle linéaire avec effets additifs des facteurs de risque;
- GLM et effets multiplicatifs des variables sur les OR de Y:

```
>  ## comparaison entre prevision par modele lineaire et glm
>  esoph.logit <- glm(cbind(ncases,ncontrols) ~ agegp + tobgp * alcgp, family=bi
>  cbind(obs=esoph$ncases/esoph$ncontrols, LM=fitted(esoph.lm), GLM=fitted(esoph
         obs        LM        GLM
32 0.0000000 0.1539088 0.09441146
33 0.0000000 0.1767439 0.09993717
34 0.0000000 0.1827688 0.16478482
35 0.1578947 0.2299179 0.15056445
36 0.1904762 0.2227619 0.17211332
37 0.3333333 0.2153707 0.18260428
38 0.7142857 0.4899419 0.30463314
39 0.1875000 0.3742752 0.20184238
40 0.4285714 0.3986418 0.23163685
```

Coefficients de régression

La qualité d'ajustement et le sens de l'impact des facteurs de risque est donnée par la fonction summary().

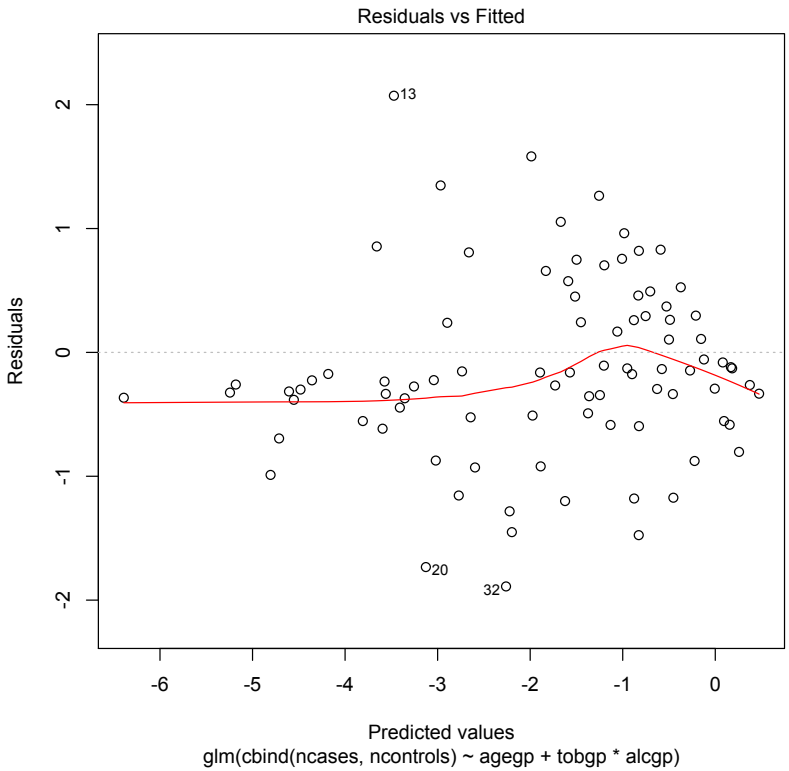
```
> summary(esoph.logit) (...)
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -1.75985    0.19822  -8.878  < 2e-16 ***
agegp.L          2.99646    0.65386   4.583 4.59e-06 ***
agegp.Q         -1.35008    0.59197  -2.281   0.0226 *
agegp.C          0.13436    0.45056   0.298   0.7655
...
agegp^5         -0.21347    0.19627  -1.088   0.2768
tobgp.L          0.63846    0.19710   3.239   0.0012 **
...
tobgp.Q:alcgp.C  0.04843    0.36211   0.134   0.8936
tobgp.C:alcgp.C -0.13905    0.35754  -0.389   0.6973
---
Signif. codes:  0 Ô***Õ 0.001 Ô**Õ 0.01 Ô*Õ 0.05 Ô.Õ 0.1 Ô Õ 1
```

