Chapitre 9

Tarification a priori

9.1 Introduction

Dans la tarification dite *a priori*, l'idée est de séparer les contrats (et les assurés) en plusieurs catégories, de façon qu'à l'intérieur d'une catégorie, les risques puissent être considérés comme "équivalents". Les bases de la tarification en univers segmenté ont été jetées dans le Tome 1 (voyez la Section 3.8).

Comme nous l'avons vu à la Section 3.7, l'hétérogénéité au sein d'un portefeuille pose un grand nombre de problèmes, en particulier d'antisélection : si la même prime est appliquée à l'ensemble du portefeuille, les *mauvais* risques s'assureront (à un prix d'ailleurs moins élevé que celui qui devrait leur être réclamé), mais les *bons* pourraient être découragés par le tarif trop élevé, ce qui aura tendance à dégrader le résultat [1]. L'idée naturelle qui est développée dans les premières sections de ce chapitre est de partitionner le portefeuille afin de constituer des sous-portefeuilles sur lesquels les risques peuvent être considérés comme indépendants et de même loi. On parle alors de classes de risques. Les classes sont dites *a priori* lorsqu'il s'agit de classer le risque a partir d'information disponible a priori (sur l'assuré, le bien assuré...) et *a posteriori* si l'information sur l'historique des sinistres de l'assuré est prise en considération (comme cela sera fait dans les deux chapitres suivants).

Les Chapitres 3 et 4 du Tome 1 ont présenté les principes généraux de calcul des primes, reposant essentiellement sur le cal-

1. Insistons encore une fois sur l'absence de caractère péjoratif dans le qualificatif "mauvais". Il ne s'agit pas de stigmatiser une certaine catégorie d'individus mais bien de reconnaître techniquement que certains assurés causent davantage de sinistres, ou des sinistres plus coûteux, que d'autres.

cul de la prime pure, correspondant à l'espérance mathématique du coût annuel des sinistres déclarés à l'assureur. Cette prime pure se décomposait toutefois en deux composantes : la fréquence et le coût moyen, en notant simplement que

$$\mathbb{E}\left[\sum_{i=1}^{N} X_i\right] = \mathbb{E}\left[N\right] \times \mathbb{E}\left[X_i\right],$$

pour des montants de sinistres $X_1, X_2, \ldots$ indépendants et de même loi, indépendants de leur nombre N (voyez la Propriété 3.2.11). Il convient en effet de séparer ces deux notions pour plusieurs raisons : les facteurs explicatifs ne sont pas toujours les mêmes (en assurance automobile, la fréquence est une variable liée essentiellement au conducteur, et le coût moyen au véhicule), le coût moyen est soumis à l'inflation alors que la fréquence suit des cycles plus complexes, et si la fréquence peut être connue rapidement, le coût peut être long à estimer lors d'accidents corporels par exemple (ce dernier point sera abordé en détails dans le Chapitre **??** sur le provisionnement).

Exemple 9.1.1. *La Figure* **??** *compare, sur des données françaises, le coût moyen des sinistres et la fréquence pour des segments de population différents, avec un découpage prenant en compte le sexe, l'âge (18-25, 25-40, 40-65, plus de 65 ans), l'utilisation du véhicule (commerciale ou non), la puissance (graduée de A à K, K désignant les grosses cylindrées), pour quelques catégories socio-professionelle (profession libérale, fonction publique ou agriculture). Les deux axes permettent de représenter le comportement moyen sur l'ensemble de la population (fréquence de l'ordre de* $13,5\%$ *et coût moyen de* $1\ 100$ € *correspondant respectivement aux axes verticaux et horizontaux de la figure). On note que le point qui se détache le plus du nuage est celui des jeune conducteurs (hommes de 18 à 25 ans, classe H-18-25 en haut à droite), qui ont à la fois plus de sinistres et des sinistres plus coûteux. Les hommes de 25 à 40 ans se détachent également, en particulier dès lors qu'ils ont une voiture puissante ou qu'ils ont une profession libérale et qu'ils utilisent leur véhicule dans le cadre de leur travail. Les femmes exerçant une profession libérale sont égale mises en avant, avec une fréquence de sinistres très élevée, mais un coût moyen dans la moyenne. En revanche, les personnes âgées ou les hommes inactifs ont une fréquence relativement faible pour un coût moyen standard.*

De plus, si la modélisation directe de la prime pure peut paraître plus robuste et plus rapide, elle s'avère relativement complexe à

mettre en oeuvre puisqu'il est difficile de trouver des lois simples modélisant correctement la prime pure. En revanche, des lois simples permettent de modéliser les fréquences, et les coûts moyens (voir Chapitre 2).

9.2 Les variables tarifaires

En tarification, par habitude, mais aussi pour des raisons techniques que nous allons présenter ici, les variables tarifaires sont généralement des variables qualitatives. Les variables continues sont alors généralement regroupées par classes.

En assurance automobile, la tarification peut dépendre des caractéristiques propres du véhicule (puissance et vitesse de pointe, ancienneté du véhicule), de son usage, de la zone de circulation (densité du trafic), ou de certains traits spécifiques du conducteur habituel.

FIGURE 9.1 – Statistiques relatives au couple fréquence / coût moyen en assurance automobile, observation d'un portefeuille d'assurance français.

Les informations que tente d'avoir l'assureur sur l'assuré, afin de contrer l'antisélection doivent, d'un point de vue pratique, être vérifiables. Certaines informations qui pourraient être révélatrices d'un comportement à risque (c'est-à-dire fortement corrélées à la sinistralité) ne peuvent être utilisées car elles induiraient une forte fraude (point que nous n'aborderons pas dans ce chapitre) : ainsi, la distance parcourue par le véhicule chaque année est difficilement vérifiable.

Les variables explicatives composant $\boldsymbol{X}$ peuvent être de différents types. Certaines d'entre elles peuvent être quantitatives et continues (comme la puissance de la voiture ou l'âge de l'assuré par exemple). D'autres variables explicatives dont l'assureur dispose à propos de ses assurés peuvent être quantitatives discrètes (le nombre d'enfants de l'assuré, par exemple). D'autres encore sont qualitatives ou catégorielles (comme le sexe de l'assuré, par exemple). On peut également utiliser des indices résumant les caractéristiques de la zone d'habitation de l'assuré à partir de données publiques, comme celles issues d'un recensement. En condensant l'information pertinente à propos de la sinistralité disponible auprès d'un institut national de statistique, on peut ainsi obtenir de nouvelles variables explicatives

dont le pouvoir prédictif (reconnu par le modèle) est souvent très important. Nous reviendrons sur ce point à la Section **??**.

Les variables quantitatives n'appellent pas de commentaires particuliers. En ce qui concerne les variables catégorielles, on convient de coder tout facteur partitionnant la population en k catégories par les entiers $0, 1, \ldots, k-1$. Certains facteurs peuvent être ordinaux et résulter d'une variable quantitative (comme la puissance d'un véhicule que l'on a codée en différentes classes), soit ordinaux mais sans échelle quantitative (comme le niveau d'études) ou encore être purement qualitatif, sans induire d'ordre (comme le sexe). Une variable catégorielle à k facteurs est généralement codée par $k-1$ variables binaires qui sont toutes nulles pour le niveau de référence.

Exemple 9.2.1. *La plupart du temps, les variables explicatives sont toutes catégorielles dans un tarif commercial. Considérons par exemple un compagnie segmentant selon le sexe, le caractère sportif du véhicule et l'âge de l'assuré (3 classes d'âges, à savoir moins de 30 ans, 30-65 ans et plus de 65 ans). Un assuré sera représenté par un vecteur binaire donnant les valeurs des variables*

$$X_1 = \begin{cases} 0, \text{ si l'assuré est un homme,} \\ 1, \text{ si l'assuré est une femme,} \end{cases}$$

$$X_2 = \begin{cases} 0, \text{ si le véhicule n'a pas de caractère sportif,} \\ 1, \text{ si le véhicule a un caractère sportif,} \end{cases}$$

$$X_3 = \begin{cases} 1, \text{ si l'assuré a moins de 30 ans,} \\ 0, \text{ sinon,} \end{cases}$$

$$X_4 = \begin{cases} 1, \text{ si l'assuré a plus de 65 ans,} \\ 0, \text{ sinon.} \end{cases}$$

Pour chaque variable, on choisit comme niveau de référence (i.e. celui pour lequel toutes les variables binaires utilisées pour la coder valent simultanément 0) la modalité la plus représentée dans le portefeuille. Les résultats s'interprèteront ensuite comme une sur- ou sous-sinistralité par rapport à cette classe de référence. Ainsi, le vecteur (0,1,1,0) représente un assuré masculin de moins de 30 ans conduisant un véhicule sportif.

9.3 Principes de base de la statistique

Cette section rappelle brièvement les méthodes statistiques élémentaires. Notre présentation est intuitive et informelle. Pour plus de détails, nous renvoyons le lecteur par exemple à Monfort (1996).

9.3.1 Fonction de répartition empirique

Le cas non-groupé

Supposons disposer des observations $x_1, x_2, \ldots, x_n$ et considérons-les comme autant de réalisations de variables aléatoires $X_1, X_2, \ldots, X_n$, indépendantes et de même fonction de répartition F. Ces observations peuvent par exemple représenter le coût de sinistres déclarés à la compagnie [2].

La fonction de répartition empirique, notée $\hat{F}_n$, donne une idée de l'allure de F à partir des observations $x_1, x_2, \ldots, x_n$. Elle s'obtient en accordant une masse de probabilité de $1/n$ à chacun des x_i, i.e.

$$\hat{F}_n(x) = \frac{\#\{x_i \text{ tels que } x_i \leq x\}}{n}, \quad x \in I\!R.$$

En d'autres termes, $\hat{F}_n(x)$ est la proportion d'observations dans l'échantillon qui sont inférieures ou égales à $x \in I\!R$. La fonction $x \mapsto \hat{F}_n(x)$ est une fonction "en escaliers", présentant un saut d'amplitude k/n à chaque valeur apparaissant k fois dans l'échantillon.

L'approche empirique consiste à utiliser $\hat{F}_n$ en lieu et place de F pour tous les calculs actuariels. Cette façon de faire est justifiée par le théorème de Glivenko-Cantelli qui assure que

$$\Pr\Big[\sup_{x \in I\!R} |\hat{F}_n(x) - F(x)| \to 0 \text{ lorsque } n \text{ tend vers } +\infty\Big] = 1.$$

En d'autres termes, le graphe de $\hat{F}_n$ épouse d'autant mieux celui de F à mesure que le n nombre d'observations, i.e. notre information, augmente ; le graphe de $\hat{F}_n$ devrait donc fournir une bonne approximation de celui de F pour n "grand".

La loi de $n\widehat{F}_n(x)$ est $\mathcal{B}in(n, F(x))$. Dès lors, en vertu du théorème central-limite, on a quel que soit x,

$$\sqrt{n}\Big(\widehat{F}_n(x) - F(x)\Big) \to_{\text{loi}} \mathcal{N}or(0, \sigma_x^2) \text{ lorsque } n \to +\infty$$

où $\sigma_x^2 = F(x)(1-F(x))$. Pour autant que n soit suffisamment grand, ceci permet d'obtenir un intervalle de confiance pour la valeur inconnue $F(x)$.

2. Lorsque les observations se rapportent à une longue période, il convient le cas échéant de corriger les montants x_i de l'effet de l'inflation, et de prendre en compte les modifications légales ou réglementaires de même que les changements dans les clauses des polices de nature à influencer le montant des sinistres.

Remarque 9.3.1. *On peut également "lisser" la fonction de répartition empirique et en déduire une fonction de densité empirique $\widehat{f}_n$. Ceci conduit à un estimateur à noyau de la densité, défini par*

$$\widehat{f}_K(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right)$$

où K est appelé noyau et h est un paramètre positif (appelé bandwidth *en anglais) donnant l'importance du lissage. Dans une première approche, on peut retenir un noyau gaussien, i.e. $K(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)$. Notez que $\widehat{f}_K$ constitue un estimateur convergent de la densité lorsque $n \to +\infty$ et $h \to 0$ à une vitesse adéquate.*

Le cas groupé

Bien souvent, les données dont dispose l'actuaire sont groupées en classes, plus ou moins larges. Ceci revient à ne pas distinguer les sinistres ne différant que de quelques unités monétaires. L'actuaire ne dispose dès lors plus des observations $x_1, x_2, \ldots, x_n$. Par contre, si l'on note $c_0 < c_1 < c_2 < \ldots < c_r$ les limites des r classes, il sait que n_j sinistres ont un montant compris entre c_{j-1} et c_j, $j = 1, 2, \ldots, r$; on appelle n_j l'effectif de la classe $C_j =]c_{j-1}, c_j]$. Souvent, $c_0 = 0$, de sorte que la première classe est du type "sinistres de montant $\leq c_1$". Parfois, la limite supérieure c_r de C_r n'est pas spécifiée ; la dernière classe est alors du type "sinistres de montant $> c_{r-1}$" et on considère que $c_r = +\infty$. Parfois, la moyenne m_j des montants des sinistres tombant dans la classe C_j est également fournie.

Le groupement des données a pour conséquence que la fonction de répartition empirique $\hat{F}_n$ n'est connue avec précision qu'aux limites de classes c_j, où elle vaut $\hat{F}_n(c_0) = 0$ et

$$\hat{F}_n(c_j) = \begin{cases} 0, \text{ si } j = 0 \\ \frac{1}{n} \sum_{i=1}^{j} n_i, \quad j = 1, 2, \ldots, r. \end{cases}$$

On approxime alors $\hat{F}_n$ par interpolation linéaire sur les segments $]c_{j-1}, c_j]$ pour obtenir

$$\hat{F}_n(x) = \begin{cases} 0, \text{ si } x < c_0, \\ \frac{(c_j - x)\hat{F}_n(c_{j-1}) + (x - c_{j-1})\hat{F}_n(c_j)}{c_j - c_{j-1}}, \text{ si } c_{j-1} \leq x < c_j, \\ 1, \text{ si } x \geq c_r. \end{cases}$$

Il est à noter que dans le cas groupé, $\hat{F}_n$ n'est pas définie sur $]c_{r-1}, +\infty[$ lorsque $c_r = +\infty$, à moins que $n_r = 0$.

Remarque 9.3.2. *Puisque $\hat{F}_n$ est à présent une fonction linéaire par morceaux, $\hat{F}_n$ est dérivable partout, excepté aux extrémités des classes $c_0, c_1, \ldots, c_r$, où les dérivées à droite et à gauche existent néanmoins. Cette dérivée estime la fonction de densité f associée à F ; nous noterons $\hat{f}_n$ cet estimateur, encore appelé histogramme. L'estimateur $\hat{f}_n$ est donné par*

$$\hat{f}_n(x) = \begin{cases} 0, \ si\ x < c_0, \\ \frac{\hat{F}_n(c_j) - \hat{F}_n(c_{j-1})}{c_j - c_{j-1}} = \frac{n_j}{n(c_j - c_{j-1})}, \ si\ c_{j-1} \leq x < c_j, \\ 0 \ si\ x \geq c_r. \end{cases}$$

Le graphe de la fonction $x \mapsto \hat{f}_n(x)$ apparaît cette fois comme une succession de blocs. L'aire sous le graphe de $\hat{f}_n$ vaut 1, par construction, excepté lorsque $c_r = +\infty$, auquel cas on ne peut pas représenter la probabilité de la classe C_r. Il est important de noter que c'est l'aire, et non pas la hauteur, des blocs qui est proportionnelle au nombre d'observations de chaque classe. Ceci revient à tenir compte de la longeur des classes dans le calcul des fréquences relatives.

De manière générale, par souci de précision, l'actuaire prendra soin de conserver les données brutes $x_1, x_2, \ldots, x_n$ pour effectuer les calculs, mais aura souvent recours à un groupement pour la présentation de ses résultats.

Estimation des paramètres principaux

Une fois que l'on dispose de $\hat{F}_n$, on peut facilement estimer les paramètres principaux de F, à savoir la moyenne, la variance, le coefficient de variation ainsi que les différents quantiles. A chaque fois, il suffit de remplacer la fonction de répartition F inconnue par son équivalent empirique $\hat{F}_n$.

La moyenne Afin d'estimer la moyenne, on a naturellement recours à la moyenne échantillon

$$\hat{\mu}_1 = \int_{x \in \mathbb{R}} x d\hat{F}_n(x) = \begin{cases} \frac{1}{n} \sum_{i=1}^{n} x_i \text{ (cas non-groupé)} \\ \sum_{j=1}^{r} \frac{n_j(c_j + c_{j-1})}{2n} \text{ (cas groupé).} \end{cases}$$

Bien souvent, on utilise la notation $\overline{x}$ en lieu et place de $\hat{\mu}_1$. La moyenne observée $\overline{x}$ apparaît ainsi comme le centre de gravité des données, et est fort sensible aux valeurs "extrêmes".

La variance En notant,

$$\hat{\mu}_2 = \int_{x\in I\!R} x^2 d\hat{F}_n(x) = \begin{cases} \frac{1}{n}\sum_{i=1}^{n} x_i^2 \text{ (cas non-groupé)} \\ \sum_{j=1}^{r} \frac{n_j(c_j^3 - c_{j-1}^3)}{3n(c_j - c_{j-1})} \text{ (cas groupé)} \end{cases}$$

l'estimateur naturel de la variance est alors donné par $s^2 = \hat{\mu}_2 - \hat{\mu}_1^2$. On appelle écart-type observé la racine carrée positive de la variance échantillon s^2 ; il est noté s.

Le coefficient de variation Bien souvent, l'actuaire a recours au coefficient de variation cv, défini comme le rapport de l'écart-type échantillon à la moyenne échantillon, à savoir $cv = \frac{s}{\bar{x}}$. Le coefficient de variation a le grand avantage d'être un nombre sans dimension, ce qui facilite les comparaisons (en excluant par exemple les effets des différentes unités monétaires). Il peut se voir comme une normalisation de l'écart-type.

Quantiles Les quantiles empiriques, notés $\hat{q}_p$, sont simplement obtenus grâce à la relation

$$\hat{q}_p = \hat{F}_n^{-1}(p) = \inf\{x \in I\!R | \hat{F}_n(x) \geq p\}, \;\; 0 < p < 1.$$

Dans le cas non-groupé, $\hat{F}_n^{-1}$ fait correspondre à p, $0 < p < 1$, la plus petite des observations laissant à sa gauche au moins $100p\%$ des données. Si l'on note $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ les observations rangées dans l'ordre croissant, à savoir $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$, on a que

$$\hat{q}_p = x_{(i)} \text{ pour } p \in \left]\frac{i-1}{n}, \frac{i}{n}\right].$$

Dans le cas groupé, l'estimation du $p$ème quantile se fait comme suit. On détermine tout d'abord j_0 tel que

$$\hat{F}_n(c_{j_0-1}) \leq p < \hat{F}_n(c_{j_0}).$$

Il vient alors

$$\hat{F}_n(\hat{q}_p) = \frac{(c_{j_0} - \hat{q}_p)\hat{F}_n(c_{j_0-1}) + (\hat{q}_p - c_{j_0-1})\hat{F}_n(c_{j_0})}{c_{j_0} - c_{j_0-1}} = p,$$

d'où l'on tire

$$\hat{q}_p = \frac{c_{j_0} - c_{j_0-1}}{\hat{F}_n(c_{j_0}) - \hat{F}_n(c_{j_0-1})} p - \frac{c_{j_0}\hat{F}_n(c_{j_0-1}) - c_{j_0-1}\hat{F}_n(c_{j_0})}{\hat{F}_n(c_{j_0}) - \hat{F}_n(c_{j_0-1})}.$$

9.3.2 L'approche paramétrique

L'approche empirique ne permet pas de répondre à toutes les questions que se pose l'actuaire. Afin de s'en convaincre, examinons l'exemple suivant.

Exemple 9.3.3. *Un assureur commercialise des polices prévoyant un découvert obligatoire de 50 € par sinistre. Il a enregistré les coûts suivants (en €) : 141, 16, 46, 40, 351, 259, 317, 1 511, 107 et 567. Le montant moyen $\overline{x}$ payé par l'assureur par sinistre s'élève à 335.5 € . Supposons à présent que l'assureur augmente le découvert obligatoire pour le porter à 100 € . On peut alors estimer le coût moyen par sinistre à*

$$\frac{91 + 301 + 209 + 267 + 1461 + 57 + 517}{7} = 414.71.$$

L'assureur qui payait autrefois 3 355 € pour dédommager les assurés, ne doit plus désormais débourser que 2 903 € .

Le passage du découvert obligatoire de 50 à 100 € permettrait ainsi de réduire les coûts de

$$\frac{3\,355 - 2\,903}{3\,355} = 13.47\%.$$

Supposons à présent au contraire que l'assureur désire supprimer la clause de découvert obligatoire, par exemple dans le but de conquérir de nouvelles parts de marché. L'approche empirique ne nous permet pas d'évaluer le surcoût engendré par cette modification contractuelle. En effet, les montants des sinistres enregistrés vont être augmentés de 50 € chacun, portant le coût total à 3 855 € , mais il ne faut pas perdre de vue que les sinistres dont le coût est inférieur à 50 € , à propos desquels nous ne disposons d'aucune information (puisqu'ils n'étaient pas rapportés à la compagnie), vont à présent ouvrir le droit à l'indemnisation. Tout ce que l'approche empirique permet de dire c'est que le surcoût engendré par l'abandon de la clause de découvert obligatoire s'élève au moins à

$$\frac{500}{3\,355} = 14.9\%.$$

Supposons à présent que la compagnie introduise dans les conditions générales de sa police une limite supérieure à l'indemnisation, fixée à 1 500 € par sinistre, disons. Ceci n'engendrera aucune modification quant au coût moyen de l'assureur, puisqu'aucun sinistre

de montant supérieur à 1 500 € n'a été observé. Toutefois, tout assuré rationnel attendra très certainement une réduction de prime compensant le surcroît de risque à sa charge, ou à tout le moins un avantage quelconque.

Supposons à présent que les montants de sinistre subissent une inflation de 10%. Que peut-on dire de la charge moyenne de sinistre ? On peut facilement obtenir les montants corrigés d'inflation comme suit :

$$\begin{aligned} 1.1(141+50)-50 &= 160.1 \\ 1.1(16+50)-50 &= 22.6 \\ \text{etc.} &\quad , \end{aligned}$$

ce qui fournit la nouvelle série 160.1, 22.6, 55.6, 391.1, 289.9, 353.7, 1 667.1, 122.7 et 628.7. On serait tenté d'affirmer que le montant moyen payé par la compagnie par sinistre s'élève à présent à 374.05 € . Néanmoins, ce serait faire fi des sinistres tombant dans la tranche 45.45-50, qui n'ont pas été rapportés à l'assureur, du fait du découvert obligatoire de 50 € , et qui, suite à l'inflation tombent désormais sous la garantie. L'inflation entraîne donc non seulement une modification des montants des sinistres, mais également une modification du nombre de sinistres rapportés à la compagnie.

L'alternative permettant de traiter les questions restées sans réponse dans l'exemple ci-dessus est l'approche paramétrique. Dans cette approche, on suppose que la fonction de répartition inconnue F du montant payé par l'assureur suite à un sinistre fait partie d'une famille $\mathcal{F} = \{F_{\boldsymbol{\theta}},\ \boldsymbol{\theta} \in \Theta\}$ de fonctions de répartition dont la forme analytique est connue, mais qui dépendent d'un ou de plusieurs paramètres inconnus $\boldsymbol{\theta}$; Θ note l'espace paramétrique, c'est-à-dire l'ensemble de toutes les valeurs permises pour le paramètre. Notez que $\boldsymbol{\theta}$ est tantôt unidimensionnel, tantôt vectoriel, selon la famille paramétrique considérée. Dès lors, identifier F dans $\mathcal{F}$ revient à déterminer la valeur de $\boldsymbol{\theta}$ telle que $F \equiv F_{\boldsymbol{\theta}}$.

Afin de bien saisir la complémentarité des approches paramétriques et empiriques, poursuivons l'étude du cas de figure présenté ci-dessus dans l'optique paramétrique.

Exemple 9.3.4 (Suite de l'Exemple **??**)**.** *Supposons que l'on puisse valablement considérer que le montant d'un sinistre suit une loi exponentielle négative de moyenne θ. Notons X le montant du sinistre et Y le montant de l'indemnité versée par l'assureur. Une manière*

commode de fixer la valeur de θ consiste à égaler $\mathbb{E}[Y]$ et $\overline{x}$, puisque

$$\begin{aligned}\mathbb{E}[Y] &= \mathbb{E}[X-50|X>50] \\ &= \int_{x=50}^{+\infty}(x-50)\frac{\exp(-x/\theta)}{\theta\exp(-50/\theta)}dx=\theta.\end{aligned}$$

Ainsi, on choisit $\theta = 335.5$. Suite à l'inflation, la charge moyenne d'un sinistre pour l'assureur passe à

$$\mathbb{E}[1.1X-50|1.1X>50]=369.05.$$

D'autre part, la probabilité qu'un sinistre tombe sous la couverture était de

$$\Pr[X>50]=0.86154$$

sans inflation, alors qu'elle passe à

$$\Pr[1.1X>50]=0.87329$$

suite à l'effet de l'inflation.

Comme le montre cet exemple très simple, l'approche paramétrique permet de répondre à toutes les questions que se pose l'assureur.

9.3.3 L'information de Fisher

Conditions sur le modèle statistique

Considérons un modèle statistique paramétrique $\mathcal{F}=\{F_{\boldsymbol{\theta}},\ \boldsymbol{\theta}\in\Theta\}$ de support commun $\mathcal{S}$, dont l'espace paramétrique Θ est un ouvert de $\mathbb{R}^p$. On note $f_{\boldsymbol{\theta}}$ la densité de probabilité (discrète ou continue) associée à $F_{\boldsymbol{\theta}}$ et on supposera désormais que les trois conditions suivantes sont satisfaites :

(i) $f_{\boldsymbol{\theta}}(x)>0$ pour tout $x\in\mathcal{S}$ et $\boldsymbol{\theta}\in\Theta$;

(ii) $\frac{\partial}{\partial\boldsymbol{\theta}}f_{\boldsymbol{\theta}}(x)$ existe pour tout $x\in\mathcal{S}$ et $\boldsymbol{\theta}\in\Theta$;

(iii) quel que soit $\boldsymbol{\theta}$ dans Θ, on peut dériver $\int_{x\in A}dF_{\boldsymbol{\theta}}(x)$ sous le signe intégral par rapport aux composantes de $\boldsymbol{\theta}$.

Définition

Considérons une variable aléatoire X de fonction de répartition $F_{\boldsymbol{\theta}}$, pour $\boldsymbol{\theta}\in\Theta$. On appelle information de Fisher du modèle la matrice variance-covariance, si elle existe, du vecteur aléatoire

$$\frac{\partial}{\partial\boldsymbol{\theta}}\ln f_{\boldsymbol{\theta}}(X)=\left(\frac{\partial}{\partial\theta_1}\ln f_{\boldsymbol{\theta}}(X),\ldots,\frac{\partial}{\partial\theta_p}\ln f_{\boldsymbol{\theta}}(X)\right)^t;$$

cette matrice sera notée $\mathcal{I}(\boldsymbol{\theta})$.

Propriété 9.3.5. *Le vecteur* $\frac{\partial}{\partial \boldsymbol{\theta}} \ln f_{\boldsymbol{\theta}}(X)$ *est centré, i.e.*

$$\mathbb{E}\left[\frac{\partial}{\partial \boldsymbol{\theta}} \ln f_{\boldsymbol{\theta}}(X)\right] = \mathbf{0}.$$

Démonstration. Considérons le cas continu. Partant de

$$\int_{x\in\mathbb{R}} f_{\boldsymbol{\theta}}(x)dx = 1$$

on tire en dérivant les deux membres par rapport à θ_i, $i = 1, 2, \ldots, p$,

$$\begin{aligned} 0 &= \frac{\partial}{\partial \theta_i} \int_{x\in\mathbb{R}} f_{\boldsymbol{\theta}}(x)dx \\ &= \int_{x\in\mathbb{R}} \frac{\partial}{\partial \theta_i} f_{\boldsymbol{\theta}}(x)dx \\ &= \int_{x\in\mathbb{R}} \left\{\frac{\partial}{\partial \theta_i} \ln f_{\boldsymbol{\theta}}(x)\right\} f_{\boldsymbol{\theta}}(x)dx \\ &= \mathbb{E}\left[\frac{\partial}{\partial \theta_i} \ln f_{\boldsymbol{\theta}}(X)\right]. \end{aligned}$$

□

Par conséquent, la Propriété **??** permet d'obtenir l'expression de $\mathcal{I}(\boldsymbol{\theta})$, puisque

$$\mathbb{C}\left[\frac{\partial}{\partial \theta_i} \ln f_{\boldsymbol{\theta}}(X), \frac{\partial}{\partial \theta_j} \ln f_{\boldsymbol{\theta}}(X)\right] = \mathbb{E}\left[\frac{\partial}{\partial \theta_i} \ln f_{\boldsymbol{\theta}}(X) \frac{\partial}{\partial \theta_j} \ln f_{\boldsymbol{\theta}}(X)\right].$$

Ceci donne sous forme matricielle

$$\mathcal{I}(\boldsymbol{\theta}) = \mathbb{E}\left[\frac{\partial}{\partial \boldsymbol{\theta}} \ln f_{\boldsymbol{\theta}}(X) \left(\frac{\partial}{\partial \boldsymbol{\theta}} \ln f_{\boldsymbol{\theta}}(X)\right)^t\right].$$

Expression alternative

De manière générale, pour autant que $\int_{x\in\mathbb{R}} dF_{\boldsymbol{\theta}}(x)$ soit dérivable deux fois sous le signe intégral, nous obtenons deux expressions

équivalentes pour $\mathcal{I}(\boldsymbol{\theta})$. En effet, considérons le cas continu et partons de

$$\begin{aligned} 0 &= \frac{\partial^2}{\partial\theta_i\partial\theta_j}\int_{x\in\mathbb{R}} f_{\boldsymbol{\theta}}(x)dx \\ &= \int_{x\in\mathbb{R}} \frac{\partial^2}{\partial\theta_i\partial\theta_j} f_{\boldsymbol{\theta}}(x)dx \\ &= \int_{x\in\mathbb{R}} \frac{\frac{\partial^2}{\partial\theta_i\partial\theta_j} f_{\boldsymbol{\theta}}(x)}{f_{\boldsymbol{\theta}}(x)} f_{\boldsymbol{\theta}}(x)dx \\ &= \mathbb{E}\left[\frac{\frac{\partial^2}{\partial\theta_i\partial\theta_j} f_{\boldsymbol{\theta}}(X)}{f_{\boldsymbol{\theta}}(X)}\right]. \end{aligned}$$

On obtient alors

$$\begin{aligned} \mathbb{E}\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ln f_{\boldsymbol{\theta}}(X)\right] &= \mathbb{E}\left[\frac{\partial}{\partial\theta_i}\frac{\frac{\partial}{\partial\theta_j} f_{\boldsymbol{\theta}}(X)}{f_{\boldsymbol{\theta}}(X)}\right] \\ &= \mathbb{E}\left[\frac{\frac{\partial^2}{\partial\theta_i\partial\theta_j} f_{\boldsymbol{\theta}}(X)}{f_{\boldsymbol{\theta}}(X)}\right] - \mathbb{E}\left[\frac{\frac{\partial}{\partial\theta_i} f_{\boldsymbol{\theta}}(X)\frac{\partial}{\partial\theta_j} f_{\boldsymbol{\theta}}(X)}{\{f_{\boldsymbol{\theta}}(X)\}^2}\right] \\ &= -\mathbb{E}\left[\frac{\frac{\partial}{\partial\theta_i} f_{\boldsymbol{\theta}}(X)\frac{\partial}{\partial\theta_j} f_{\boldsymbol{\theta}}(X)}{\{f_{\boldsymbol{\theta}}(X)\}^2}\right] \\ &= -\mathbb{E}\left[\frac{\partial}{\partial\theta_i}\ln f_{\boldsymbol{\theta}}(X)\frac{\partial}{\partial\theta_j}\ln f_{\boldsymbol{\theta}}(X)\right]. \end{aligned}$$

Nous pouvons alors écrire sous forme matricielle

$$\begin{aligned} \mathcal{I}(\boldsymbol{\theta}) &= \mathbb{E}\left[\frac{\partial}{\partial\boldsymbol{\theta}}\ln f_{\boldsymbol{\theta}}(X)\left(\frac{\partial}{\partial\boldsymbol{\theta}}\ln f_{\boldsymbol{\theta}}(X)\right)^t\right] \\ &= -\mathbb{E}\left[\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^t}\ln f_{\boldsymbol{\theta}}(X)\right]. \end{aligned} \tag{9.1}$$

L'information de Fisher $\mathcal{I}(\boldsymbol{\theta})$ peut donc s'obtenir soit comme l'espérance du produit des vecteurs gradients de $\ln f_{\boldsymbol{\theta}}(X)$, soit comme l'opposé de l'espérance de la matrice hessienne de $\ln f_{\boldsymbol{\theta}}(X)$.

Information de Kullback

Supposons que $\widetilde{\boldsymbol{\theta}}$ soit la vraie valeur du paramètre et définissons, pour le résultat x, le pouvoir discriminant de x entre la vraie valeur

$\widetilde{\boldsymbol{\theta}}$ du paramètre et une autre valeur possible $\boldsymbol{\theta}$ pour le paramètre comme

$$\ln \frac{f_{\widetilde{\boldsymbol{\theta}}}(x)}{f_{\boldsymbol{\theta}}(x)}. \tag{9.2}$$

Il faut interpéter (**??**) comme le logarithme du rapport entre la "chance" d'observer x pour la vraie valeur du paramètre $\widetilde{\boldsymbol{\theta}}$ et la "chance" d'observer x si le paramètre vaut $\boldsymbol{\theta}$. Pour les valeurs de x telles que $f_{\boldsymbol{\theta}}(x) > f_{\widetilde{\boldsymbol{\theta}}}(x)$, la quantité (**??**) est négative, ce qui est naturel puisqu'au vu d'un tel résultat, l'actuaire serait plutôt enclin à conclure en faveur de $\boldsymbol{\theta}$.

Kullback a ainsi proposé de définir un "pouvoir discriminant moyen", ou une "information discriminante moyenne" comme suit : l'information de Kullback de $\widetilde{\boldsymbol{\theta}}$ contre $\boldsymbol{\theta}$ est définie par

$$\mathbb{E}_{\widetilde{\boldsymbol{\theta}}}\left[\ln \frac{f_{\widetilde{\boldsymbol{\theta}}}(X)}{f_{\boldsymbol{\theta}}(X)}\right] = \int_{x\in\mathbb{R}} \ln \frac{f_{\widetilde{\boldsymbol{\theta}}}(x)}{f_{\boldsymbol{\theta}}(x)} f_{\widetilde{\boldsymbol{\theta}}}(x)dx.$$

On la note $\mathcal{I}(\widetilde{\boldsymbol{\theta}}|\boldsymbol{\theta})$. On voit facilement que $\mathcal{I}(\widetilde{\boldsymbol{\theta}}|\boldsymbol{\theta}) \geq 0$, en effet,

$$\mathcal{I}(\widetilde{\boldsymbol{\theta}}|\boldsymbol{\theta}) = -\mathbb{E}_{\widetilde{\boldsymbol{\theta}}}\left[\ln \frac{f_{\boldsymbol{\theta}}(X)}{f_{\widetilde{\boldsymbol{\theta}}}(X)}\right] \geq -\ln \mathbb{E}_{\widetilde{\boldsymbol{\theta}}}\left[\frac{f_{\boldsymbol{\theta}}(X)}{f_{\widetilde{\boldsymbol{\theta}}}(X)}\right]$$

en vertu de l'inégalité de Jensen appliquée à la fonction $-\ln$. Or,

$$\mathbb{E}_{\widetilde{\boldsymbol{\theta}}}\left[\frac{f_{\boldsymbol{\theta}}(X)}{f_{\widetilde{\boldsymbol{\theta}}}(X)}\right] = \int_{x\in\mathbb{R}} f_{\boldsymbol{\theta}}(x)dx = 1,$$

ce qui achève la justification. De plus,

$$\mathcal{I}(\widetilde{\boldsymbol{\theta}}|\boldsymbol{\theta}) = 0 \Leftrightarrow f_{\widetilde{\boldsymbol{\theta}}} \equiv f_{\boldsymbol{\theta}}.$$

Notons cependant que $\mathcal{I}(\widetilde{\boldsymbol{\theta}}|\boldsymbol{\theta})$ n'est en général pas une métrique car elle n'est pas nécessairement symétrique et ne vérifie pas toujours l'inégalité triangulaire.