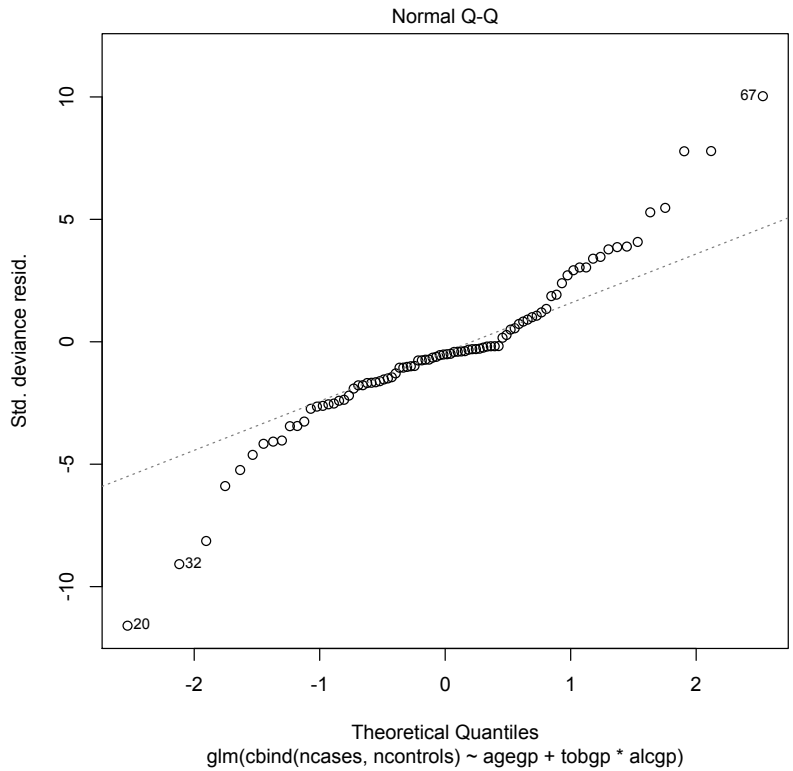
Résidus de la modélisation

Résidus (Pearson/déviance) ne doivent pas dégager de tendance.

Ces résidus ne sont pas forcément gaussiens...(cf Q-Q plot de normalité slide suivante).

```
> plot(esoph.logit, which = 1:2)
```


Residuals vs Fitted
Residuals
Predicted values
glm(cbind(ncases, ncontrols) ~ agegp + tobgp * alcgp)
13
20
32


Normal Q-Q
Std. deviance resid.
Theoretical Quantiles
glm(cbind(ncases, ncontrols) ~ agegp + tobgp * alcgp)
67
32
20

Déviance

Elle mesure la qualité d'adéquation du modèle (en comparant la vraisemblance du modèle courant à celle du modèle saturé, du modèle nul, d'un modèle emboîté).

Les résidus de déviance doivent être aussi petits que possible.

```
    Null deviance: 227.241  on 87  degrees of freedom
Residual deviance:  47.484  on 67  degrees of freedom
AIC: 236.96
Number of Fisher Scoring iterations: 6

## p-valeur du test de significativite:
> 1-pchisq(residual.deviance,df)
...
```

Rq: la déviance suit $\approx$ le Khi-deux (McCullagh and Nelder (1989)).

3 Les Modèles Linéaires Généralisés (GLM)

- Les GLM: brefs rappels
- Caractérisation et formalisation
- Validation
- Implémentation
- Lecture des résultats de la calibration
- Sélection de modèle et de variables

Sélection de modèle par analyse de variance

- Comparaison du modèle courant au modèle nul:

```
>  anova(esoph.logit)  # comparaison modele courant au modele nul
Analysis of Deviance Table
Terms added sequentially (first to last)

            Df Deviance Resid. Df Resid. Dev
NULL                           87    227.241
agegp        5   88.128         82    139.112
tobgp        3   19.085         79    120.028
alcgp        3   66.054         76     53.973
tobgp:alcgp  9    6.489         67     47.484
```

- Comparaison entre deux modèles GLM:

```
>  esoph.logit2 <- glm(cbind(ncases,ncontrols) ~ agegp + tobgp + alcgp, fam
>  anova(esoph.logit2, esoph.logit1, test="Chisq")
Analysis of Deviance Table

Model 1: cbind(ncases, ncontrols) ~ agegp + tobgp + alcgp
Model 2: cbind(ncases, ncontrols) ~ agegp + tobgp * alcgp
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        76     53.973
2        67     47.484  9   6.4895   0.6901
```

Sélection de variable

→ Tests d'hypothèses pour connaitre la pertinence des variables (test de Wald basé sur les prop. du MLE).
→ Ici approche descendante (modèle saturé et suppression): fonction stepAIC() du package MASS (AIC à minimiser).

```
>  esoph.backward <- stepAIC(esoph.logit, direction="backward")
Start:  AIC=236.96 % AIC du modele calibre et stocke dans l'objet
cbind(ncases, ncontrols) ~ agegp + tobgp * alcgp
              Df Deviance    AIC
- tobgp:alcgp  9   53.973 225.45
<none>             47.484 236.96 % supprime rien, garde le modele actuel
- agegp        5  123.950 303.43

Step:  AIC=225.45
cbind(ncases, ncontrols) ~ agegp + tobgp + alcgp
        Df Deviance    AIC
<none>       53.973 225.45  % fin de la procedure puisque AIC est minimise ici
- tobgp  3   64.572 230.05
- alcgp  3  120.028 285.51
- agegp  5  131.484 292.96
```

]

Choix de la fonction de lien

Suivant la distribution choisie pour la variable réponse, on peut considérer différentes fonctions de lien. Citons les:

- distribution normale:
  - liens identité, log et inverse;
- distribution gamma:
  - liens inverse, log, identité;
- distribution inverse gaussienne:
  - liens inverse carré, inverse, log, identité;
- distribution binomiale:
  - liens logit, probit, cauchit, cloglog;
- distribution poisson:
  - liens log, identité, racine carré, inverse.

Remarque: les premiers liens cités sont les liens canoniques.

Choix de la fonction de lien (suite)

- La détection d'une tendance systématique des résidus indique probablement un mauvais choix de lien,
- Suivant la distribution de l'erreur, il y a un choix limité de fonctions de lien possibles.