Lien entre les informations de Fisher et de Kullback

Faisons à présent le lien entre les deux types d'information ; pour ce faire supposons à nouveau que les trois hypothèses techniques énumérées en début de section sont verifiées et que Θ est un ouvert de $\mathbb{R}^p$. Plaçons-nous dans le cas continu et partons de

$$\mathcal{I}(\widetilde{\boldsymbol{\theta}}|\boldsymbol{\theta}) = \int_{x\in\mathbb{R}} \ln \frac{f_{\widetilde{\boldsymbol{\theta}}}(x)}{f_{\boldsymbol{\theta}}(x)} f_{\widetilde{\boldsymbol{\theta}}}(x)dx$$

et dérivons par rapport à θ_i pour obtenir

$$\frac{\partial}{\partial\theta_i}\mathcal{I}(\widetilde{\boldsymbol{\theta}}|\boldsymbol{\theta}) = \int_{x\in\mathbb{R}} \left\{ -\frac{\partial}{\partial\theta_i} \ln f_{\boldsymbol{\theta}}(x) \right\} f_{\widetilde{\boldsymbol{\theta}}}(x)dx.$$

Au passage, on vérifie bien que pour $i = 1, 2, \ldots, p$,

$$\frac{\partial}{\partial\theta_i}\mathcal{I}(\widetilde{\boldsymbol{\theta}}|\boldsymbol{\theta})\bigg|_{\boldsymbol{\theta}=\widetilde{\boldsymbol{\theta}}} = 0,$$

ce qui est naturel puisque $\mathcal{I}(\widetilde{\boldsymbol{\theta}}|\boldsymbol{\theta})$ atteint son minimum 0 pour $\boldsymbol{\theta} = \widetilde{\boldsymbol{\theta}}$. Considérons les dérivées secondes ; on a

$$\frac{\partial^2}{\partial\theta_i\partial\theta_j}\mathcal{I}(\widetilde{\boldsymbol{\theta}}|\boldsymbol{\theta}) = \int_{x\in\mathbb{R}} \left\{ \frac{\frac{\partial}{\partial\theta_i} f_{\boldsymbol{\theta}}(x) \frac{\partial}{\partial\theta_j} f_{\boldsymbol{\theta}}(x)}{\{f_{\boldsymbol{\theta}}(x)\}^2} - \frac{\frac{\partial^2}{\partial\theta_i\partial\theta_j} f_{\boldsymbol{\theta}}(x)}{f_{\boldsymbol{\theta}}(x)} \right\} f_{\widetilde{\boldsymbol{\theta}}}(x)dx$$

d'où l'on tire

$$\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^t}\mathcal{I}(\widetilde{\boldsymbol{\theta}}|\boldsymbol{\theta})\bigg|_{\boldsymbol{\theta}=\widetilde{\boldsymbol{\theta}}} = \mathbb{V}\left[\frac{\partial}{\partial\boldsymbol{\theta}} \ln f_{\widetilde{\boldsymbol{\theta}}}(X)\right] = \mathcal{I}(\widetilde{\boldsymbol{\theta}}) \tag{9.3}$$

puisque $\int_{x\in\mathbb{R}} \frac{\partial^2}{\partial\theta_i\partial\theta_j} f_{\boldsymbol{\theta}}(x)dx = 0$. On peut aussi vérifier que la matrice des dérivées secondes de $\mathcal{I}(\widetilde{\boldsymbol{\theta}}|\boldsymbol{\theta})$ est semi-définie positive en $\boldsymbol{\theta} = \widetilde{\boldsymbol{\theta}}$.

La relation (**??**) donne le lien entre l'information de Fisher et l'information de Kullback. C'est cette relation qui nous permet d'interpréter $\mathcal{I}(\boldsymbol{\theta})$, qui jouera un rôle très important dans la suite. En effet, la matrice $\mathcal{I}(\widetilde{\boldsymbol{\theta}})$ décrit le comportement local de la fonction $\boldsymbol{\theta} \mapsto \mathcal{I}(\widetilde{\boldsymbol{\theta}}|\boldsymbol{\theta})$ au voisinage du point $\widetilde{\boldsymbol{\theta}}$ pour lequel elle atteint son minimum 0. Ainsi, pour $p = 1$ (i.e. il y a un seul paramètre)

(i) si $\mathcal{I}(\widetilde{\theta})$ est voisin de 0, la courbe $\mathcal{I}(\widetilde{\theta}|\theta)$ est très plate au voisinage de $\widetilde{\theta}$ et on peut mal discriminer $\widetilde{\theta}$ des valeurs θ de son voisinage ;

(ii) si $\mathcal{I}(\widetilde{\theta})$ est élevée, au contraire, on peut facilement distinguer la vraie valeur $\widetilde{\theta}$ du paramètre des valeurs au voisinage de celles-ci.

9.3.4 Estimation des paramètres par la méthode du maximum de vraisemblance

Fonction de vraisemblance

Nous considérerons dorénavant des modèles identifiables, c'est-à-dire tels que l'application $\boldsymbol{\theta} \mapsto F_{\boldsymbol{\theta}}$ est injective. Supposons avoir

sélectionné la famille paramétrique $\{F_{\boldsymbol{\theta}},\ \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p\}$. Nous cherchons à déterminer la valeur de $\boldsymbol{\theta}$ la plus plausible compte tenu des observations $x_1, x_2, \ldots, x_n$ dont on dispose ; nous noterons $\hat{\boldsymbol{\theta}}$ cette valeur, dite estimation du paramètre d'intérêt.

La méthode du maximum de vraisemblance requiert la définition de la fonction de vraisemblance, notée $\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{x})$, valant

1. dans le cas non-groupé :

$$\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{x}) = \prod_{j=1}^{n} f_{\boldsymbol{\theta}}(x_j),$$

   où $f_{\boldsymbol{\theta}}$ est la densité de probabilité (discrète ou continue) associée à $F_{\boldsymbol{\theta}}$;

2. dans le cas groupé :

$$\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{x}) = \prod_{j=1}^{r} (F_{\boldsymbol{\theta}}(c_j) - F_{\boldsymbol{\theta}}(c_{j-1}))^{n_j}.$$

Intuitivement, il faut considérer $\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{x})$ comme la "chance" d'observer les valeurs $x_1, x_2, \ldots, x_n$ pour la valeur $\boldsymbol{\theta}$ du paramètre. Il est très important de noter que $\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{x})$ est une fonction de $\boldsymbol{\theta}$, les observations $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^t$ étant données, alors que la densité jointe des observations $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \prod_{i=1}^{n} f_{\boldsymbol{\theta}}(x_i)$ est une fonction des observations $x_1, x_2, \ldots, x_n$, paramétrée par $\boldsymbol{\theta}$.

Méthode d'estimation

L'estimateur du maximum de vraisemblance $\widehat{\boldsymbol{\theta}}$ de $\boldsymbol{\theta}$ est obtenu en maximisant la "chance" d'observer $x_1, x_2, \ldots, x_n$, i.e.

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \Theta} \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{x}). \tag{9.4}$$

Il s'agit somme toute d'une méthode assez intuitive lorsqu'on dispose d'un échantillon fiable : on estimera le paramètre $\boldsymbol{\theta}$ à la valeur $\widehat{\boldsymbol{\theta}}$ maximisant la probabilité de recueillir les observations $x_1, x_2, \ldots, x_n$ de l'échantillon. En pratique, il est souvent plus facile de passer au logarithme avant l'étape de maximisation. On définit ainsi la log-vraisemblance $L(\boldsymbol{\theta}|\boldsymbol{x})$ associée à l'échantillon, donnée par

$$L(\boldsymbol{\theta}|\boldsymbol{x}) = \ln \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{x}).$$

L'estimateur du maximum de vraisemblance $\widehat{\boldsymbol{\theta}}$ de $\boldsymbol{\theta}$ est obtenu grâce au programme de maximisation

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}|\boldsymbol{x}). \tag{9.5}$$

Remarque 9.3.6 (Censure)**.** *Selon les cas, la fonction de vraisemblance peut encore prendre d'autres formes que celles évoquées ci-dessus. Dans la plupart des branches d'assurance, le montant des indemnités versées par l'assureur ne correspond pas exactement au préjudice subi par l'assuré, suite à l'introduction de clauses de franchise ou de découvert obligatoire dans les conditions des polices, voire la spécification d'un plafond d'intervention. Ceci a un effet important sur les méthodes d'estimation. En effet, l'actuaire désire modéliser le montant du sinistre, pas celui de l'intervention de l'assureur. Par exemple, si l'assureur a introduit un plafond à son obligation de garantie, fixé à ω disons, et que l'actuaire dispose des données x_1, x_2, ..., x_k, ω, ω, ..., ω, la vraisemblance $\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{x})$ s'écrit sous la forme*

$$\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{x}) = \big(\overline{F}_{\boldsymbol{\theta}}(\omega)\big)^{n-k} \prod_{j=1}^{k} f_{\boldsymbol{\theta}}(x_j).$$

En effet, une observation ω signifie en réalité que le montant du sinistre s'élevait au moins à ω, mais que l'assureur n'a payé qu'une indemnité de ω.

Propriétés

Les estimateurs du maximum de vraisemblance possèdent d'excellente propriétés théoriques. Sous des hypothèses assez générales, ils sont asymptotiquement sans biais et efficaces (c'est-à-dire qu'ils sont les plus précis), et font toujours le meilleur usage des informations contenues dans l'échantillon.

Intéressons-nous à présent au comportement des estimateurs du maximum de vraisemblance en grands échantillons. Notons $\widehat{\boldsymbol{\theta}}_n$ l'estimateur du maximum de vraisemblance de $\boldsymbol{\theta}$ obtenu à partir d'un échantillon de taille n. Lorsque Θ est un ouvert de $\mathbb{R}^p$ et pour autant que le modèle soit identifiable, montrons que lorsque le nombre n d'observations est suffisamment grand, $\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}$ est approximativement de loi normale multivariée de moyenne $\mathbf{0}$ et de matrice variance-covariance l'inverse de la matrice d'information de Fisher $\mathcal{I}$, i.e.

$$\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta} \approx_{\text{loi}} \mathcal{N}or(\mathbf{0}, \mathcal{I}^{-1}). \tag{9.6}$$

Afin de justifier cette affirmation, supposons qu'il existe un unique maximum à la log-vraisemblance et que $\widehat{\boldsymbol{\theta}}_n$ est proche de la vraie valeur $\boldsymbol{\theta}$ du paramètre. Notons $\boldsymbol{U}$ le vecteur gradient de la

log-vraisemblance, et $\boldsymbol{H}$ la matrice hessienne correspondante. Un développement de Taylor limité au premier ordre donne alors

$$\boldsymbol{U}(\boldsymbol{\theta}) \approx \underbrace{\boldsymbol{U}(\widehat{\boldsymbol{\theta}}_n)}_{=0 \text{ par définition de } \widehat{\boldsymbol{\theta}}_n} + \boldsymbol{H}(\widehat{\boldsymbol{\theta}}_n)\Big(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_n\Big).$$

Asymptotiquement, $\boldsymbol{H}$ est égale à sa valeur moyenne $-\mathcal{I}$ de sorte que

$$\boldsymbol{U}(\boldsymbol{\theta}) \approx -\mathcal{I}\Big(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_n\Big) = \mathcal{I}\Big(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\Big)$$

$$\Rightarrow \widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta} \approx \mathcal{I}^{-1}\boldsymbol{U}(\boldsymbol{\theta}).$$

Cette dernière relation nous permet d'obtenir la matrice de variance-covariance asymptotique de l'estimateur du maximum de vraisemblance $\widehat{\boldsymbol{\theta}}_n$ de $\boldsymbol{\theta}$, laquelle est donnée par

$$\mathbb{E}\big[(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})^t\big] \approx \mathcal{I}^{-1}\underbrace{\mathbb{E}[\boldsymbol{U}\boldsymbol{U}^t]}_{=\mathcal{I}}\mathcal{I} = \mathcal{I}^{-1}.$$

Le théorème central-limite garantit que $\boldsymbol{U}(\boldsymbol{\theta})$ est approximativement gaussien (comme somme de n variables aléatoires indépendantes), de sorte que, en grand échantillon, on a bien (**??**).

9.3.5 Test du rapport de vraisemblance

Ce test est souvent très utile pour répondre à certaines questions concernant les paramètres. Ainsi, si $\boldsymbol{\theta}$ est p-dimensionnel, et si la question posée (H_0) est qu'il y a j restrictions sur le domaine paramétrique du type $R_i(\boldsymbol{\theta}) = 0$, $i = 1, 2, \ldots, j$, où chacune des fonctions R_i possède des dérivées partielles premières continues par rapport aux composantes de $\boldsymbol{\theta}$, et si l'alternative (H_1) consiste à dire qu'il n'y a pas de telles restrictions, on calcule alors l'estimateur du maximum de vraisemblance contraint (sous H_0), noté $\widetilde{\boldsymbol{\theta}}$, et l'estimateur du maximum de vraisemblance non contraint, noté $\widehat{\boldsymbol{\theta}}$. La statistique de test

$$\mathcal{RV}_n = 2\Big\{L(\widehat{\boldsymbol{\theta}}|\boldsymbol{X}) - L(\widetilde{\boldsymbol{\theta}}|\boldsymbol{X})\Big\}$$

est approximativement de loi khi-deux à j degrés de liberté (pour n suffisamment grand). On rejettera H_0 si T est "trop grande", i.e. si $\mathcal{RV}_n > \chi^2_{j;1-\alpha}$.

9.3.6 Autres méthodes d'estimation

Comme nous le verrons dans la suite, les équations de vraisemblance n'admettent souvent pas de solution explicite. On a dès lors recours à des méthodes numériques de résolution, lesquelles procèdent par itération. Il faut donc disposer d'une valeur initiale aussi précise que possible pour les paramètres, pouvant être obtenue à l'aide de méthodes "ad hoc".

Les méthodes "ad hoc" représentent un ensemble de méthodes largement utilisées en pratique, souvent sans réel fondement théorique, et consistant à obtenir $\widehat{\boldsymbol{\theta}}$ en égalant un certain nombre de valeurs échantillons (c'est-à-dire calculées sur base de $\hat{F}_n$) à leurs analogues population (c'est-à-dire calculées sur base de $F_{\boldsymbol{\theta}}$). Le choix de ces quantités est guidé par des considérations pratiques : ce sont celles sur lesquelles l'actuaire désire mettre l'emphase car leur importance est primordiale dans le problème traité. Parmi cet ensemble de méthodes, on distingue la méthode des moments et celle des quantiles.

Méthode des moments

Supposons que $\boldsymbol{\theta}$ est p-dimensionnel. La méthode des moments consiste à égaler les p premiers moments observés à leurs homologues théoriques, i.e. $\widehat{\boldsymbol{\theta}}$ est la solution du système

$$\int_{x\in\mathbb{R}} x^j d\hat{F}_n(x) = \int_{x\in\mathbb{R}} x^j dF_{\boldsymbol{\theta}}(x), \quad j = 1, 2, \ldots, p. \tag{9.7}$$

Notez que rien ne garantit que la solution du système soit bien dans Θ.

Méthode des quantiles

La méthode des quantiles pour sa part consiste à sélectionner un certain nombre de quantiles observés, $\hat{q}_{\pi_1}$, $\hat{q}_{\pi_2}$, ..., $\hat{q}_{\pi_p}$, disons, obtenus par $\hat{F}_n^{-1}(\pi_i) = \hat{q}_{\pi_i}$, $i = 1, 2, \ldots, p$, et ensuite à prendre pour $\widehat{\boldsymbol{\theta}}$ la solution du système

$$F_{\boldsymbol{\theta}}(\hat{q}_{\pi_i}) = \pi_i, \quad i = 1, 2, \ldots, p. \tag{9.8}$$

On peut évidemment mélanger la méthode des moments et celles des quantiles en exigeant que les équations (**??**) soient satisfaites pour $j = 1, 2, \ldots, \ell$ et celles de (**??**) pour $\ell + 1, \ldots, p$.

Pour terminer, signalons que les méthodes “ad hoc” doivent être considérées avec prudence. Leur intérêt indéniable est de fournir des valeurs de départ aux algorithmes itératifs permettant d'obtenir les estimations du maximum de vraisemblance.

Méthode de type “distance minimum”

Il s'agit d'un autre type d'approche, basée sur des distances probabilistes : cette classe de méthodes consiste à choisir $\boldsymbol{\theta}$ afin de minimiser une “distance” entre $F_{\boldsymbol{\theta}}$ et $\hat{F}_n$. Epinglons quelques cas particuliers intéressants dans le cas groupé :

1. méthode du type Cramér-Von Mises :

$$\hat{\theta} = \arg\min_{\boldsymbol{\theta}\in\Theta} \sum_{j=1}^{r} w_j \left(F_{\boldsymbol{\theta}}(c_j) - \hat{F}_n(c_j) \right)^2,$$

   où les poids $w_1, w_2, \ldots, w_r$ seront sélectionnés par l'actuaire afin de mettre l'emphase sur certaines régions où la qualité de l'ajustement est primordiale (ceci revient à ajuster la fonction non-linéaire $F_{\boldsymbol{\theta}}$ au nuage de points $(c_j, \hat{F}_n(c_j))$ par une méthode des moindres carrés pondérés) ;

2. méthode du type χ^2 :

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\Theta} \sum_{j=1}^{r} \frac{\left(F_{\boldsymbol{\theta}}(c_j) - F_{\boldsymbol{\theta}}(c_{j-1}) - \hat{F}_n(c_j) + \hat{F}_n(c_{j-1}) \right)^2}{F_{\boldsymbol{\theta}}(c_j) - F_{\boldsymbol{\theta}}(c_{j-1})}.$$

La plupart du temps, $\widehat{\boldsymbol{\theta}}$ ne poura être obtenu que grâce à un algorithme itératif (du type simplexe, par exemple). Il est fortement conseillé de vérifier la solution proposée par la méthode numérique en évaluant la fonction objectif en quelques points au voisinage de celles-ci. De plus, l'actuaire devra garder à l'esprit que la solution proposée peut ne correspondre qu'à un minimum local.

9.4 Analyse des données

9.4.1 Principe

L'actuaire dispose souvent de quantités impressionnantes de données à analyser. Avant d'opter pour un modèle paramétrique $\mathcal{F}$, il est souvent utile d'analyser les données sans formuler d'hypothèses

à leur égard. Il y a plusieurs types de méthodes permettant d'analyser les données (en statistique multidimensionelle) : les méthodes dite factorielles, qui consistent à projeter le nuage de points sur un sous-espace, en perdant le moins d'information possible, et les méthodes dites de classification, qui tentent de regrouper les points.

Parmi les méthodes factorielles, trois groupes de techniques sont généralement distingués : l'analyse en composantes principales (ACP, basée sur plusieurs variables quantitatives, idéalement continues), l'analyse des correspondances binaires (ACOBI, deux variables qualitatives, représentées par un tableau de contingence) et l'analyse factorielle des correspondances multiples (AFCM, avec plus de deux variables qualitatives, et aucune quantitative). Cette section rappelle succinctement les principes d'analyse des données. Pour plus de détails, nous renvoyons le lecteur à Lebart, Morineau & Piron (2000).

9.4.2 Analyse en composantes principales (ACP)

Variables et individus

L'ACP fournit des représentations et des réductions de l'information contenue dans de volumineux tableaux de données numériques $\boldsymbol{X}$. L'élément x_{ij} de la matrice $\boldsymbol{X}$ de dimension $n \times p$ représente la valeur numérique de la $j$ème variable sur le $i$ème individu ($j = 1, 2, \ldots, p$ et $i = 1, 2, \ldots, n$). La matrice des données $\boldsymbol{X}$ a par conséquent np éléments, nombre en général très grand.

L'espace naturel du statisticien pour représenter les données est l'espace euclidien $\mathbb{R}^p$ dans lequel l'échantillon prend la forme d'un nuage de n points (chaque point correspondant à un individu). Nous l'appelons *espace des variables* et le point i est le vecteur p-dimensionnel $\boldsymbol{x}_i^{\mathrm{v}}$ défini par la $i$ème ligne de $\boldsymbol{X}$, i.e.

$$\boldsymbol{x}_i^{\mathrm{v}} = (x_{i1}, \ldots, x_{ip})^t, \quad i = 1, \ldots, n. \tag{9.9}$$

Ceci fournit un premier nuage de points, dans $\mathbb{R}^p$, dit nuage des individus ou nuage des points-lignes. Chacun des n points de ce nuage correspond à une ligne de $\boldsymbol{X}$ et donc résume les mesures des p variables sur un des n individus.

Une des originalités de l'analyse des données est de considérer en plus un deuxième espace appelé *espace des observations* dans lequel se trouve un nuage de p points représentant les vecteurs-colonnes de $\boldsymbol{X}$, i.e. les

$$\boldsymbol{x}_j^{\mathrm{o}} = (x_{1j}, \ldots, x_{nj})^t, \quad j = 1, \ldots, p. \tag{9.10}$$

Ceci fournit un second nuage de points, dans $\mathbb{R}^n$ cette fois, dit nuage des variables ou nuage des points colonnes. Chacun des p points de ce nuage correspond donc à une colonne de $\boldsymbol{X}$ et donc reprend les mesures d'une même variable effectuées sur chacun des n individus.

Les interprétations de chacun de ces espaces sont simples : dans l'espace des variables $\mathbb{R}^p$, $\boldsymbol{x}_i^{\mathrm{v}}$ représente les p caractéristiques ou variables mesurées sur le $i$ème individu et dans l'espace des observations $\mathbb{R}^n$, $\boldsymbol{x}_j^{\mathrm{o}}$ représente les valeurs prises par la $j$ème variable sur l'ensemble des n individus. Les points de l'espace des variables $\mathbb{R}^p$ représentent donc des individus et ceux de l'espace des observations $\mathbb{R}^n$ des variables.

Les proximités géométriques entre points-lignes et entre points-colonnes traduisent en fait des associations statistiques, soit entre les individus, soit entre les variables. Ainsi, une proximité de deux points-individus de $\mathbb{R}^p$ signifie que ces deux individus ont un comportement analogue par rapport à l'ensemble des p variables ; une proximité entre deux points-variables de $\mathbb{R}^n$ signifie que les n individus ont un comportement analogue en ce qui concerne ces deux variables.

Ajustement du nuage des individus dans l'espace des variables

Une façon simple de se rendre compte visuellement de la forme d'un nuage de points est de le projeter sur des droites, ou mieux sur des plans, en minimisant les déformations que la projection implique.

Ayant le nuage de n points dans $\mathbb{R}^p$, on cherche la droite passant par l'origine et déterminée par le vecteur directeur unitaire $\boldsymbol{u}$ réalisant le meilleur ajustement au sens des moindres carrés c'est-à-dire minimisant $\sum_{i=1}^n d_i^2(\boldsymbol{u})$, $d_i(\boldsymbol{u})$ représentant la distance du point i à la droite déterminée par le vecteur $\boldsymbol{u}$. Si $p_i(\boldsymbol{u})$ désigne la projection du vecteur $\boldsymbol{x}_i^{\mathrm{v}}$ sur la droite déterminée par le vecteur $\boldsymbol{u}$, du théorème de Pythagore, il résulte qu'il revient au même de

$$\max_{\boldsymbol{u}} \sum_{i=1}^n p_i^2(\boldsymbol{u}) \tag{9.11}$$

sur l'ensemble des vecteurs normés. En exprimant $p_i(\boldsymbol{u})$ à l'aide du produit scalaire, il vient $p_i(\boldsymbol{u}) = \boldsymbol{u}^t \boldsymbol{x}_i^{\mathrm{v}}$. Notant $||\cdot||$ la norme

euclidienne, (**??**) revient donc à

$$\begin{aligned}\max_{\boldsymbol{u}} \sum_{i=1}^{n} ||\boldsymbol{u}^t \boldsymbol{x}_i^{\mathrm{v}}||^2 &\Leftrightarrow \max_{\boldsymbol{u}} \sum_{i=1}^{n} \boldsymbol{u}^t \boldsymbol{x}_i^{\mathrm{v}} \boldsymbol{x}_i^{\mathrm{v}t} \boldsymbol{u} \\ &\Leftrightarrow \max_{\boldsymbol{u}} \boldsymbol{u}^t \left(\sum_{i=1}^{n} \boldsymbol{x}_i^{\mathrm{v}} \boldsymbol{x}_i^{\mathrm{v}t} \right) \boldsymbol{u} \\ &\Leftrightarrow \max_{\boldsymbol{u}} \boldsymbol{u}^t \boldsymbol{X}^t \boldsymbol{X}\ \boldsymbol{u}\end{aligned}$$

sous la contrainte $\boldsymbol{u}^t \boldsymbol{u} = 1$ Le problème à résoudre s'énonce donc comme suit :

$$\max_{\boldsymbol{u}} \boldsymbol{u}^t \boldsymbol{X}^t \boldsymbol{X}\ \boldsymbol{u} \text{ sous la contrainte } \boldsymbol{u}^t \boldsymbol{u} = 1. \tag{9.12}$$

Il s'agit d'un problème classique du calcul différentiel qui se résout par l'introduction de la fonction lagrangienne

$$\Psi(\boldsymbol{u}; \lambda) = \boldsymbol{u}^t\ \boldsymbol{X}^t \boldsymbol{X}\ \boldsymbol{u} - \lambda(\boldsymbol{u}^t \boldsymbol{u} - 1)$$

fonction de $(p+1)$ variables, les p coordonnées de $\boldsymbol{u}$ et λ. En annulant la dérivée partielle de Ψ par rapport à λ, on retrouve évidemment la contrainte $\boldsymbol{u}^t \boldsymbol{u} = 1$; l'annulation des p dérivées partielles de Ψ par rapport aux coordonnées de $\boldsymbol{u}$ conduit, après quelques manipulations de calcul matriciel, au système

$$\boldsymbol{X}^t \boldsymbol{X}\ \boldsymbol{u} - \lambda \boldsymbol{u} = 0 \Leftrightarrow \boldsymbol{X}^t \boldsymbol{X} \boldsymbol{u} = \lambda \boldsymbol{u} \tag{9.13}$$

où l'on voit que la résolution de (**??**) est identique à la recherche des valeurs propres et vecteurs propres de la matrice symétrique $\boldsymbol{X}^t \boldsymbol{X}$ de dimension $p \times p$. Pour déterminer quel couple propre convient, il suffit de prémultiplier les deux membres de la relation (**??**) par $\boldsymbol{u}^t$,

$$\boldsymbol{u}^t\ \boldsymbol{X}^t \boldsymbol{X}\ \boldsymbol{u} = \lambda \boldsymbol{u}^t \boldsymbol{u} = \lambda.$$

Revenant à (**??**), on en déduit qu'il faut prendre le couple correspondant à la plus grande valeur propre de $\boldsymbol{X}^t \boldsymbol{X}$.

L'algèbre linéaire nous apprend que toutes les valeurs propres de $\boldsymbol{X}^t \boldsymbol{X}$ sont non-négatives et que le nombre de celles-ci strictement positives est donnée par le rang r de $\boldsymbol{X}$ (avec $r \leq \min\{n, p\}$). De plus, si λ est d'ordre de multiplicité k, il existe k vecteurs propres orthogonaux liés à cette valeur ; enfin, l'ensemble des p vecteurs propres de $\boldsymbol{X}^t \boldsymbol{X}$ est orthogonal.

Désignons par $\lambda_1, \ldots, \lambda_r$ les valeurs propres non-nulles de $\boldsymbol{X}^t \boldsymbol{X}$ rangées dans l'ordre décroissant ; dans la plupart des applications,

elles sont distinctes c'est-à-dire $\lambda_1 > \lambda_2 > \ldots > \lambda_r$. Soient $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_r$ les vecteurs propres normés correspondant à ces r valeurs propres. Ainsi $\boldsymbol{u}_1$ détermine la droite cherchée, solution du problème (**??**), appelée premier axe factoriel et notée F_1.

Si l'on veut ajuster le nuage des n points de $\mathbb{R}^p$ par un hyperplan, optimum au sens des moindres carrés, le même formalisme que celui utilisé pour la droite montre que l'hyperplan cherché est celui engendré par les vecteurs propres $\boldsymbol{u}_1$ et $\boldsymbol{u}_2$ de $\boldsymbol{X}^t\boldsymbol{X}$ correspondant aux valeurs propres λ_1 et λ_2 ; il est donc engendré par les droites déterminées par $\boldsymbol{u}_1$ et $\boldsymbol{u}_2$ (c'est-à-dire les deux premiers axes factoriels F_1 et F_2).

Exemple 9.4.1. *La Figure* **??** *montre avec une bonne projection, mettant en avant la grande dispersion du nuage, et une mauvaise pour laquelle la projection n'apporte pas beaucoup d'information.*

FIGURE 9.2 – Principe de l'ACP : projection d'un nuage de points (en haut) sur un sous-espace, avec une *bonne* (en bas à gauche) et une *mauvaise* (en bas à droite) projection.

Ajustement du nuage des variables dans l'espace des observations

Plaçons-nous à présent dans l'espace des observations $\mathbb{R}^n$, où le tableau $\boldsymbol{X}$ est représenté par un nuage de points-variables dont les n coordonnées représentent les colonnes de $\boldsymbol{X}$. Par analogie, le meilleur ajustement par la méthode des moindres carrés du nuage de p points par une droite déterminée par le vecteur normé $\boldsymbol{v}$ conduit au problème

$$\begin{aligned} \max_{\boldsymbol{v}} \sum_{j=1}^{p} ||\boldsymbol{v}^t \boldsymbol{x}_j^{\circ}||^2 &\Leftrightarrow \max_{\boldsymbol{v}} \boldsymbol{v}^t \left(\sum_{j=1}^{p} \boldsymbol{x}_j^{\circ} \boldsymbol{x}_j^{\circ t} \right) \boldsymbol{v} \\ &\Leftrightarrow \max_{\boldsymbol{v}} \boldsymbol{v}^t \boldsymbol{X}\boldsymbol{X}^t \boldsymbol{v} \end{aligned}$$

sous la contrainte $\boldsymbol{v}^t\boldsymbol{v} = 1$ où $\boldsymbol{X}\boldsymbol{X}^t$ est cette fois une matrice symétrique $n \times n$.

Par analogie avec le problème (**??**), on trouve que $\boldsymbol{v}$ est solution de

$$\boldsymbol{X}\boldsymbol{X}^t\boldsymbol{v} = \mu\boldsymbol{v} \tag{9.14}$$

C'est-à-dire que $\boldsymbol{v}$ est vecteur propre normé de $\boldsymbol{X}\boldsymbol{X}^t$ correspondant, comme pour (**??**), à la plus grande valeur propre μ_1 de $\boldsymbol{X}\boldsymbol{X}^t$.

Remarque 9.4.2. *Il est facile de voir que toute valeur propre λ de $\boldsymbol{X}^t\boldsymbol{X}$ l'est pour $\boldsymbol{X}\boldsymbol{X}^t$ et réciproquement. En effet, supposons par exemple que $\boldsymbol{X}^t\boldsymbol{X}\boldsymbol{u} = \lambda\boldsymbol{u}$. En prémultipliant par $\boldsymbol{X}$, il vient $(\boldsymbol{X}\boldsymbol{X}^t)\boldsymbol{X}\boldsymbol{u} = \lambda\boldsymbol{X}\boldsymbol{u}$, ce qui prouve que $\boldsymbol{X}\boldsymbol{u}$ est vecteur propre de $\boldsymbol{X}\boldsymbol{X}^t$ correspondant à la valeur propre λ. Inversément si $\boldsymbol{X}\boldsymbol{X}^t\boldsymbol{v} = \mu\boldsymbol{v}$, en prémultipliant par $\boldsymbol{X}^t$, il vient $(\boldsymbol{X}^t\boldsymbol{X})\boldsymbol{X}^t\boldsymbol{v} = \mu\boldsymbol{X}^t\boldsymbol{v}$, d'où $\boldsymbol{X}^t\boldsymbol{v}$ est valeur propre de $\boldsymbol{X}^t\boldsymbol{X}$ correspondant à la valeur propre μ. Il résulte de cela que les n valeurs propres (dans l'ordre décroissant) de $\boldsymbol{X}\boldsymbol{X}^t$ sont $\lambda_1, \ldots, \lambda_r$ et les $n - r$ restantes étant nulles.*

Remarque 9.4.3 (Eléments supplémentaires)**.** *Les variables et individus qui servent à la construction des sous-espaces optimaux de représentation des proximités sont appelées éléments actifs. Rien n'empêche cependant l'actuaire de positionner dans ces sous-espaces des éléments (points-lignes ou points-colonnes) n'ayant pas participé à l'analyse ; ceux-ci sont appelés éléments supplémentaires ou illustratifs.*

Les éléments supplémentaires interviennent a posteriori pour enrichir l'interprétation des facteurs. Ils n'interviennent pas dans les calculs d'ajustement et ne participent donc pas à la formation des axes factoriels. Ils sont positionnés dans le nuage des individus ou dans celui des variables en calculant a posteriori leurs coordonnées sur les axes factoriels. On pourrait également vouloir représenter une variable nominale supplémentaire. Pour ce faire, on constitue autant de groupes d'individus qu'il y a de niveaux à cette variable, et on en détermine les centres de gravité. Ce sont ces points-moyens qui vont être positionnés parmi les points individus comme autant d'éléments supplémentaires.

Application à la construction d'indices

Des quantités importantes d'informations sont disponibles, notamment auprès d'instituts nationaux de statistique (INSEE en France, INS en Belgique), des services d'appui policier, des banques nationales, ainsi qu'auprès d'organismes privés de sondage ou de marketing. L'intégration de telles données dans un modèle de tarification peut parfois permettre de réelles améliorations de la précision du calcul de la prime pure.

Les quelques exemples suivants montrent comment l'incorporation de telles statistiques dans le schéma de tarification est susceptible de conduire à une meilleure évaluation du risque :

1. en assurance vol-habitation, l'information à propos du voisinage de l'habitation peut être fort utile. En effet, on peut

s'attendre à ce que le type d'habitat (villas isolées, unifamiliales entre pignons, immeubles à appartements, etc.), le profil socio-économique des habitants, etc. puisse influencer le risque couvert.

2. en assurance vol-véhicules, l'actuaire pourra se procurer les statistiques des services policiers, afin de déterminer les modèles les plus volés et les zones les plus sensibles.

3. en assurance automobile, il est utile d'exploiter les caractéristiques techniques du véhicule (disponibles auprès des constructeurs), notamment en vue d'apprécier leur caractère sportif (et donc moins facilement maniable). On pourra consulter Ingenbleek & Lemaire (1988) pour une application de l'ACP à la détermination d'un indice reflétant les performances mécaniques des automobiles.

Néanmoins, il ne suffit pas d'incorporer telles quelles les informations dans les modèles de tarification que nous développerons dans la suite de ce chapitre. En effet, les caractéristiques des véhicules ou des secteurs statistiques sont contenues dans plusieurs centaines, voire milliers, de variables, alors que les caractéristiques de l'assuré et du risque couvert sont le plus souvent résumées dans quelques dizaines de variables. Avant toute incorporation dans un modèle de tarification, il faut donc résumer les informations en quelques indices pertinents obtenus grâce, notamment, aux techniques ACP.

9.4.3 Analyse factorielle des correspondances multiples - AFCM

Analyse descriptive de grands ensembles de données qualitatives

L'analyse factorielle des correspondances multiples, AFCM en abrégé, est une technique puissante de description de grands ensembles de données qualitatives. L'AFCM peut se voir comme l'analogue de l'ACP pour des variables qualitatives. Comme pour l'ACP, les résultats apparaissent sous forme graphique (représentation dans les plans factoriels).

On considère ici N individus décrits par Q variables qualitatives à J_1, J_2, …, J_Q modalités. On note $J = \sum_{q=1}^{Q} J_q$ le nombre total des modalités pour toutes les variables.

Tableau disjonctif complet

Le tableau de départ est un tableau croisant des variables qualitatives et des individus. Chaque individu est décrit par les modalités des Q variables auxquelles il appartient. Ces données brutes se présentent donc sous la forme d'un tableau à N lignes et Q colonnes.

Remarque 9.4.4. *En pratique, on dispose souvent simultanément de variables catégorielles et de variables quantitatives. Nous supposerons dorénavant que les variables quantitatives ont été rendues catégorielles. Pour ce faire, on distingue*

1. *les variables ne prenant que quelques valeurs entières (comme le nombre d'enfants à charge, par exemple) qu'on rend catégorielle en regroupant les valeurs en classes qui ont un sens concret (par exemple, 0 enfant, 1-2 enfants et ≥ 3 enfants).*
2. *les variables continues (puissance du véhicule, par exemple) que l'on rend catégorielles en choisissant par exemple des quantiles comme limites des classes (une partition en 4 ou 5 classes est généralement suffisante en pratique).*

A chaque variable q correspond un tableau $\boldsymbol{Z}_q$, à N lignes et J_q colonnes. Ce tableau est tel que sa $i$ème ligne contient $J_q - 1$ fois la valeur 0 et une fois la valeur 1 (dans la colonne correspondant à la modalité de la variable q pour l'individu i).