Ex: pour une erreur de loi de Poisson, nous pouvons considérer comme liens: *identite*, *sqrt*, *inverse* et *log*.

```
esoph.logit <- glm(cbind(ncases,ncontrols) ~ agegp + tobgp * alcgp,
                   family=binomial(link="logit"), data=esoph)
esoph.logit <- glm(cbind(ncases,ncontrols) ~ agegp + tobgp * alcgp,
                   family=binomial(link="probit"), data=esoph)
esoph.logit <- glm(cbind(ncases,ncontrols) ~ agegp + tobgp * alcgp,
                   family=binomial(link="cauchit"), data=esoph)
esoph.logit <- glm(cbind(ncases,ncontrols) ~ agegp + tobgp * alcgp,
                   family=binomial(link="cloglog"), data=esoph)
```

4 Usage pratique des GLM: les écueils récurrents

- Quelques notions opérationnelles importantes sur les GLM
- Surdispersion et masse en 0
- Segmentation et modélisation: limites à garder en tête
- Tenir compte de l'exposition: l'offset
- Réponse catégorielle: sur-représentation d'une modalité

4 Usage pratique des GLM: les écueils récurrents

- Quelques notions opérationnelles importantes sur les GLM
- Surdispersion et masse en 0
- Segmentation et modélisation: limites à garder en tête
- Tenir compte de l'exposition: l'offset
- Réponse catégorielle: sur-représentation d'une modalité

Choix de la loi de l'erreur et fonctions de lien en actuariat

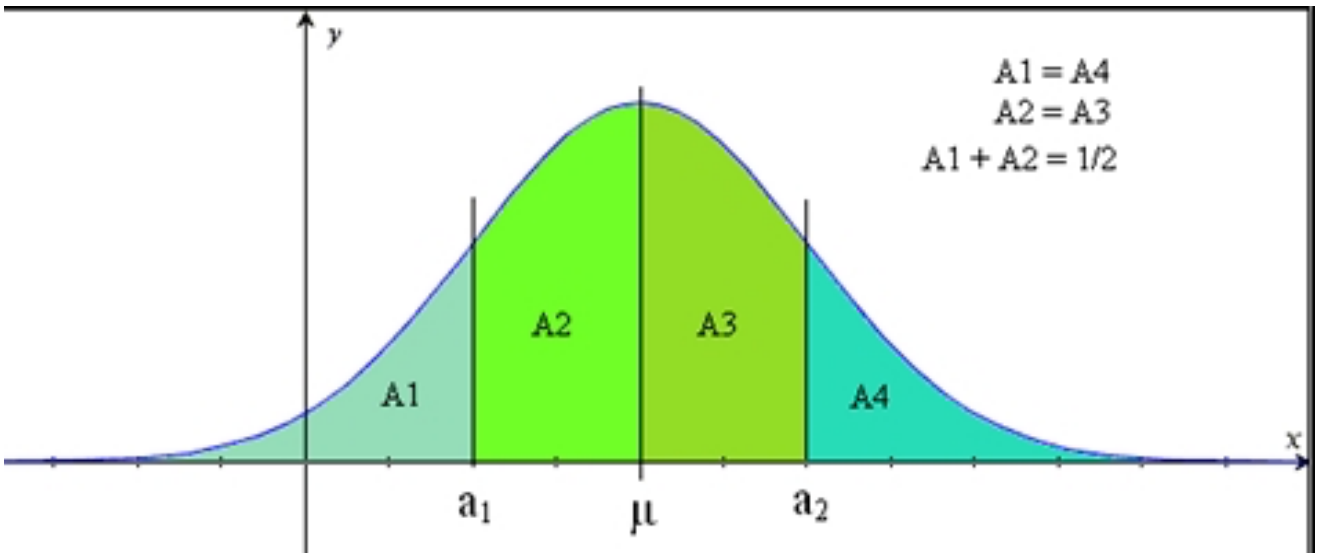
Adapter le lien en fonction du domaine de définition de Y.

Loi	Lien naturel	Moyenne	Utilisation
$\mathcal{N}(\mu, \sigma^2)$	Id: $\eta = \mu$	$\mu = X\beta$	Rég. lin.
$\mathcal{B}(\mu)$	logit: $\eta = ln(\frac{\mu}{1-\mu})$	$\mu = \frac{exp(X\beta)}{1+exp(X\beta)}$	Taux
$\mathcal{P}(\mu)$	log: $\eta = ln(\mu)$	$\mu = exp(X\beta)$	Fréquence
$\mathcal{G}(\alpha, \beta)$	inverse: $\eta = \frac{1}{\mu}$	$\mu = (X\beta)^{-1}$	Sévérité
$\mathcal{IN}(\mu, \lambda)$	inverse^2: $\eta = -\frac{1}{\mu^2}$	$\mu = (X\beta)^{-2}$	Sévérité

La gaussienne

L'utilisation d'une loi Normale est encore très répandue... Mais cela implique des erreurs fondamentales de raisonnement, notamment

- la densité de la loi est symétrique,
- sa queue de distribution est **fine**,
- support non adapté à des charges sinistres $\Rightarrow \mathbb{P}(Y < 0)$.



Valeur des coefficients calibrés: impact sur la réponse

En général, on interprète les résultats de la manière suivante:

- $\beta_j > 0$: $\nearrow$ du facteur de risque X_j provoque $\nearrow$ de $g(\mathbb{E}[Y])$;
- $\beta_j < 0$: $\nearrow$ du facteur de risque X_j provoque $\searrow$ de $g(\mathbb{E}[Y])$
- $\beta_j = 0$: effet nul de la variation dudit X_j.

Evidemment, cela **dépend aussi du type de modélisation**!

- Pour des modèles à effets additifs, la valeur de réf. sera 0;
- Pour des modèles multiplicatifs, la valeur de référence sera 1 (à une transformation près parfois, cf modèle log-Poisson).

Pour connaitre le type d'effet, on réécrit le modèle sous la forme

$$\mathbb{E}[Y \,|\, \mathbf{X}] = g^{-1}(\mathbf{X}^T \beta).$$

Comparateur en ligne et odd-ratio (OR)

En souscrivant en ligne, vous pouvez par ex. avoir une idée de la calibration de certains assureurs pour certains facteurs de risque: comparer le tarif en faisant évoluer **1 seule** caractéristique (ex: âge, ancienneté du permis, couleur de la voiture, ...)

Cela correspond à l'**odd-ratio**, un rapport sur la quantité d'intérêt:

$$\frac{\mathbb{E}[Y \mid X_j = x_j + 1]}{\mathbb{E}[Y \mid X_j = x_j]} = h(\beta_j),$$

avec h une fonction à déterminer.

Exemple log-poisson: $Y \sim \mathcal{P}(\lambda)$, donc $\lambda = e^{\mathbf{x}^T\beta} \Rightarrow h(\beta_j) = e^{\beta_j}$.

Validation d'un modèle GLM

Il faut garder en tête que la validation d'une modélisation de type GLM passe par plusieurs étapes:

1. construction de 2 échantillons $\perp\!\!\!\perp$ par tirage aléatoire: un d'apprentissage (construction) et un de validation;
2. validation de la significat. globale du modèle (déviance, LRT);
3. validation de la significativité des coef. de régression un à un;
4. allure des résidus (doit être aléatoire);
5. confrontation "modélisé / empirique" sur l'échantillon de validation par prévisions données par le modèle.

Transformations au sein du prédicteur

Il peut être utile d'**introduire une transfo. dans le prédicteur** sur certaines covariables en fonction du type d'impact sur Y.

Cette transformation sera choisie en fonction de l'effet du facteur de risque sur Y lors de la visualisation des statistiques desc.

Prenons un ex. concret: supposons que l'âge x a un impact exponentiel sur le taux de mortalité q_x, mais que la CSP joue de manière linéaire. Ainsi on posera un modèle de la forme

$$ln(q_x) = a + b\,x + ln(c\,CSP) \;\Leftrightarrow\; q_x = A \times \exp(bx) \times c\,CSP$$

Tweedie or not Tweedie?

Boucher and Danail (2011)

La densité est donnée par

$$f(y;\mu,\phi) = a(y,\phi)\ \exp\left(\frac{1}{\phi}\Big[y\theta(\mu) - \kappa(\theta(\mu))\Big]\right),$$

$$\theta(\mu) = \begin{cases} \frac{\mu^{1-p}}{1-p} & \text{si } p \neq 1 \\ \log\mu & \text{si } p = 1 \end{cases} \qquad \kappa(\theta(\mu)) = \begin{cases} \frac{\mu^{2-p}}{2-p} & \text{si } p \neq 2 \\ \log\mu & \text{si } p = 2 \end{cases}$$

Dans cette formalisation, $\mathbb{E}[Y] = \mu$ et $Var(Y) = \psi\mu^p = \psi\mathbb{E}[Y]^p$, avec ψ un parametre de dispersion > 0.

L'ordre $p \in \mathbb{R}^+$ (*paramètre d'indice*), choisi (en fonction de l'application) avant d'estimer μ et ϕ, définit le **type de distribution**:

- $\rightarrow$ $p < 0$: réalisations dans $\mathbb{R}$; $p = 0$: loi gaussienne,
- $\rightarrow$ $0 < p < 1$: pas de distribution (pas de modèle Tweedie),
- $\rightarrow$ $p = 1$ avec $\phi = 1$: loi de Poisson,
- $\rightarrow$ $1 < p < 2$: loi composée Poisson-Gamma (réalisations ≥ 0),
- $\rightarrow$ $2 < p < 3$ ou $p > 3$: positive stable distributions ($x > 0$),
- $\rightarrow$ $p = 2$: loi Gamma, $p = 3$: loi inverse gaussienne.

En pratique, $1 < p < 2$ pour modéliser fréq. et coût en mm tps!
Inconvénient: mêmes var. explicatives prises en compte dans les lois de fréq. et de coût, or les praticiens savent qu'elles sont $\neq$.