Le tableau $\boldsymbol{Z}$ à N lignes et J colonnes décrivant les Q caractéristiques des N individus à l'aide d'un codage binaire s'obtient alors en juxtaposant $\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_Q$; $\boldsymbol{Z}$ est appelé tableau disjonctif complet.

Tableau de Burt

Le tableau disjonctif complet $\boldsymbol{Z}$ est ensuite transformé en un tableau de contingence multiple $\boldsymbol{B}$ (dit tableau de Burt) afin d'être utilisable pour l'AFCM. Ce tableau s'obtient grâce à $\boldsymbol{Z}$ via $\boldsymbol{B} = \boldsymbol{Z}^t\boldsymbol{Z}$; $\boldsymbol{B}$ apparaît comme une juxtaposition de tableaux de contingence. Plus précisément, ce tableau est formé par la juxtaposition de Q^2 blocs où l'on distingue

1. le bloc $\boldsymbol{Z}_q^t\boldsymbol{Z}_{q'}$ indicé par (q, q') de dimension $J_q \times J_{q'}$ qui n'est autre que la table de contingence croisant les modalités des variables q et q'.
2. le $q$ème bloc carré $\boldsymbol{Z}_q^t\boldsymbol{Z}_q$ qui apparaît comme une matrice diagonale de dimension $J_q \times J_q$ (puisque deux modalités d'une

même variable ne peuvent être choisies simultanément). Les termes diagonaux sont les effectifs des J_q modalités de la variable q.

L'AFCM est l'analyse des correspondances d'un tableau disjonctif complet ou, de manière équivalente, du tableau de Burt correspondant.

Analyse factorielle des correspondances binaires

Le tableau de Burt pouvant se voir comme une juxtaposition de tableaux de contingence croisant les variables deux à deux, il s'agit de pouvoir analyser un tableau de contingence. C'est l'objet de l'analyse des correspondances binaires, ou ACOBI.

Supposons que l'on observe sur un ensemble de N individus les valeurs de deux caractères notés I et J prenant respectivement n et p valeurs possibles. Ces caractères peuvent être quantitatifs (auquel cas, il a été procédé à des regroupements en n et en p groupes) ou qualitatifs.

Le résultat de l'observation des N individus peut se mettre sous la forme d'un tableau à n lignes et p colonnes croisant les deux caractères de façon qu'à l'intersection de la $i$ème de la $j$ème colonne se trouve le nombre total de fois N_{ij} que l'on a observé à la fois la valeur i de I et la valeur j de J. Ce tableau porte le nom de *tableau de contingence*.

En divisant tous les N_{ij} par N, on obtient les fréquences relatives f_{ij} définies par

$$f_{ij} = \frac{N_{ij}}{N}; \tag{9.15}$$

le tableau des f_{ij} est appelé le *tableau des fréquences*. On le complète en lui adjoignant une ligne et une colonne supplémentaires donnant les fréquences marginales associées aux n niveaux de I et aux p niveaux de J :

$$f_{i\bullet} = \sum_j f_{ij}, \quad i = 1, \ldots, n, \tag{9.16}$$

$$f_{\bullet j} = \sum_i f_{ij}, \quad j = 1, \ldots, p. \tag{9.17}$$

L'ACOBI revient en fait à effectuer une ACP sur la matrice $\boldsymbol{X}$ dont l'élément x_{ij} est défini comme

$$x_{ij} = \frac{f_{ij}}{f_{i\bullet}\sqrt{f_{\bullet j}}}.$$

9.5 Méthodes de scoring

9.5.1 Méthodes de classification

La régression linéaire qui fera l'objet de la Section **??** permet de prédire une variable continue à partir de variables explicatives, en faisant intervenir une combinaison linéaire de ces variables explicatives. Cette méthode ne permet toutefois pas de modéliser une variable dichotomique du type *bon-mauvais* client (indicatrice d'avoir ou non un sinistre durant l'année, par exemple).

De façon schématique, on cherche à modéliser une telle variable dichotomique (notée Y) à partir de deux variables quantitatives X_1 et X_2. En première approche, on peut considérer la variable qualitative comme une variable quantitative (prenant deux valeurs, 0 si la personne n'a pas eu de sinistre, et 1 si elle en a au moins eu un), et utiliser une régression linéaire par rapport aux deux autres. On a alors recours au modèle

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon,$$

où ε désigne le terme d'erreur.

Cette technique permet ainsi de séparer l'espace (X_1, X_2) en deux par un hyperplan (ici une droite). De façon plus formelle, on cherche à construire une fonction $y = f(x_1, x_2)$ telle que l'ensemble des points $\{(x_1, x_2) \in \mathbb{R}^2 | f(x_1, x_2) < 1/2\}$ doit correspondre aux assurés non sinistrés (i.e. ceux pour lesquels $Y = 0$) et $\{(x_1, x_2) \in \mathbb{R}^2 | f(x_1, x_2) > 1/2\}$ doit correspondre aux assurés sinistrés (i.e. ceux pour lesquels $Y = 1$). La frontière sera alors $\{(x_1, x_2) \in \mathbb{R}^2 | f(x_1, x_2) = 1/2\}$. Notons que le facteur $1/2$ est arbitraire, puisqu'il correspond au point limite naturel permettant de séparer 0 et 1. Choisir une valeur plus proche de 1 permet de mieux isoler les individus tels que $Y = 1$, mais beaucoup d'assurés sinistrés ne seront alors plus dans la bonne région de l'espace. La régression linéaire correspond au cas où f est une fonction affine, et on sépare l'espace par la droite

$$f(x_1, x_2) = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 = 1/2,$$

où les $\widehat{\beta}_i$ sont les estimateurs obtenus par moindres carrés. De façon plus générale, on peut considérer une régression dite curviligne, en régressant Y non seulement sur X_1 et X_2, mais aussi sur $X_1 \cdot X_2$, X_1^2 et X_2^2. On sépare alors par

$$f(x_1, x_2) = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \widehat{\beta}_{1,1} x_1^2 + \widehat{\beta}_{1,2} x_1 \cdot x_2 + \widehat{\beta}_{2,2} x_2^2 = 1/2,$$

La Figure **??** illustre ainsi les deux méthodes, où les points $Y = 0$ sont représentés en blanc, et les points $Y = 1$ en noir.

FIGURE 9.3 – Classification, séparation linéaire (à gauche) et curvilinéaire (à droite) entre deux nuages de points, avec ou sans sinistres.

Une autre méthode peut également être considérée, la méthode dite des *plus proches voisins*. On regarde pour un point (X_1, X_2) la valeur moyenne de Y sur un voisinage, puis on arrondit à 0 ou 1. Le résulat est visible sur la Figure **??**. Plus formellement, on choisit une distance dans l'espace des variables explicatives (X_1, X_2). Il s'agit classiquement de la distance euclidienne, c'est-à-dire que la distance séparant deux individus, i et j disons, vaut

$$d(X^i, X^j) = \sqrt{(X_1^i - X_1^j)^2 + (X_2^i - X_2^j)^2}.$$

Pour un point (x_1, x_2) de l'espace, la fonction de classification $f(x_1, x_2)$ est ici la moyenne des Y obtenus sur les k plus proches voisins du point (x_1, x_2) au sein de l'échantillon.

FIGURE 9.4 – Classification, séparation entre deux nuages de points, avec ou sans sinistres, méthode des plus proches voisins avec deux notions différentes de voisinage.

Aussi, si deux classes de clients existent au sein de la population (les *bons* et les *mauvais*), à partir des informations *a priori* sur l'individu (correspondant aux valeurs des variables tarifaires X_1 et X_2), l'individu sera classé dans la région des *bons* clients (partie inférieure gauche sur les graphiques, où prédomine une majorité de points blancs) ou dans la région des *mauvais* (partie supérieure droite).

9.5.2 Définition d'un score

Le score correspond à un classement entre les individus en fonction de leurs caractéristiques. Ces techniques sont ainsi utilisées par les banques pour l'attribution de crédits à leurs clients (le score reflétant la probabilité de ne pas pouvoir rembourser un crédit). Les assureurs peuvent également utiliser ces techniques, en modélisant la probabilité d'avoir un accident : plus faible sera cette probabilité, meilleur sera le client.

Comme nous l'avons expliqué dans le Tome I, le cycle de production est inversé en assurance. En effet, les assureurs promettent des prestations en cas de sinistre alors que les primes sont fixées *a priori*. Les assureurs doivent donc soigner leur politique d'acceptation. Une fois la proposition d'assurance complétée, la compagnie voudrait pouvoir prédire si l'assuré déclarera ou non des sinistres, compte tenu de ses caractéristiques observables. Plusieurs méthodologies sont envisageables : inspection des dossiers, entretiens, recherches d'informations complémentaires, etc. Cependant, en raison des coûts, du nombre de dossiers qu'il faudrait étudier mais aussi de la dose d'arbitraire introduite dans l'examen au cas par cas, il est de plus en plus fait recours aux méthodes statistiques de classification qui sont moins coûteuses, plus rapides et plus systématiques pour les risques de masse.

Remarque 9.5.1. *Même si la suite de cette section sera consacrée au problème de l'acceptation, les méthodes qui y sont décrites peuvent être appliquées à la résolution de bien d'autres problèmes qui se posent au sein des compagnies d'assurance, au nombre desquels la détection des assurés du portefeuille qui risquent de résilier leur contrat à la prochaine échéance, par exemple. Ceci se fait en analysant le fichier contenant les polices du portefeuille avec comme variable dépendante l'indicatrice de l'événement "la police i quitte le portefeuille en fin de période". On veillera à faire apparaître dans ce fichier des informations cruciales, telles que le nombre d'années que la police est en portefeuille, la sinistralité, le nombre de modifications apportées au contrat, etc.*

Définissons la variable indicatrice Y, prenant les valeurs 0 ou 1 suivant que le risque est bon ou non (typiquement, le bon risque est l'assuré ne déclarant aucun sinistre sur la période). Comme pour les méthodes de classification, la règle de décision est basée sur un ensemble de variables explicatives $\boldsymbol{X}$: on considère que l'assuré est un *bon* risque si $\boldsymbol{X} \in \mathcal{A}$, un *mauvais* risque sinon, où $\mathcal{A}$ désigne un domaine d'acceptabilité. Formellement, un classement des candidats preneurs en acceptation et rejet revient à choisir une partition de $\mathbb{R}^p$ en une zone d'acceptation $\mathcal{A}$ et une zone de rejet $\overline{\mathcal{A}} = \mathbb{R}^p \setminus \mathcal{A}$. Ainsi, pour un assuré dont les caractéristiques sont résumées dans le vecteur $\boldsymbol{x}$,

$$\begin{aligned} \boldsymbol{x} \in \mathcal{A} &\Rightarrow \text{acceptation} \\ \boldsymbol{x} \in \overline{\mathcal{A}} &\Rightarrow \text{refus.} \end{aligned}$$

La partition la plus simple est celle engendrée par un hyperplan, mais on peut imaginer des ensembles beaucoup plus compliqués. Il y a donc autant de classificateurs que de partitions de $\mathbb{R}^p$ en $\mathcal{A}$ et $\overline{\mathcal{A}}$.

Comme les variables qualitatives peuvent toujours être remplacées par des codes numériques comme décrit plus haut, $\boldsymbol{X}$ est à valeurs dans $\mathbb{R}^p$. Nous commencerons par supposer que $\boldsymbol{X}$ est continu et nous verrons par la suite comment gérer les cas où certaines composantes de $\boldsymbol{X}$ sont discrètes ou catégorielles.

9.5.3 Principe du scoring

Le score dit canonique est le score qui classe les individus en fonction de la probabilité d'être ou non un bon client. C'est donc la fonction qui associe aux variables tarifaire $\boldsymbol{X}$ le score S^* défini par

$$S^*(\boldsymbol{x}) = \Pr[Y = 1|\boldsymbol{X} = \boldsymbol{x}].$$

On peut alors définir deux classes, les *bons* et les *mauvais* risques, les bons correspondant aux individus ayant obtenu un score inférieur à une seuil s - fixé *a priori*[3]. Notons qu'à s fixé $\Pr[S \leq s]$ est la proportion des individus retenus, et $\Pr[Y = 0|S \leq s]$ désigne la proportion des bons parmi les individus retenus.

De façon générale, on notera $p_1 = \Pr[Y = 1]$ et $p_0 = 1 - p_1$ les probabilités marginales inconnues (mais qui peuvent être estimées à partir de l'historique). Supposons que les covariables soient continues, et de densité $f(\boldsymbol{x})$. On notera f_0 et f_1 les densités conditionnelles

$$f_j(\boldsymbol{x}) = f(\boldsymbol{x}|Y = j), \text{ où } j = 0, 1.$$

Le pouvoir de discrimination (ou de segmentation) sera alors d'autant plus fort que ces densités seront différentes. On notera enfin $p_j(\boldsymbol{x})$ les probabilités de $Y = j$ sachant $\boldsymbol{x}$,

$$p_j(\boldsymbol{x}) = \Pr[Y = j|\boldsymbol{X} = \boldsymbol{x}], \text{ où } j = 0, 1.$$

Le théorème de Bayes (rappelé à la Section 2.2.8) nous permet de lier ces quantités par

$$p_j(\boldsymbol{x}) = \frac{p_j f_j(\boldsymbol{x})}{f(\boldsymbol{x})}, \text{ avec } f(\boldsymbol{x}) = p_0 f_0(\boldsymbol{x}) + p_1 f_1(\boldsymbol{x}).$$

3. Notez que deux notions différentes sont utilisées ici : il y a d'une part les bons et les mauvais risques, correspondant respectivement à $Y = 0$ et à $Y = 1$, et d'autre part les individus considérés comme tels par l'assureur, correspondant respectivement à $S \leq s$ et $S > s$. Ces deux notions se distingueront par le contexte.

9.5.4 Classification optimale et choix du seuil

Un certain nombre de méthodes permettent de fixer le seuil s. Pour cela, supposons que l'assureur puisse refuser d'assurer les mauvais clients. Soit g le gain de l'assureur en cas de bonne décision : accepter d'assurer un bon client (ce qui arrive avec probabilité $\Pr[\boldsymbol{X} \in \mathcal{A}, Y = 0]$). Soit c_0 le coût associé à l'assurance d'un mauvais risque (avec probabilité $\Pr[\boldsymbol{X} \in \mathcal{A}, Y = 1]$) ; et soit c_1 le manque à gagner en refusant d'assurer un bon client (avec probabilité $\Pr[\boldsymbol{X} \notin \mathcal{A}, Y = 0]$).

Il est essentiel de tenir compte des conséquences très différentes des mauvaises classifications pour la compagnie. En effet, ranger un mauvais assuré parmi les bons revient à accepter de couvrir un assuré qui causera un sinistre (voire plusieurs), ce qui expose la compagnie à de lourdes pertes. Au contraire, refuser un bon risque coûte peu à la compagnie (mais pourrait avoir des conséquences commerciales désastreuses si cela se répétait trop souvent). Il est essentiel d'introduire cette distinction dans le modèle.

En définissant la rentabilité de l'assureur comme

$$\begin{aligned} R(\mathcal{A}) &= g \Pr[\boldsymbol{X} \in \mathcal{A}, Y = 0] - c_0 \Pr[\boldsymbol{X} \in \mathcal{A}, Y = 1] \\ &\quad -c_1 \Pr[\boldsymbol{X} \notin \mathcal{A}, Y = 0], \end{aligned}$$

la région d'acceptation optimale est alors donnée par

$$\mathcal{A}^* = \underset{\mathcal{A}}{\operatorname{argmax}} \{R(\mathcal{A})\}.$$

On peut noter que la rentabilité de l'assureur peut se réécrire

$$R(\mathcal{A}) = \int_{\boldsymbol{x} \in \mathcal{A}} \Big((g + c_1) p_0 f_0(\boldsymbol{x}) - c_0 p_1 f_1(\boldsymbol{x}) \Big) d\boldsymbol{x} - c_1 p_0.$$

Le domaine d'acceptation optimal $\mathcal{A}^*$ est celui pour lequel l'intégrand est toujours positif, à savoir

$$\mathcal{A}^* = \left\{ \boldsymbol{x} \middle| \frac{f_1(\boldsymbol{x})}{f_0(\boldsymbol{x})} \leq \frac{(g + c_1) p_0}{c_0 p_1} \right\},$$

ou encore, un utilisant la formule de Bayes avec $p_j(\boldsymbol{x}) = p_j f_j(\boldsymbol{x}) / f(\boldsymbol{x})$,

$$\mathcal{A}^* = \left\{ \boldsymbol{x} \middle| p_1(\boldsymbol{x}) \leq \frac{g + c_1}{g + c_0 + c_1} \right\}.$$

On peut alors sélectionner les assurés à l'aide soit des lois conditionnelles de $\boldsymbol{X}$ sachant Y (c'est-à-dire f_0 et f_1), soit des lois conditionnelles de Y sachant $\boldsymbol{X}$ (c'est-à-dire $p_0(\boldsymbol{X})$ et $p_1(\boldsymbol{X})$). Notons toutefois que si ces deux approches correspondent à deux formalisations équivalentes d'un point de vue mathématique, elles témoignent de deux optiques sensiblement différentes :

- la loi de $\boldsymbol{X}|Y$ est utilisée en analyse discriminante (la question est de savoir si un individu, dont on sait s'il est un bon ou un mauvais risque sera classé parmi les *bons* ou les *mauvais* clients) ;
- la loi de $Y|\boldsymbol{X}$ est utilisée dans une problématique prévisionnelle (à quelle population - des *bons* ou des *mauvais* clients - un individu dont on connaît les caractéristiques $\boldsymbol{X}$ a-t-il le plus de chance d'appartenir ?).

Remarque 9.5.2. *On peut bien entendu imaginer que les coûts de mauvaises classification c_0 et c_1, de même que le gain g, dépendent des caractéristiques de l'assuré, i.e. $c_0 = c_0(\boldsymbol{x})$, $c_1 = c_1(\boldsymbol{x})$ et $g = g(\boldsymbol{x})$. Le raisonnement est en tout point similaire à celui suivi ci-dessus.*

9.5.5 La pratique de la construction d'un score

Cinq grandes étapes sont fondamentales lors de la construction d'un score.

1. le choix du critère dichotomique à modéliser : ne pas avoir de sinistre dans l'année, ne pas avoir de sinistre grave dans l'année, ne pas avoir de sinistre responsable dans l'année, ...
2. le choix de la population : la difficulté vient du fait que la population d'assurés dans le portefeuille peut être sensiblement différente de la population qui demande à être assurée (de par la sélection à l'aide d'un score, précisément). Une partie de la population d'assurés est utilisée pour calculer le score, l'autre servant à tester les performances.
3. le choix des covariables $\boldsymbol{X}$.
4. l'estimation du modèle : schématiquement, il est possible d'utiliser des modèles logit ou probit, et d'estimer les paramètres par maximum de vraisemblance.
5. analyse des performances : un certain nombre de tests et de critères permettent de juger de la qualité de discrimination du score (courbes de performance, de sélection, de discrimination, par exemple).

Désignons par $\mathcal{H}$ les données historiques qui serviront à la constuction du classificateur (c'est-à-dire les observations $(\boldsymbol{x}, Y)$ réalisées par la compagnie dans le passé sur un grand nombre d'individus). Cet ensemble sera partitionné en deux sous-ensembles (le plus souvent de façon aléatoire) qui serviront respectivement à estimer les paramètres et à évaluer le classificateur, à savoir

le "training set" Il s'agit du sous-ensemble de $\mathcal{H}$ utilisé pour déterminer les paramètres. Le plus souvent, le training set est de la forme

$$\{(\boldsymbol{x}_{0;k}, 0)\ k = 1, ..., n_0, (\boldsymbol{x}_{1;k}, 1)\ k = 1, ..., n_1\} \tag{9.18}$$

où les $\boldsymbol{x}_{0,j}$ en nombre n_0 correspondent à des assurés répertoriés parmi les bons, et les $\boldsymbol{x}_{1,j}$ en nombre n_1 correspondent à des assurés répertoriés parmi les mauvais.

le "test set" Il s'agit du sous-ensemble de $\mathcal{H}$ qui sert à évaluer les performances du classificateur.

Si on retient un modèle paramétrique, on estime les densités des caractéristiques observables pour les bons et les mauvais assurés par

$$\widehat{f_1(\boldsymbol{x})} = f_1(\boldsymbol{x}; \widehat{\boldsymbol{\theta}}) \text{ et } \widehat{f_0(\boldsymbol{x})} = f_0(\boldsymbol{x}; \widehat{\boldsymbol{\theta}}).$$

où $\widehat{\boldsymbol{\theta}}$ est l'estimateur du maximum de vraisemblance de $\boldsymbol{\theta}$. La vraisemblance est donnée par

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{j=1}^{n_0} \Big(f_0(\boldsymbol{x}_{0;j}, \boldsymbol{\theta}) p_0 \Big) \prod_{j=1}^{n_1} \Big(f_1(\boldsymbol{x}_{1;j}, \boldsymbol{\theta}) p_1 \Big). \tag{9.19}$$

Afin d'estimer le vecteur de paramètres $(\boldsymbol{\theta}, p_0, , p_1)$, écrivons la log-vraisemblance :

$$L(\boldsymbol{\theta}, \boldsymbol{p}) = \sum_{j=1}^{n_0} \ln f_0(\boldsymbol{x}_{0,j}; \boldsymbol{\theta}) + \sum_{j=1}^{n_1} \ln f_1(\boldsymbol{x}_{1,j}; \boldsymbol{\theta}) + n_0 \ln p_0 + n_1 \ln p_1. \tag{9.20}$$

On maximise d'abord $n_0 \ln p_0 + n_1 \ln p_1$ pour obtenir

$$\hat{p}_0 = \frac{n_0}{n_0 + n_1} \text{ et } \hat{p}_1 = 1 - \hat{p}_0 = \frac{n_1}{n_0 + n_1}.$$

Il s'agit bien de l'estimateur naturel de p_0, à savoir la proportion des bons clients dans le training set. Il faut bien se rendre compte qu'on postule le fait que la population cible pour les nouvelles affaires est semblable aux individus de $\mathcal{H}$.

Il suffit ensuite de maximiser

$$\sum_{j=1}^{n_0} \ln f_0(\boldsymbol{x}_{0,j};\boldsymbol{\theta}) + \sum_{j=1}^{n_1} \ln f_1(\boldsymbol{x}_{1,j};\boldsymbol{\theta}) \tag{9.21}$$

sur $\boldsymbol{\theta}$ pour obtenir l'estimateur du maximum de vraisemblance de $\boldsymbol{\theta}$.

9.5.6 Analyse discriminante

L'objectif est ici de décrire l'appartenance d'un individu à une classe prédéfinie en fonction de ses modalités. On s'intéresse alors à la loi de $\boldsymbol{X}|Y$ décrite par les fonctions de densité f_0 et f_1. Rappelons que la fonction de score canonique S^* conduit à une règle de classification basée sur le rapport f_1/f_0.

Exemple 9.5.3. *Supposons que, conditionnellement à $Y = j$, $\boldsymbol{X} \sim \mathcal{N}or_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$. Notez que la matrice variance-covariance ne dépend pas de la valeur j de Y. Le rapport des densités est alors une fonction croissante de*

$$(\boldsymbol{X}-\boldsymbol{\mu}_0)^t \boldsymbol{\Sigma}^{-1} (\boldsymbol{X}-\boldsymbol{\mu}_0) - (\boldsymbol{X}-\boldsymbol{\mu}_1)^t \boldsymbol{\Sigma}^{-1} (\boldsymbol{X}-\boldsymbol{\mu}_1).$$

On définira le score

$$S(\boldsymbol{X}) = \boldsymbol{X}^t \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0).$$

Si les matrices de variance-covariance ne sont pas égales, le score devient

$$S(\boldsymbol{X}) = (\boldsymbol{X}-\boldsymbol{\mu}_0)^t \boldsymbol{\Sigma}_0^{-1} (\boldsymbol{X}-\boldsymbol{\mu}_0) - (\boldsymbol{X}-\boldsymbol{\mu}_1)^t \boldsymbol{\Sigma}_1^{-1} (\boldsymbol{X}-\boldsymbol{\mu}_1),$$

où $\boldsymbol{\Sigma}_0$ est la matrice variance-covariance de $\boldsymbol{X}$ si $Y = 0$, et $\boldsymbol{\Sigma}_1$ celle de $\boldsymbol{X}$ si $Y = 1$. On n'a alors plus un critère linéaire (comme dans le cas précédent), mais quadratique.

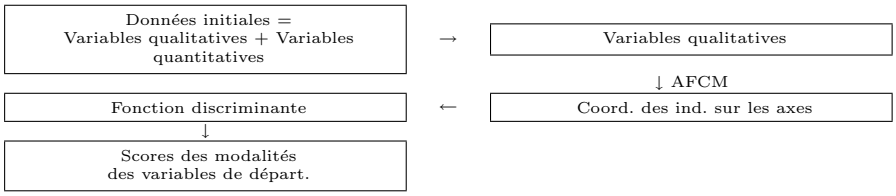
9.5.7 La méthode DISQUAL

Les méthodes d'analyse discriminante que nous venons de présenter semblent particulièrement bien adaptées à la résolution du problème posé par l'acceptation. Il y a pourtant une difficulté de

taille qu'il faudra surmonter. En pratique, l'information dont dispose l'actuaire à propos des futurs assurés est majoritairement composée de variables qualitatives (binaires, comme le sexe, ou présentant plusieurs modalités, comme la catégorie socio-professionnelle), et de variables entières ne prenant que quelques valeurs distinctes (comme le nombre d'enfants à charge, ou le nombre de véhicules du ménage, par exemple).

La densité jointe de ces variables sera donc fort loin de ressembler à celle d'une loi normale multivariée. Afin de se rapprocher des hypothèses de validité de l'analyse discriminante, on peut tout d'abord transformer les variables qualitatives en variables continues et non-corrélées en réalisant une AFCM sur l'ensemble des variables qualitatives. On travaillera ensuite avec les coordonnées des individus sur les axes factoriels. L'AFCM permet de substituer aux caractères qualitatifs de départ, qui ne se prêtent pas toujours bien à un scoring, des variables continues. Ensuite, une analyse discriminante est effectuée. Il suffit alors de repasser aux variables de départ pour en déduire un scoring.

Cette stratégie peut encore se représenter schématiquement comme suit :



9.5.8 Le modèle PROBIT

Le fait que Y prenne ses valeurs dans $\{0,1\}$ rend toute modélisation linéaire esquissée à la Section **??** inappropriée. L'utilisation d'une variable latente est, comme nous allons le voir, beaucoup plus pertinente.

On suppose ici qu'il existe une variable latente Y^* telle que

$$Y^* = \boldsymbol{X}^t\boldsymbol{\beta} + \varepsilon \text{ et } Y = \mathbb{I}[Y^* \geq 0],$$

où $\mathbb{I}[A]$ est l'indicatrice de l'événement A, valant 1 si cet événement s'est réamlisé et 0 sinon, et où $\varepsilon \sim \mathcal{N}or(0,1)$. Aussi, le score associé

à ce modèle est-il donné par

$$\begin{aligned} S(\boldsymbol{x}) &= \Pr[Y=1|\boldsymbol{X}=\boldsymbol{x}] = \Pr[Y^*\geq 0|\boldsymbol{X}=\boldsymbol{x}] \\ &= \Pr[\boldsymbol{x}^t\beta+\varepsilon\geq 0] = 1-\Phi(-\boldsymbol{x}^t\boldsymbol{\beta}) = \Phi(\boldsymbol{x}^t\boldsymbol{\beta}), \end{aligned}$$

où Φ désigne la fonction de répartition de loi $\mathcal{N}or(0,1)$.

Pour ce modèle, l'estimation des paramètres $\boldsymbol{\beta}$ se fait par maximum de vraisemblance. Disposant d'observations $(y_i, \boldsymbol{x}_i)$, $i = 1, 2, \ldots, n$, effectuées sur n individus, ceci revient à maximiser

$$\mathcal{L} = \prod_{i=1}^{n} \Big(\Phi(\boldsymbol{x}_i^t\boldsymbol{\beta})\Big)^{1-y_i} \Big(1-\Phi(\boldsymbol{x}_i^t\boldsymbol{\beta})\Big)^{y_i}.$$

Il suffit en fait de résoudre le système des équations de vraisemblance obtenu en annulant le gradient de la log-vraisemblance, ce qui donne

$$\frac{\partial}{\partial\boldsymbol{\beta}} \ln\mathcal{L} = \sum_{i=1}^{n} \left((1-y_i)\frac{\boldsymbol{x}_i\phi(\boldsymbol{x}_i^t\boldsymbol{\beta})}{\Phi(\boldsymbol{x}_i^t\boldsymbol{\beta})} - y_i\frac{\boldsymbol{x}_i\phi(\boldsymbol{x}_i^t\boldsymbol{\beta})}{1-\Phi(\boldsymbol{x}_i^t\boldsymbol{\beta})} \right) = \mathbf{0},$$

où ϕ est la densité associée à la loi $\mathcal{N}or(0,1)$. Ce système n'admet pas de solution explicite. Nous reviendrons plus longuement dans la Section **??** sur ce modèle de régression non linéaire.

9.5.9 Le modèle LOGIT

L'idée est là aussi de recourir à une variable latente Y^*, mais on suppose ici que ε obéit à la loi logistique, de fonction de répartition

$$F(x) = \frac{1}{1+\exp(-x)}, \quad x\in\mathbb{R}.$$

Dans ce cas, le score vaut

$$S^*(\boldsymbol{x}) = \Pr[Y=1|\boldsymbol{X}=\boldsymbol{x}] = F\left(\boldsymbol{x}^t\boldsymbol{\beta}\right) = \frac{1}{1+\exp(-\boldsymbol{x}^t\boldsymbol{\beta})}.$$

De plus,

$$\Pr[Y=0|\boldsymbol{X}=\boldsymbol{x}] = F\left(-\boldsymbol{x}^t\boldsymbol{\beta}\right) = \frac{1}{1+\exp(\boldsymbol{x}^t\boldsymbol{\beta})}.$$

Notons que le ratio $\Pr[Y=1|\boldsymbol{X}=\boldsymbol{x}] / \Pr[Y=0|\boldsymbol{X}=\boldsymbol{x}]$ s'appelle la cote (*odd ratio*, en anglais). Pour les turfistes, une cote à 5 contre 1 signifie que la probabilité de perdre est 5 fois plus grande que celle de gagner. Ainsi, lorsqu'un cheval est coté 5 contre 1, cela signifie que 5 parieurs l'ont joué perdant contre 1 gagnant. En d'autres termes, il y a 5 parieurs sur 6 qui ont joué le cheval perdant, d'où une cote de $\frac{5/6}{1/6} = 5$.

Nous reviendrons plus longuement à la Section **??** sur les détails de ce modèle de régression non linéaire.

9.5.10 Dualité des approches

Les approches par analyse discriminante et par régression logistique sont en fait des approches duales ; voyez Celeux & Nakache (1994) et Gouriéroux (1992,1999). L'analyse discriminante est basée sur la spécification des lois conditionnelles de $\boldsymbol{X}|Y$, i.e. f_0 et f_1. Ainsi, le score canonique s'écrit

$$S^*(\boldsymbol{x}) = \Pr[Y = 1|\boldsymbol{X} = \boldsymbol{x}] = \frac{p_1 f_1(\boldsymbol{x})}{p_0 f_0(\boldsymbol{x}) + p_1 f_1(\boldsymbol{x})}.$$

Sous les hypothèses de l'Exemple **??** (lois conditionelles gaussiennes de même matrice variance-covariance), ce score canonique s'écrit après simplifications

$$S^*(\boldsymbol{x}) = \frac{1}{1 + \exp\big(-\Delta(\boldsymbol{x})\big)}$$

où

$$\begin{aligned}\Delta(\boldsymbol{x}) &= \boldsymbol{X}^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) - \ln\left(\frac{1-p_1}{p_1}\right) \\ &\quad -\frac{1}{2}\Big(\boldsymbol{\mu}_0^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1\Big).\end{aligned}$$

Ceci correspond à un modèle logit avec un terme constant, i.e. $\beta_0 + \boldsymbol{X}^t\boldsymbol{\beta}$ où $\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$ et

$$\beta_0 = -\ln\left(\frac{1-p_1}{p_1}\right) - \frac{1}{2}\Big(\boldsymbol{\mu}_0^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1\Big).$$

Autrement dit, l'analyse discriminante linéaire décrite dans l'Exemple **??** est un cas particulier du modèle LOGIT. De façon équivalente, l'analyse discriminante quadratique obtenue avec des matrices de variance-covariance différentes (deuxième partie de l'Exemple **??**) apparaît comme un cas particulier du modèle LOGIT, lorsque les variables explicatives comprennent des transformations quadratiques des composantes de $\boldsymbol{X}$.

9.5.11 Les courbes de performance et de sélection

Supposons que le score S soit utilisé pour discriminer en deux sous-populations, de bons et de mauvais clients, à l'aide d'un seuil s. L'idée de Gouriéroux (1992) est alors de représenter la performance du score par la courbe de performance

$$\mathcal{P} = \big\{(x(s), y(s))\big|s \in [0,1]\big\},$$

où

$$x(s) = \Pr[S \leq s] \text{ et } y(s) = \frac{\Pr[Y = 0|S \leq s]}{\Pr[Y = 0]},$$

dont l'équation explicite est $y = \mathcal{P}(x)$. On définit la courbe de sélection par

$$\mathcal{S} = \big\{(x(s), y(s))\big|s \in [0, 1]\big\},$$

où

$$x(s) = \Pr[S \leq s] \text{ et } y(s) = \Pr[S \leq s|Y = 0],$$

dont l'équation explicite est $y = \mathcal{S}(x)$. La courbe de performance est alors nécessairement croissante, alors que la courbe de sélection d'un score canonique est toujours croissante et convexe. Notons également que les deux courbes sont liées par la relation $\mathcal{S}(x) = x\mathcal{P}(x)$. Ces deux courbes sont représentées à la Figure **??**.

Les courbes de performance sont invariantes par transformation strictement croissante du score : soit h une transformation strictement croissante,

$$x_h(s) = \Pr[h(S) \leq s] = \Pr[S \leq h^{-1}(s)] = x(h^{-1}(s)),$$

et

$$y_h(s) = \frac{\Pr[Y = 0|h(S) \leq s]}{\Pr[Y = 0]} = y(h^{-1}(s)).$$

Autrement dit, ces courbes ne tiennent pas compte de la valeur du score, mais uniquement de l'ordre qu'il établit.

FIGURE 9.5 – Courbes de sélection et de performance pour un score S.

9.5.12 Propriétés (souhaitables) d'un score

Quel que soit le score S (non nécessairement canonique), il est souhaitable qu'il soit fortement dépendant de Y. En particulier, il est possible de montrer que les variables aléatoires Y et S sont associées (cette notion a été présentée dans la Section 8.5.3 du Tome 1) si, et seulement si, la coube de performance $\mathcal{P}$ est en dessous de la droite $y = 1$, ou de façon équivalente, si, et seulement si, la courbe de sélection $\mathcal{S}$ est en-dessous de la première bissectrice.

De plus, si la courbe de sélection est croissante et convexe, alors la courbe de sélection peut être vue comme la courbe de Lorenz associée à $\Pr[Y = 0|S]$.

9.5.13 Comparaison de scores

Les courbes de performance peuvent être utilisées pour comparer des scores entre eux. En particulier, un score S_1 est dit plus performant qu'un score S_2 si, et seulement si, sa courbe de performance est en dessous de celle de S_2.

On peut également définir la courbe dite de discrimination du score : si on note $G_0(s) = \Pr[S \leq s|Y = 0]$ et $G_1(s) = \Pr[S \leq s|Y = 1]$, alors la courbe de discrimination du score est l'application $[0,1] \to [0,1]$ définie par

$$\mathcal{D}(x) = G_1 \circ G_0^{-1}(x), \quad x \in [0,1].$$

Notons que cette application est croissante, et est invariante par transformation strictement croissante du score. Notons que la concavité de cette application est équivalente à avoir $\Pr[Y = 1|S = s]$ croissante en s.

Le terme de *discrimination* de cette courbe vient de la propriété suivante : si le score n'est pas discriminant, et que donc $G_0 = G_1$ (Y et S sont indépendantes), alors la courbe $\mathcal{D}$ est confondue avec la première bissectrice. En revanche, pour une population partitionnée en deux sous-ensembles, où G_0 et G_1 seraient alors concentrées en deux points, la courbe de discrimination serait alors la courbe $y = 0$ sur $[0,1[$. Entre ces deux cas limites, il est possible de considérer le préordre suivant : le score S_1 est plus discriminant que le score S_2 si, et seulement si, sa courbe de discrimination est en dessous de celle de S_2.