4 Usage pratique des GLM: les écueils récurrents

- Quelques notions opérationnelles importantes sur les GLM
- Surdispersion et masse en 0
- Segmentation et modélisation: limites à garder en tête
- Tenir compte de l'exposition: l'offset
- Réponse catégorielle: sur-représentation d'une modalité

Pratique courante

Dans les compagnies d'assurance, on penche souvent pour la loi de Poisson dans la modélisation de la fréquence des sinistres lorsqu'on adopte une modélisation de type fréquence-coût.

En pratique,

- cela simplifie le calcul global de sinistralité à l'échelle du portefeuille: loi **Poisson composée stable par addition**;
- souvent on observe que la **variance empirique du nombre de sinistres est bien supérieure à sa moyenne empirique**: cela va à l'encontre de la propriété fondamentale de cette loi.

On réalise donc que cette modélisation n'est pas adaptée!

Cas classique de surdispersion: la Binomiale Négative

Elle peut être construite comme un mélange de lois de Poisson:

$$(N \mid \Lambda = \lambda) \sim \mathcal{P}(\lambda) \qquad \text{et} \qquad \Lambda \sim \mathcal{G}a(\alpha, \delta).$$

La densité jointe de N et Λ vaut

$$f_{N,\Lambda}(n,\lambda) = f_{N\mid\Lambda=\lambda}(n)\, f_\Lambda(\lambda) = e^{-\lambda}\frac{\lambda^n}{n!}\,\frac{\delta^\alpha \lambda^{\alpha-1} e^{-\delta\lambda}}{\Gamma(\alpha)} \quad (\lambda, \alpha, \delta > 0, n \in \mathbb{N}).$$

Λ est continue et N discrète: la distribution marginale de N est

$$\begin{aligned}
\mathbb{P}(N=n) &= \int_0^\infty f_{N,\Lambda}(n,\lambda)\,d\lambda = \int_0^\infty e^{-\lambda}\frac{\lambda^n}{n!}\,\frac{\delta^\alpha\lambda^{\alpha-1}e^{-\delta\lambda}}{\Gamma(\alpha)}\,d\lambda \\
&= \frac{\delta^\alpha}{n!\,\Gamma(\alpha)}\int_0^\infty \lambda^{n+\alpha-1}e^{-(\delta+1)\lambda}\,d\lambda = \frac{\delta^\alpha\Gamma(\alpha+n)}{n!\,\Gamma(\alpha)\,(\delta+1)^{\alpha+n}}
\end{aligned}$$

Posons ensuite $p = \frac{\delta}{\delta+1}$, et $q = 1 - p = \frac{1}{\delta+1}$. Alors

$$\boxed{\mathbb{P}(N = n) = \frac{\Gamma(\alpha + n)}{n!\,\Gamma(\alpha)} p^{\alpha} q^{n}.}$$

La v.a. $N \sim \mathcal{NB}(\alpha; p)$ prend ses valeurs dans $\{0, 1, 2, ...\}$.

Remarques:

- La queue de distribution est plus épaisse que celle d'une loi de Poisson.
- Sa variance est plus grande qu'une loi de Poisson: loi utilisée en cas de surdispersion des observations.

Les modèles de comptage Zero-Inflated (ZI)

Frees (2009)

Utilisé lorsque la survenance des sinistres est **rare**...
Les "0" observés viennent de loi de comptage + masse en 0 (π_0):

- une composante regroupe les deux "sources" de 0,
- l'autre regroupe les obs. $\neq 0$ provenant de la loi de comptage.

$$\boxed{\mathbb{P}(N=k) \quad = \quad \pi_0 \, \mathbb{1}_{\{k=0\}} + (1-\pi_0) \, f_{\text{count}}(k).}$$

Ex: $N \sim ZI\mathcal{P}(\lambda)$: $\mathbb{P}(N=k) = \begin{cases} \pi_0 + (1-\pi_0)\, e^{-\lambda} & \text{si } k=0, \\ (1-\pi_0)\, e^{-\lambda} \, \dfrac{\lambda^k}{k!} & \text{si } k>0. \end{cases}$

Les modèles de type “hurdle-at-zero”

Frees (2009)

Idem que précédemment pour l’utilisation, sauf que l’on **maitrise mieux la proportion de 0** ici (– d’aléa sur cette quantité):

- masse en 0 (ne proviennent plus du tout de la loi comptage),
- à laquelle on ajoute une loi de comptage tronquée.

$$\mathbb{P}(N=k) = \begin{cases} \pi_0 & \text{si } k=0, \\ (1-\pi_0)\,\dfrac{f_{\text{count}}(k)}{1-f_{\text{count}}(0)} & \text{si } k>0. \end{cases}$$

Zero-trunc. $\mathcal{P}$: $$\mathbb{P}(N=k) = \begin{cases} \pi_0 & \text{si } k=0, \\ (1-\pi_0)\,\dfrac{e^{-\lambda}\lambda^k}{(1-e^{-\lambda})k!} & \text{si } k>0. \end{cases}$$

4 Usage pratique des GLM: les écueils récurrents

- Quelques notions opérationnelles importantes sur les GLM
- Surdispersion et masse en 0
- Segmentation et modélisation: limites à garder en tête
- Tenir compte de l'exposition: l'offset
- Réponse catégorielle: sur-représentation d'une modalité

Création de poches d'assurés

La segmentation amène à créer des poches d'assurés ayant les mêmes caractéristiques. Il y a un arbitrage naturel entre

- une segmentation "**grossière**": peu de poches différentes, donc peu de tarifs $\neq$;
- une segmentation **précise**: beaucoup de profils de risque considérés $\neq$, des tarifs très personnalisés (cf pb Big Data).