9.6 Modèle linéaire

9.6.1 Définition

Les modèles linéaires généralisés étendant le modèle linéaire gaussien, il semble naturel de commencer par rappeler brièvement les principaux résultats relatifs à l'approche de régression linéaire classique. Nous ne pouvons que conseiller au lecteur soucieux d'approfondir ses connaissances en la matière l'excellent ouvrage de Dodge & Rousson (2004).

Pendant longtemps, les modèles utilisés pour expliquer les variations des variables continues $Y_1, Y_2, \ldots, Y_n$ en présence de variables explicatives résumées dans les vecteurs $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$, ont pris la

forme

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i \text{ avec } \epsilon_i \sim \mathcal{N}or(0, \sigma^2).$$

De manière équivalente, on a

$$Y_i \sim \mathcal{N}or\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}, \sigma^2\right), \quad i = 1, 2, \ldots, n.$$

Les observations Y_i sont donc supposées de loi normale de moyenne $\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$, une fonction affine des variables explicatives, et de variance constante σ^2. La combinaison linéaire des variables explicatives $\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$ qui donne $\mathbb{E}[Y_i]$ est appelée score (ou prédicteur linéaire) et sera notée par la suite η_i.

Même si le modèle linéaire impose de sérieuses limitations, et que son réalisme est discutable dans beaucoup de problèmes auxquels l'actuaire est confronté, il n'en demeure pas moins fort important, car la plupart des modèles plus réalistes (dont les modèles linéaires généralisés qui feront l'objet de la section suivante) empruntent bon nombre de techniques à celui-ci.

9.6.2 Formalisme matriciel

Le formalisme matriciel est fort utile pour analyser le modèle linéaire général. On voit facilement que le modèle de régression linéaire peut se réécrire vectoriellement comme

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{9.22}$$

où $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^t$ est un vecteur $n \times 1$ reprenant les variables à expliquer, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^t$ est un vecteur $(p+1) \times 1$ de paramètres,

$$\boldsymbol{X} = \begin{pmatrix} 1 & \boldsymbol{x}_1^t \\ 1 & \boldsymbol{x}_2^t \\ \vdots & \vdots \\ 1 & \boldsymbol{x}_n^t \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix}$$

est une matrice $n \times (p+1)$ reprenant les variables explicatives, et $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)^t \sim \mathcal{N}or_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$ est un vecteur $n \times 1$ reprenant les erreurs. La matrice $\boldsymbol{X}$ est supposée être de rang $p+1$, i.e. la matrice carrée $\boldsymbol{X}^t\boldsymbol{X}$ de dimension $(p+1) \times (p+1)$ est supposée être inversible.

Avec ces notation, la valeur observée du vecteur aléatoire $\boldsymbol{Y}$ est la somme d'une composante déterministe $\boldsymbol{X}\boldsymbol{\beta}$ et d'une composante aléatoire $\boldsymbol{\epsilon}$ qui modélise le bruit. La composante déterministe du vecteur $\boldsymbol{Y}$ représent les observations qui auraient été faites en l'absence de bruit. L'hypothèse $\mathbb{E}[\boldsymbol{\epsilon}] = \mathbf{0}$ signifie que la composante déterministe du vecteur $\boldsymbol{Y}$ est sa moyenne.

9.6.3 Estimation des paramètres

Supposons disposer des réalisations $y_1, y_2, \ldots, y_n$ des variables $Y_1, Y_2, \ldots, Y_n$. La fonction de vraisemblance relative aux observations $y_1, y_2, \ldots, y_n$ est

$$\begin{aligned}\mathcal{L}(\boldsymbol{\beta}, \sigma|\boldsymbol{y}) &= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \prod_{i=1}^n \exp\left(-\frac{1}{2\sigma^2}(y_i - \boldsymbol{x}_i^t\boldsymbol{\beta})^2\right) \\ &= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^t(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right).\end{aligned}$$

Comme expliqué plus haut, l'estimateur du maximum de vraisemblance de $\boldsymbol{\beta}$ est la valeur de $\boldsymbol{\beta}$ qui maximise $\mathcal{L}(\boldsymbol{\beta}, \sigma|\boldsymbol{y})$. Il s'agit donc intuitivement de la valeur de $\boldsymbol{\beta}$ qui rend les observations $y_1, y_2, \ldots, y_n$ les plus plausibles dans le modèle (**??**). Cet estimateur s'obtient comme suit.

Proposition 9.6.1. *L'estimateur du maximum de vraisemblance $\widehat{\boldsymbol{\beta}}$ de $\boldsymbol{\beta}$ est solution des équations normales*

$$\boldsymbol{X}^t\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}^t\boldsymbol{Y} = \mathbf{0} \Leftrightarrow \boldsymbol{X}^t\Big(\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{Y}\Big) = \mathbf{0}.$$

Puisque la matrice $\boldsymbol{X}^t\boldsymbol{X}$ a été supposée inversible, le système des équations normales a une solution unique donnée par

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^t\boldsymbol{X}\right)^{-1}\boldsymbol{X}^t\boldsymbol{Y} = \left(\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^t\right)^{-1}\sum_{i=1}^n \boldsymbol{x}_i y_i. \tag{9.23}$$

qui définit l'estimateur du maximum de vraisemblance de $\boldsymbol{\beta}$.

Démonstration. La log-vraisemblance vaut

$$L(\boldsymbol{\beta}, \sigma|\boldsymbol{y}) = -\frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^t(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) - \frac{n}{2}\ln(2\pi).$$

Or,

$$\sup_{(\boldsymbol{\beta},\sigma^2)} L(\boldsymbol{\beta}, \sigma|\boldsymbol{y}) = \sup_{\sigma^2}\sup_{\boldsymbol{\beta}} L(\boldsymbol{\beta}, \sigma|\boldsymbol{y}).$$

Donc, quelle que soit la valeur de σ^2, maximiser $L(\boldsymbol{\beta}, \sigma|\boldsymbol{y})$ par rapport à $\boldsymbol{\beta}$ revient à minimiser

$$S_2(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^t(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}).$$

Pour que $\widehat{\boldsymbol{\beta}}$ minimise S_2 il faut que ce soit un point stationnaire de cette expression. Il est donc obtenu en dérivant S_2 par rapport à $\boldsymbol{\beta}$ et en identifiant le gradient à $\mathbf{0}$. Comme

$$S_2(\boldsymbol{\beta}) = \boldsymbol{y}^t\boldsymbol{y} - 2\boldsymbol{y}^t\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}^t\boldsymbol{X}^t\boldsymbol{X}\boldsymbol{\beta}.$$

Il vient alors

$$\frac{\partial S_2(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2\boldsymbol{X}^t\boldsymbol{Y} + 2\boldsymbol{X}^t\boldsymbol{X}\boldsymbol{\beta},$$

d'où le système des équations normales. Remarquons aussi que toute solution des équations normales correspond bien à un minimum de la fonction S_2 car la matrice hessienne est la matrice définie positive $2\boldsymbol{X}^t\boldsymbol{X}$. □

Si la matrice $\boldsymbol{X}^t\boldsymbol{X}$ n'est pas inversible, le système des équations normales peut avoir plus d'une solution et on utilise alors la notion de matrice inverse généralisée (notons toutefois que l'estimateur de la moyenne $\boldsymbol{X}\widehat{\boldsymbol{\beta}}$ est unique). Ce sera le cas si les colonnes de $\boldsymbol{X}$ sont liées entre elles par une relation linéaire. Cela peut arriver par exemple si pour chaque individu sont mesurées le nombre d'années d'études pré-universitaires, le nombre d'années d'études supérieures, ainsi que le nombre total d'années d'études. En pratique, il suffit donc de passer les variables en revue et d'éliminer toute redondance d'information. Souvent cependant, les problèmes sont causés par le fait que le déterminant de $\boldsymbol{X}^t\boldsymbol{X}$, bien que non nul, soit très proche de 0, causant une instabilité numérique lors de l'estimation de $\boldsymbol{\beta}$ et de σ^2.

Les équations normales peuvent encore s'écrire comme

$$\sum_{i=1}^{n} \boldsymbol{x}_i^t(y_i - \boldsymbol{\beta}^t\boldsymbol{x}_i) = \mathbf{0} \text{ pour } i = 1, 2, \ldots, n.$$

Ecrites de cette manière, elles possèdent une signification intuitive très simple : $y_i - \boldsymbol{\beta}^t\boldsymbol{x}_i$ est le résidu relatif à l'observation i et les équations normales reviennent donc à imposer l'orthogonalité entre le vecteur des résidus et le vecteur des variables explicatives. Cette orthogonalité signifie intuitivement qu'il n'y a plus rien dans les variables explicatives qui puisse apporter de l'information sur les

résidus. Nous verrons à la section suivante que cette interprétation sera conservée dans les modèles linéaires généralisés.

En calculant la dérivée partielle de la log-vraisemblance $L(\boldsymbol{\beta}, \sigma|\boldsymbol{y})$ par rapport à σ^2, on vérifie que l'estimateur du maximum de vraisemblance de σ^2 est

$$\widetilde{\sigma}^2 = \sigma^2(\widehat{\boldsymbol{\beta}}) = \frac{S_2(\widehat{\boldsymbol{\beta}})}{n} = \frac{R_0^2}{n}.$$

Nous verrons plus loin qu'on préfère souvent à $\widetilde{\sigma}^2$ un estimateur non biaisé de σ^2 qui sera défini en (**??**).

9.6.4 Matrice de prédiction

Ayant obtenu l'estimation du vecteur $\boldsymbol{\beta}$, on peut définir un estimateur $\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ de la moyenne du vecteur $\boldsymbol{Y}$ et un estimateur $\widehat{\boldsymbol{\epsilon}} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}$ du vecteur non observable $\boldsymbol{\epsilon}$, appelé vecteur des résidus. Notez qu'on peut encore écrire

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\big(\boldsymbol{X}^t\boldsymbol{X}\big)^{-1}\boldsymbol{X}^t\boldsymbol{Y},$$

ce qui conduit à la définition suivante.

Définition 9.6.2. *On appelle matrice de projection (hat matrix) associée à une matrice de données $\boldsymbol{X}$, la matrice $\boldsymbol{H}$ carrée $n \times n$ définie par*

$$\boldsymbol{H} = \boldsymbol{X}\big(\boldsymbol{X}^t\boldsymbol{X}\big)^{-1}\boldsymbol{X}^t.$$

Cette matrice fait passer de $\boldsymbol{Y}$ à $\widehat{\boldsymbol{Y}}$ puisque $\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{H}\boldsymbol{Y}$. C'est pourquoi cette matrice s'appelle aussi matrice de prédiction. En conséquence,

$$\widehat{\boldsymbol{\epsilon}} = \boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}.$$

La matrice de projection $\boldsymbol{H}$ associée à $\boldsymbol{X}$ possède les propriétés suivantes.

Proposition 9.6.3. *Soit $\boldsymbol{X}$ une matrice réelle de dimension $n \times (p+1)$, de rang $p+1$ et $\boldsymbol{H}$ la matrice de prédiction associée. Alors :*

(i) $\sum_{i=1}^{n} h_{ii} = p + 1$, *la trace de $\boldsymbol{H}$ est donc égale au nombre de coefficients de régression,*

(ii) $\sum_{i=1}^{n}\sum_{j=1}^{n} h_{ij}^2 = p + 1$,

(iii) $0 \leq h_{ij} \leq 1$ *pour tout* i.
(iv) $-1/2 \leq h_{ij} \leq 1/2$ *quel que soit* $j \neq i$,
(v) *si* $h_{ii} = 1$ *ou* $h_{ii} = 0$, *alors* $h_{ij} = 0$ *pour* $j \neq i$,
(vi) $(1 - h_{ii})(1 - h_{jj}) - h_{ij}^2 \geq 0$
(vii) $h_{ii}h_{jj} - h_{ij}^2 \geq 0$.

9.6.5 Estimation des moyennes et de la variance

Le résultat suivant donne les propriétés des estimateurs $\widehat{\boldsymbol{Y}}$ de la moyenne de $\boldsymbol{Y}$ et $\widehat{\boldsymbol{\epsilon}}$ du vecteur des résidus.

Proposition 9.6.4. *Le vecteur aléatoire* $\widehat{\boldsymbol{Y}}$ *est un estimateur sans biais de la moyenne de* $\boldsymbol{Y}$ *de matrice variance-covariance* $\sigma^2\boldsymbol{H}$. *Le vecteur* $\widehat{\boldsymbol{\epsilon}}$ *des résidus estimés est centré ; sa matrice variance-covariance est* $\sigma^2(\boldsymbol{I} - \boldsymbol{H})$. *De plus, ces deux vecteurs sont non-corrélés.*

Les composantes du vecteur $\widehat{\boldsymbol{\epsilon}}$ sont généralement corrélées ; leur corrélation dépend de la matrice $\boldsymbol{X}$ du plan d'expérience.

Comme

$$\begin{aligned}\mathbb{E}\left[\sum_{i=1}^{n}\widehat{\epsilon}_i^2\right] &= \mathbb{E}\left[\boldsymbol{Y}^t\boldsymbol{Y} - \widehat{\boldsymbol{\beta}}^t\boldsymbol{X}^t\boldsymbol{X}\widehat{\boldsymbol{\beta}}\right] \\ &= \mathbb{E}\left[\boldsymbol{Y}^t(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}\right] \\ &= \mathrm{Trace}\big(\sigma^2(\boldsymbol{I} - \boldsymbol{H})\big) = \sigma^2(n - p - 1),\end{aligned}$$

il suffit pour avoir un estimateur sans biais de σ^2, de considérer

$$\widehat{\sigma}^2 = \frac{1}{n - p - 1}\sum_{i=1}^{n}\widehat{\epsilon}_i^2. \tag{9.24}$$

9.6.6 Mesure de la qualité de l'ajustement : le coefficient de détermination

Afin d'apprécier la qualité de l'ajustement fourni par le modèle, on se sert généralement du coefficient de détermination, ou pourcentage de variance expliquée par le modèle, défini par

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(\widehat{y}_i - y_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2} = \frac{\sum_{i=1}^{n}\left(\widehat{y}_i - \overline{y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2}.$$

La valeur de R^2 est comprise entre 0 et 1, le modèle étant d'autant meilleur que R^2 est proche de 1. Le coefficient de détermination

mesure "la part de variabilité de Y due à sa régression linéaire par moindres carrés sur les variables explicatives X_i". Dans le cas où il n'y a qu'une seule variable explicative (i.e., $p = 2$), R^2 est le carré du coefficient de corrélation linéaire entre Y et X_1. Il convient pour terminer de noter que R^2 n'est utile que si le modèle comprend un terme indépendant β_0.

9.6.7 Résidus standardisés

Contrairement aux résidus théoriques ϵ_i (non observables), les résidus estimés $\widehat{\epsilon_i}$ n'ont pas une variance constante et sont généralement corrélés. On préfère donc les résidus standardisés, donnés par

$$T_i = \frac{\widehat{\epsilon_i}}{\widehat{\sigma}\sqrt{1 - h_{ii}}}$$

qui remédient à ces inconvénients.

9.6.8 Résultats inférentiels pour les paramètres

Poursuivons l'analyse statistique faite précédemment en complétant les propriétés des estimateurs $\widehat{\boldsymbol{\beta}}$ et $\widehat{\sigma}^2$. Les résultats suivants sont fondamentaux.

Proposition 9.6.5. *(i) L'estimateur $\widehat{\boldsymbol{\beta}}$ donné en (**??**) a pour loi $\mathcal{N}or_{p+1}\left(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^t\boldsymbol{X})^{-1}\right)$.*

*(ii) L'estimateur $\widehat{\sigma}^2$ donné en (**??**) est tel que*

$$\frac{(n-p-1)\widehat{\sigma}^2}{\sigma^2}$$

suit la loi khi-carrée à $n - p - 1$ degrés de liberté.

(ii) Les estimateurs $\widehat{\boldsymbol{\beta}}$ et $\widehat{\sigma}^2$ sont indépendants.

9.6.9 Tests d'une hypothèse simple

Soit $\boldsymbol{\beta}_0$ une valeur fixée du vecteur des coefficients du modèle de régression. Pour tester l'hypothèse nulle $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$, contre sa négation, nous allons appliquer le test du rapport des maximums de vraisemblance.

Proposition 9.6.6. *Soit $\boldsymbol{\beta}_0$ une valeur fixée du paramètre. On rejette H_0 au niveau de confiance α lorsque*

$$\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^t\boldsymbol{\beta}_0)^2 - \sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^t\widehat{\boldsymbol{\beta}})^2 \geq \widehat{\sigma}^2 F_{p+1,n-p-1;1-\alpha},$$

où $F_{p+1,n-p-1;1-\alpha}$ est le quantile d'ordre $1-\alpha$ de la loi de Fisher-Snedecor à $p+1$ et $n-p-1$ degrés de liberté.

9.6.10 Comparaison de modèles emboîtés

Supposons que nous ayons à choisir entre un modèle M_0 et un autre modèle M_1, tels que M_0 fait intervenir une partie seulement des variables explicatives comprises dans M_1. Formellement, M_0 est donc obtenu en posant $\beta_j = 0$ pour un ensemble d'indices $j \in \mathcal{E}_0$. On parle dans ce cas de modèles emboîtés.

Le choix entre M_0 et M_1 revient donc à tester la nullité des β_j pour $j \in \mathcal{E}_0$. Nous baserons notre décision sur la statistique du rapport de vraisemblance

$$\frac{\max_{(\boldsymbol{\beta},\sigma^2)\in M_1} \mathcal{L}(\boldsymbol{\beta},\sigma^2|\boldsymbol{y})}{\max_{(\boldsymbol{\beta},\sigma^2)\in M_0} \mathcal{L}(\boldsymbol{\beta},\sigma^2|\boldsymbol{y})}$$

et nous rejetterons M_0 au profit de M_1 si ce rapport est suffisamment grand.

En notant $\widehat{\boldsymbol{\beta}}_1$ et $\widehat{\sigma}_1^2$ les estimateurs du maximum de vraisemblance dans le modèle M_1, et $\widehat{\boldsymbol{\beta}}_0$ et $\widehat{\sigma}_0^2$ les estimateurs du maximum de vraisemblance dans le modèle M_0, on peut encore utiliser la statistique de test

$$\frac{\sum_{i=1}^n \left(y_i - \widehat{\boldsymbol{\beta}}_0^t \boldsymbol{x}_i\right)^2 - \sum_{i=1}^n \left(y_i - \widehat{\boldsymbol{\beta}}_1^t \boldsymbol{x}_i\right)^2}{\sum_{i=1}^n \left(y_i - \widehat{\boldsymbol{\beta}}_1^t \boldsymbol{x}_i\right)^2}.$$

Si le modèle M_0 compte p_0+1 paramètres, contre p_1+1 pour M_1, la statistique ci-dessus, multipliée par $(n-p_1-1)/(p_1-p_0)$ obéit à la loi de Fisher de paramètres p_1-p_0 et $n-p_1-1$.