Une question essentielle liée à cette problématique de segmentation est l'exposition... moindre dans certaines poches!
$\rightarrow$ Remise en cause du principe de mutualisation (LFGN)...
$\rightarrow$ Attention pour les GLM (MLE asymptotique), voire même pour le calcul de la sinistralité globale en espérance par agrégation...

Difficultés de calibration des coefficients

Il arrive souvent en pratique que des coefficients de régression calibrés ne soient **pas significatifs**. Cela correspond au test:

$$H_0 : \hat{\beta}_j = 0 \qquad VS \qquad H_1 : \hat{\beta}_j \neq 0.$$

But: rejeter H_0 à un certain niveau de confiance α, en se basant sur la statistique de Wald $(\hat{\beta}_j / Var(\hat{\beta}_j))^2 \quad (\sim \chi^2)$.

Lorsque l'exposition est faible dans une poche, la calibration des coefficients de régression affectés à cette poche devient ardue...

Cela est dû au fait que le MLE est **asymptotiquement gaussien**:

$$\hat{\beta}_j^{MLE} \quad \sim \quad \mathcal{N}(\beta_j,\ 1/\mathcal{I}(\beta_j)).$$

$\Rightarrow$ La variance de l'estimateur peut devenir grande si l'information de Fisher est faible (quantité d'info contenue dans les données, petite dans le cas de trop peu d'individus).

La technique consiste alors à **regrouper certaines modalités de covariables** qualitatives (ou catégorielles). La démarche statistique "propre" s'y rapportant:

1. calibration du modèle saturé (ou modèle complet),
2. pour le test de chaque coef. associé aux covariables, repérer la pire "p-valeur" au-dessus du seuil α,
3. agréger la modalité correspondante avec une autre "intelligemment";
4. recalibrer le modèle, et revenir à l'étape 2 tant que le modèle n'est pas satisfaisant.

Dimension du modèle à “minimiser”

On a tjs 2 effets inverses en modélisation (cf théorie de Vapnik):

- adéquation du modèle: plus la dimension du modèle est grande, plus l'adéquation aux données est bonne;
- qualité prédictive: plus la dimension du modèle est grande, plus la capacité prédictive du modèle est mauvaise (on capte les bruits au lieu de capter le signal principal).

L'idée est donc de rechercher un arbitrage dans la dimension qui permette d'obtenir un bon compromis dans ces 2 objectifs.

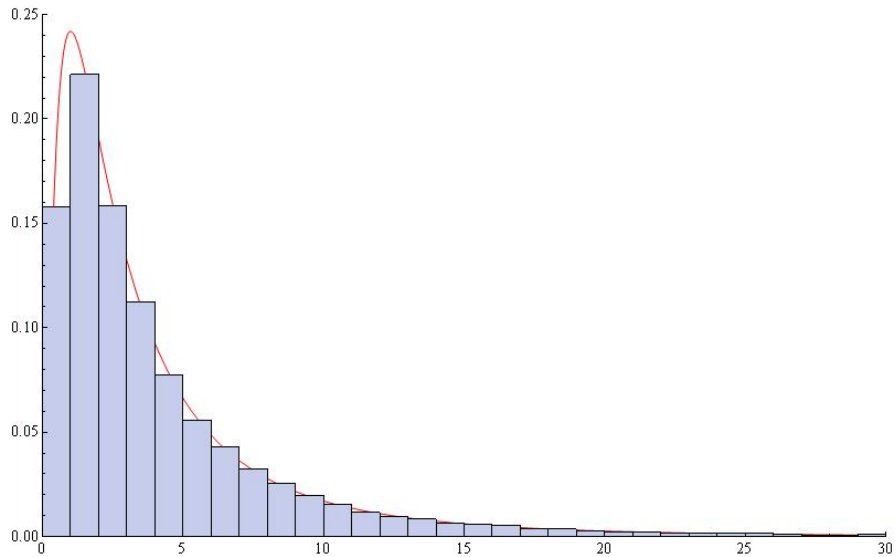
C'est ce qu'on appelle un modèle parcimonieux.

Critères de sélection de modèles emboîtés: AIC, BIC, ...

Distribution de sinistralité par poche

Au final, une question importante est d'**identifier les poches pour lesquelles la modélisation marche bien ou non**: il vaut mieux se tromper sur certains profils que sur d'autres...

Pour cela, on confronte la densité théo. construite par GLM à la densité empirique du profil (poche): dans l'idéal ça coincide presque!



4 Usage pratique des GLM: les écueils récurrents

- Quelques notions opérationnelles importantes sur les GLM
- Surdispersion et masse en 0
- Segmentation et modélisation: limites à garder en tête
- Tenir compte de l'exposition: l'offset
- Réponse catégorielle: sur-représentation d'une modalité

Qu'est-ce que l'offset?

L'offset représente une sorte d'**exposition**.

C'est une constante qui va venir modifier le risque de base, donc le risque qui n'est pas lié au profil de l'assuré en particulier.

Exemples d'offset:

- assurance auto indiv.: nb d'années d'assurance du véhicule;
- assurance collective auto: taille de la flotte assurée;
- assurance collective incapacité-invalidité: effectif de salariés, masse salariale;
- réassurance: taille du portefeuille, ...

Comment intégrer un offset dans un modèle GLM?

Tout simplement! C'est un terme commun à tous les individus, mais dont la valeur va changer en fonction des individus.

En terme explicite, l'équation devient

$$g(\mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}]) = \mathit{offset} + \mathbf{x}^T\beta.$$

- on **contraint le coefficient de l'offset à valoir 1** (c'est pourquoi il n'apparait pas dans l'équation!);
- pour la calibration, on régresse $g(\mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}]) - \mathit{offset} = \mathbf{x}^T\beta$.

Exemple d'offset dans le modèle log-Poisson

L'idée globale de l'offset est que la réponse **y est proportionnelle**.

Donc l'offset s'exprime sur la même échelle que la réponse. Dans le cas du modèle log-Poisson de paramètre λ, on aurait donc

$$ln(\mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}]) = ln(exposition) + \mathbf{x}^T\beta.$$

Soit le modèle suivant à calibrer: $ln\left(\frac{\mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}]}{exposition}\right) = \mathbf{x}^T\beta.$

On remplace donc la fréquence (au sens nb de sinistres) par une fréquence standardisée!

4 Usage pratique des GLM: les écueils récurrents

- Quelques notions opérationnelles importantes sur les GLM
- Surdispersion et masse en 0
- Segmentation et modélisation: limites à garder en tête
- Tenir compte de l'exposition: l'offset
- Réponse catégorielle: sur-représentation d'une modalité

Etude d'un taux de réponse faible

On cherche parfois à modéliser un **événement binaire "rare"** en utilisant des modèles GLM.

Quel(s) problème(s) cela pose?

Difficultés énoncées précédemment sur la calibration notamment
$\rightarrow$ +sieurs poches où on observe (très) peu ou pas l'événement...

Exemples concrets (souvent en risque comportemental):

- taux de résiliation en assurance vie et non-vie (surtout en vie où les taux de résiliation annuels sont + faibles);
- taux de conversion en assurance directe par exemple.

Formalisation du contexte

Plaçons nous dans le cadre de risque comportemental pour présenter le concept (ex: taux de conversion). Cela nous amène à considérer un modèle GLM de type logistique, à savoir

$$ln\left(\frac{p_i}{1-p_i}\right) = \mathbf{x}_i^T \beta.$$

Rappelons que

- $\mathbf{X}_i^T = (1, X_{i1}, ..., X_{iJ})$ et $\beta^T = (\beta_0, \beta_1, ..., \beta_J)$;
- $i \in 1, ..., I$: $Y_i \in \{0, 1\}$ $\Rightarrow$ $Y_i \sim \mathcal{B}(p_i)$;
- $p_i = \mathbb{P}(Y_i = 1)$.

En pratique, $\bar{p} = \frac{1}{I} \sum_i \mathbb{1}_{y_i=1}$ est **de l'ordre de quelques % au** $+$.

Les deux problèmes théoriques associés

Albert and Anderson (1984)

1. **La séparabilité**: en fait, l'existence d'un estimateur du maximum de vraisemblance est conditionné par le problème de séparation. Il n'∄ de MLE en cas de séparation complète.

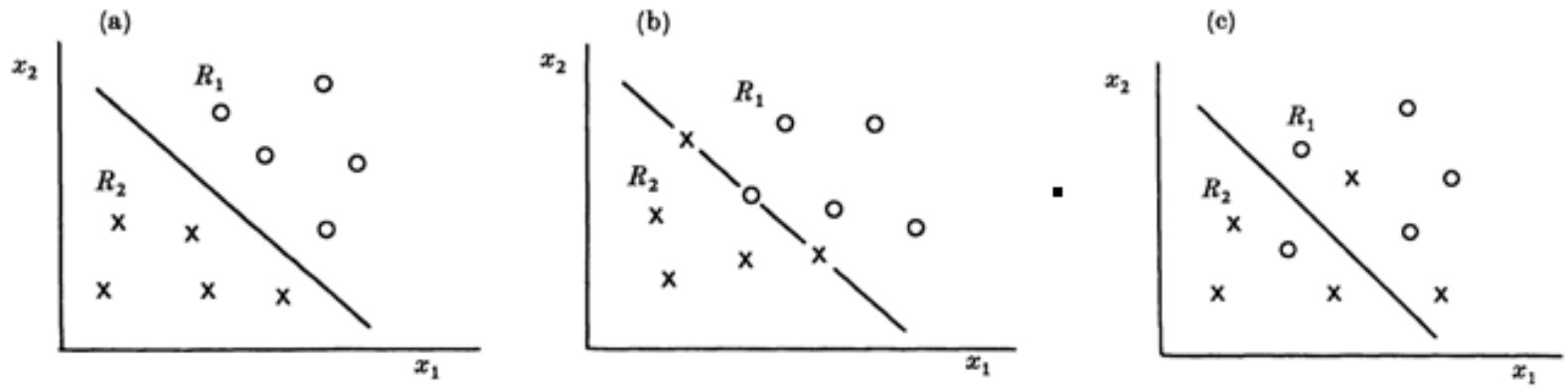


Figure 2 Possible configuration of sample points in the case of two variables, x_1 and x_2, and two groups, E_1, shown by circles, and E_2, shown by crosses. Regions R_1 and R_2 define corresponding allocation rule. (a) Complete separation. (b) Quasi-complete separation. (c) Overlap.

2. **La dimensionnalité** (“curse of dimensionality”).
   On dispose souvent de bc de covariables: la dim. de l’espace ↗ vite et les données peuvent rapidement devenir “sparse”.

Pour toute procédure statistique, la sparsité est un problème important. On entend parfois parler de

“Small N large P”

Pour avoir un résultat fiable dans la plupart des modèles statistiques, la taille des données dont nous avons besoin croit souvent exponentiellement en fonction de la dimension du modèle.

Remarque: dans le cadre de données “sparse”, on utilise plutôt la régression ridge, lasso, elastic net...

Solutions “théoriques” existantes

Pour éviter le problème de sparsité ou de non-existence du MLE pour des données qui seraient séparées (ou quasi-séparées), il existe deux principales méthodes:

- la vraisemblance pénalisée (penalized likelihood method);
- la régression logistique conditionnelle exacte (exact conditional logistic regression).

Rq: la 3^e alternative est le **response-based sampling**, artifice pour retomber sur un problème plus facile à traiter mais qui n’est pas applicable directement sur le problème d’origine (cf $+$ loin).

5 Application sur une base de données réelle

Exercice d'application en R

Importer la base de données Excel des sinistres auto.
Puis répondez aux questions suivantes:

1. extraire des statistiques descriptives de ce jeu de données (taux de sinistralité, nb de modalités par facteur de risque, corrélation, ...)
2. regrouper les modalités qu'il vous semble bon de regrouper;
3. construire aléatoirement un échantillon d'apprentissage et un échantillon de validation (de tailles respectives 2/3 et 1/3);
4. calibrer un modèle de régression linéaire multiple pour expliquer la charge sinistre (sans faire de distinction fréquence-coût):
   - par procédure de sélection de modèle backward,
   - calculer les prévisions de charges sinistres.

5. calibrer un modèle de régression log-Poisson pour expliquer la charge sinistre (sans faire de distinction fréquence-coût):
   - calculer les prévisions de charges sinistres,
   - comparer les statistiques descriptives des prévisions de la réponse par rapport à l'expérience et à la modélisation linéaire.
6. Construire un modèle de tarification fréquence-coût:
   - calibrer la loi de fréquence, et établir des prévisions;
   - calibrer la loi de sévérité, et établir des prévisions;
   - agréger les résultats pour obtenir un tarif;
   - comparer les résultats avec les modèles précédemment construits.
7. Refaire la question précédente en gérant le problème de surdispersion des données.

CONCLUSION

Il existe de nombreux écueils pratiques à la mise en place opérationnelle des modèles GLM en assurance.

Principalement:

- la segmentation et ce qu'elle induit (attention à ne pas trop segmenter!);
- le choix des lois et du lien;
- la calibration des modèles (convergence du MLE, bornitude de la vraisemblance, initialisation de l'algorithme de Newton-Raphson, etc...);
- la validation d'un modèle;
- la gestion de la surdispersion des données;
- la potentielle (très) faible sinistralité...

Bibliographie

Albert, A. and Anderson, J. A. (1984). On the Existence of Maximum Likelihood Estimates in Logistic Regression Models. *Biometrika*, 71(1):1–10.

Boucher, J. P. and Danail, D. (2011). On the Importance of Dispersion Modeling for Claims Reserving: An Application with the Tweedie Distribution. *Variance*, 5(2):158–172.

Brass, W. (1964). Uses of census and survey data for the estimation of vital rates. In *African Semin. Vital Stat.*, United Nations document E/ CN .14/CAS .4IVS/7.

Brass, W. and Macrae, S. (1984). Childhood mortality estimated from reports on previous births given by mothers at the time of a maternity: I. Preceding-births technique. In *Asian and Pacific Census Forum*, volume 11.

Frees, E. W. (2009). *Regression Modeling with Actuarial and Financial Applications*. International Series on Actuarial Science. Cambridge University Press, New York.

Lee, R. D. and Carter, L. R. (1992). Modeling and forecasting U.S. mortality. *J. Am. Stat. Assoc.*, 87(419):659–671.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models, 2nd ed.* Monographs on Statistics and Applied Probability. Chapman and Hall, London.