Exemple 9.6.7 (Test de contraintes linéaires entre paramètres)**.** *Supposons que, dans le cadre de l'analyse du risque d'impayé, nous disposions à propos d'un couple marié du montant des revenus professionnels de l'époux et du montant pour son épouse. On pourrait se demander s'il est pertinent de faire figurer ces deux variables parmi les facteurs explicatifs, donc de recourir à un modèle*

$$\ldots + \beta_j \times \text{ revenus époux } + \beta_{j+1} \times \text{ revenus épouse } + \ldots$$

ou, au contraire, si seul le revenu global du ménage importe. On testera dans ce cas l'hypothèse $H_0 : \beta_j = \beta_{j+1}$. On procèdera comme ci-dessus avec le modèle M_0 défini par la contrainte $\beta_j = \beta_{j+1}$ alors que le modèle M_1 autorise ces deux paramètres à différer.

9.6.11 Régions de confiance

Les régions de confiance sont définies par l'ensemble des paramètre $\boldsymbol{\beta}$ qui ne sont pas rejetés, au seuil donné, comme valeurs possibles du paramètre, par le test du rapport des vraisemblances maximales. Ce sont donc les régions, dans l'espace des paramètres, formées par l'ensemble des valeurs qui paraissent raisonnables au vu des observations recueillies.

Proposition 9.6.8. *La région de confiance pour $\boldsymbol{\beta}$ au niveau $1-\alpha$ est définie par l'ensemble des valeurs $\boldsymbol{\beta}$ telles que*

$$(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})^t \boldsymbol{X}^t \boldsymbol{X} (\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}) \leq \widehat{\sigma}^2 F_{p+1,n-p-1;1-\alpha}.$$

9.6.12 Intervalles de confiance

Plutôt que les régions de confiance, ce sont les intervalles de confiance des paramètres pris séparément qui apparaissent le plus souvent dans les logiciels commerciaux. On les obtient classiquement en remarquant que la $j$ème composante de $\widehat{\boldsymbol{\beta}}$, $\widehat{\beta}_j$, est de loi normale de moyenne β_j et de variance $\sigma^2\big(\boldsymbol{X}^t\boldsymbol{X}\big)^{-1}_{jj}$ où $\big(\boldsymbol{X}^t\boldsymbol{X}\big)^{-1}_{jj}$ est le $j$ème élément diagonal de l'inverse de la matrice $\boldsymbol{X}^t\boldsymbol{X}$.

Lorsque σ^2 est connu, l'intervalle de confiance pour β_j est formé par l'ensemble des valeurs $\xi \in \mathbb{R}$ qui satisfont

$$|\widehat{\beta}_j - \xi| \leq \sigma \sqrt{\big(\boldsymbol{X}^t\boldsymbol{X}\big)^{-1}_{jj}}\, z_{\alpha/2},$$

où $z_{\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ associé à la loi $\mathcal{N}or(0,1)$. Lorsque σ^2 est inconnu, un intervalle de confiance au niveau $1-\alpha$ pour β_j est donné par l'ensemble des valeurs $\xi \in \mathbb{R}$ satisfaisant à

$$|\widehat{\beta}_j - \xi| \leq \widehat{\sigma} \sqrt{\big(\boldsymbol{X}^t\boldsymbol{X}\big)^{-1}_{jj}}\, t_{n-p-1;1-\alpha/2},$$

où $t_{n-p-1;1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ associé à la loi de Student à $n-p-1$ degrés de liberté.

Remarque 9.6.9. *Les intervalles de confiance définis ci-dessus ne sont pas adéquats si on veut considérer plusieurs paramètres simultanément, car ils ne tiennent pas compte de la dépendance des paramètres. Lorsque le paramètre $\boldsymbol{\beta}$ a deux composantes β_1 et β_2, la région de confiance, qui dans ce cas est une ellipse, est formée des couples (β_1, β_2) expliquant raisonnablement les observations. Elle tient compte de la corrélation entre $\hat{\beta}_1$ et $\hat{\beta}_2$. Au contraire, les intervalles de confiance respectifs des paramètres β_1 et β_2 localisent la valeur d'une composante, sans tenir compte de la valeur prise par l'autre.*

9.6.13 Mesures d'influence

Principe

Les résultats de l'ajustement par moindres carrés du modèle linéaire à un ensemble d'observations, peuvent être sensiblement modifiés par la suppression ou la perturbation de certaines données. Différentes statistiques ont été définies dans le but de quantifier l'influence de chacune des observations sur l'ajustement du modèle. Elles se fondent essentiellement sur les résidus $\hat{\epsilon}_i$ et la matrice de projection $\boldsymbol{H}$.

Nous présentons ici l'approche par suppression, basée essentiellement sur la comparaison entre les résultats obtenus lorsqu'on ajuste le modèle à l'ensemble des données et ceux obtenus lorsqu'on ajuste le modèle après avoir supprimé une ou plusieurs observations.

Effet de la suppression d'une observation

Les caractéristiques obtenues après suppression de l'observation i seront affectées de l'indice (i) entre parenthèses. Ainsi, $\widehat{\boldsymbol{\beta}}_{(i)}$ est le vecteur des régresseurs estimé sur base des $n-1$ observations $y_1, ..., y_{i-1}, y_{i+1}, ..., y_n$. On peut étudier l'effet de chaque observation sur les estimations dans un modèle de régression.

Proposition 9.6.10. *Après suppression de la ième observation, les estimateurs des moindres carrés $\widehat{\boldsymbol{\beta}}_{(i)}$ et $\hat{\sigma}^2_{(i)}$ des paramètres $\boldsymbol{\beta}$ et σ^2 du modèle linéaire vérifient les équations suivantes :*

$$\widehat{\boldsymbol{\beta}}_{(i)} = \widehat{\boldsymbol{\beta}} - (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{x}_i.\frac{\epsilon_i}{1-h_{ii}}$$

et

$$(n-p-2)\hat{\sigma}^2_{(i)} = (n-p-1)\hat{\sigma}^2 - \frac{\epsilon_i^2}{1-h_{ii}}.$$

Les deux formules précédentes montrent que les estimateurs $\widehat{\boldsymbol{\beta}}_{(i)}$ et $\hat{\sigma}^2_{(i)}$ ne dépendent que de $\widehat{\boldsymbol{\beta}}$, $\hat{\sigma}^2$ et du $i$ème élément diagonal de $\boldsymbol{H}$. A chaque suppression d'une observation il n'est donc pas nécessaire, pour les calculer, de procéder à un ajustement du modèle appauvri.

Mesures diagnostiques basées sur les résidus

Pour vérifier l'adéquation d'une observation au modèle on peut regarder si sa suppression a une influence sur sa prédiction. Plus précisément, il s'agit de savoir si l'observation y_i est suffisamment

proche de sa valeur prédite $\hat{y}_{(i)}$ obtenue en n'utilisant pas la $i$ème observation dans le calcul. Comme y_i n'est pas utilisée dans le calcul de $\hat{y}_{(i)}$ les variables aléatoires y_i et $\hat{y}_{(i)}$ sont non corrélées. Leur différence $y_i - \hat{y}_{(i)}$ a donc pour variance

$$\sigma^2\Big(1 + \boldsymbol{x}_i^t(\boldsymbol{X}_{(i)}^t\boldsymbol{X}_{(i)})^{-1}\boldsymbol{x}_i\Big).$$

Lorsque le paramètre σ^2 est inconnu, il est estimé par la variance résiduelle $\hat{\sigma}^2_{(i)}$, obtenue à partir de l'équation de régression après élimination de la $i$ème observation. Cette dernière est indépendante de Y_i. Ceci conduit à définir les statistiques suivantes :

$$\begin{aligned} T_i^* &= \frac{Y_i - \hat{Y}_{(i)}}{\hat{\sigma}_{(i)}\sqrt{1 + \boldsymbol{x}_i^t(\boldsymbol{X}_{(i)}^t\boldsymbol{X}_{(i)})^{-1}\boldsymbol{x}_i}} \\ &= \frac{\hat{\epsilon}_i}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}}, \quad i = 1, 2, \ldots, n, \end{aligned}$$

appelées résidus par validation croisée. La loi de probabilité des T_i^* est donnée par le théorème suivant.

Proposition 9.6.11. *On suppose que la matrice $\boldsymbol{X}$ est de rang $p+1$. Si la suppression de la ième ligne de $\boldsymbol{X}$ ne modifie pas son rang, alors les résidus de validation croisée T_i^*, $i = 1, ..., n$, suivent la loi de Student à $n-p-2$ degrés de liberté.*

L'expérience montre que pour détecter des observations "anormales" les résidus standardisés T_i et les résidus par validation croisée T_i^* sont équivalents. Néanmoins, plusieurs auteurs préfèrent T_i^* à T_i pour les raisons suivantes :

1. Les T_i^*, $i = 1, 2, ..., n$, sont identiquement distribués, et suivent la loi de Student à $(n-p-2)$ degrés de liberté.
2. Un calcul simple donne :

$$T_i^* = T_i\sqrt{\frac{n-p-1}{n-p-T_i^2}}.$$

   Cette relation montre que T_i^* est une fonction monotone de T_i, et détecte mieux les observations à fort résidu.

3. Puisque $\hat{\sigma}_{(i)}$ est indépendant de y_i, cet estimateur est robuste à des erreurs grossières sur la $i$ème observation, ce qui peut arriver lors de l'acquisition des données.

Observations aberrantes

Nous nous proposons maintenant de préciser ce qu'on entend par observation aberrante dans un modèle linéaire.

Définition 9.6.12. *Une donnée aberrante est un point $(\boldsymbol{x}_i^t, y_i)$ pour lequel la valeur associée t_i^* de T_i^* est élevée (comparée au seuil donné par la loi de Student).*

Les données aberrantes sont habituellement détectées en traçant les t_i^* (ou les t_i ou les résidus partiels) séquentiellement ou en fonction d'autres variables telles que les y_i ou les $\boldsymbol{x}_i$. La détection de données aberrantes dépend uniquement de la grandeur des résidus.

Les représentations graphiques des résidus permettent souvent, non seulement de déceler des données aberrantes, mais également de s'assurer de la validité du modèle. Nous allons rappeler les plus classiques.

Représentation de la loi empirique une première manière de représenter les résidus est de tracer l'histogramme, la densité lissée, etc., de la loi empirique des résidus. L'utilisation de telles représentations pour contrôler l'hypothèse de normalité des données n'a de sens que pour de grands échantillons.

Représentations en fonction des valeurs ajustées on peut également tracer les résidus en fonction des valeurs ajustées. En effet nous avons vu que lorsque les hypothèses associées au modèle sont correctes, les résidus et les valeurs prédites sont non corrélés. Par conséquent, le tracé des points ne devrait pas avoir de structure particulière. Ce type de tracé donne des indications sur la validité des hypothèses de linéarité, ainsi que sur l'homogénéité de la variance de l'aléa. Par exemple, une courbure dans la forme des résidus suggère que l'hypothèse de linéarité n'est pas judicieuse, alors qu'un comportement monotone de la variabilité des résidus avec $\hat{y}_i$ indique une variance non constante des erreurs.

Mesures diagnostiques basées sur la matrice de projection

La disposition des points dans l'espace des régresseurs joue un rôle important. La matrice de projection $\boldsymbol{H}$ permet d'évaluer partiellement cette influence. Rappelons que les composantes de $\widehat{\boldsymbol{\epsilon}}$ sont des variables centrées, de variance égale à $\sigma^2(1 - h_{ii})$ où

$$h_{ii} = \boldsymbol{x}_i^t(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{x}_i, \quad i = 1, 2, ..., n.$$

La variance du résidu estimé est d'autant plus faible que h_{ii} est grand, et la valeur de h_{ii} mesure l'influence de l'observation y_i sur la valeur ajustée $\hat{y}_i$. La matrice $\boldsymbol{H}$ est symétrique et idempotente car elle projette $\boldsymbol{Y}$ sur le sous-espace vectoriel engendré par les vecteurs colonnes de la matrice $\boldsymbol{X}$.

La $i$ème composante de $\widehat{\boldsymbol{Y}}$ est :

$$\hat{Y}_i = \sum_{j=1}^{n} h_{ij} Y_j = h_{ii} Y_i + \sum_{j \neq i}^{n} h_{ij} Y_j, \;\; i = 1, 2, ..., n,$$

et donc h_{ii} représente le "poids" de l'observation y_i sur la détermination de la prédiction $\hat{y}_i$. En particulier, si $h_{ii} = 1$, $\hat{y}_i$ est déterminée par la seule observation y_i. Par ailleurs, si $h_{ii} = 0$, l'observation y_i n'aura aucune influence sur $\hat{y}_i$.

Comme la trace de $\boldsymbol{H}$ vaut $p+1$, la moyenne des h_{ii} est donc égale à $(p+1)/n$. Lorsque h_{ii} est "grand" nous avons vu que l'observation correspondante est influente. Ceci a conduit à qualifier d'influents les points pour lesquels $h_{ii} > 2(p+1)/n$. D'autres auteurs prèfèrent placer les bornes à 0.2 et 0.5 : des valeurs de $h_{ii} \leq 0.2$ sont normales ; des valeurs entre 0.2 et 0.5 sont considérées comme influentes.

La distance de Cook

Sous l'hypothèse de normalité des observations et lorsque σ est inconnu, la région de confiance du vecteur $\boldsymbol{\beta}$ des coefficients du modèle linéaire est :

$$C_\alpha = \left\{ \boldsymbol{\beta} \in \mathbb{R}^{p+1} | (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})^t \boldsymbol{X}^t \boldsymbol{X} (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) \leq \hat{\sigma}^2 p F_{p+1, n-p-1; 1-\alpha} \right\}.$$

Cette inégalité définit un ellipsoïde centré au point $\widehat{\boldsymbol{\beta}}$. L'influence de la $i$ème observation peut être mesurée par le décentrage de cet ellipsoïde quand on supprime la $i$ème observation.

Cook a donc proposé la statistique

$$C_i = \frac{(\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}})^t \boldsymbol{X}^t \boldsymbol{X} (\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}})}{\hat{\sigma}^2 (p+1)},$$

pour détecter l'influence de la $i$ème observation sur les coefficients de régression. Cette statistique est appelée distance de Cook. On peut la considérer comme une distance pondérée entre $\widehat{\boldsymbol{\beta}}$ et $\widehat{\boldsymbol{\beta}}_{(i)}$. A première vue, il semble que pour pouvoir calculer la distance de Cook pour toutes les observations il faille procéder à $n+1$ régressions (une

en utilisant la totalité des données et n pour les données réduites). En fait, on peut montrer que

$$C_i = \frac{h_{ii}}{(p+1)(1-h_{ii})} T_i^2.$$

Ainsi C_i s'obtient à l'aide de quantités déjà calculées au cours de l'ajustement du modèle complet. Par ailleurs, cette relation montre que la distance C_i est une fonction croissante du carré du résidu standardisé et de h_{ii}. Lorsque C_i est grand l'observation correspondante a une influence simultanée sur tous les paramètres de l'ajustement. Cook suggère de comparer chaque C_i aux quantiles de la loi de Fisher-Snedecor à $p+1$ et $n-p-1$ degrés de liberté, bien que les C_i ne suivent pas exactement une telle loi. Il ne s'agit donc pas là d'un test rigoureux.

La distance des vraisemblances

Si $L(\boldsymbol{\beta}, \sigma^2)$ désigne la log-vraisemblance pour $\boldsymbol{\beta}$ et σ^2, alors la statistique :

$$LD_i = 2.\Big\{ L(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2) - L(\widehat{\boldsymbol{\beta}}_{(i)}, \widehat{\sigma}^2_{(i)}) \Big\},$$

est appelée distance des vraisemblances. On peut encore montrer que

$$LD_i = n \ln \left(\frac{n(n-p-1-T_i^2)}{(n-1)(n-p-1)} \right) + \frac{(n-1)T_i^2}{(1-h_{ii})(n-p-1-T_i^2)} - 1.$$

Cette mesure est utile lorsque l'on s'intéresse à l'influence conjointe de l'observation i sur les estimateurs de $\boldsymbol{\beta}$ et σ^2. La distance des vraisemblances est comparée au percentile $1-\alpha$ de la loi khi-carrée à $p+1$ degrés de liberté.

9.6.14 Moindres carrés pondérés

Définition

Passons à présent au modèle où les observations $Y_1, Y_2, \ldots, Y_n$ admettent la représentation

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i \text{ avec } \epsilon_i \sim \mathcal{N}or(0, \sigma^2/w_i),$$

où w_i est un poids associé à l'observation i. Typiquement, ce poids apparaît lorsque Y_i est la moyenne de w_i observations. Notez qu'un

poids important s'accompagne d'une variance faible. En d'autres termes, les écarts par rapport à la moyenne $\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$ seront d'autant moins tolérés que les poids seront élevés. De manière équivalente, on a

$$Y_i \sim \mathcal{N}or\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}, \frac{\sigma^2}{w_i}\right), \quad i = 1, 2, \ldots, n.$$

Le formalisme matriciel est ici aussi fort utile pour analyser ce modèle. On voit facilement que le modèle de régression linéaire peut se réécrire vectoriellement comme en (**??**) où $\boldsymbol{Y}$, $\boldsymbol{\beta}$, $\boldsymbol{X}$ sont tels que définis précédemment, et $\boldsymbol{\epsilon} \sim \mathcal{N}or_n(\mathbf{0}, \sigma^2 \boldsymbol{W})$, avec

$$\boldsymbol{W} = \begin{pmatrix} 1/w_1 & 0 & \cdots & 0 \\ 0 & 1/w_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1/w_n \end{pmatrix}.$$

Estimation des paramètres

La fonction de vraisemblance relative aux observations $y_1, y_2, \ldots, y_n$ est

$$\mathcal{L}(\boldsymbol{\beta}, \sigma | \boldsymbol{y}) = \frac{1}{(2\pi)^{n/2}\sigma |\boldsymbol{W}|^{1/2}} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^t \boldsymbol{W}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right).$$

Quelle que soit la valeur de σ^2, maximiser $L(\boldsymbol{\beta}, \sigma | \boldsymbol{y}) = \ln \mathcal{L}(\boldsymbol{\beta}, \sigma | \boldsymbol{y})$ par rapport à $\boldsymbol{\beta}$ revient à minimiser

$$\begin{aligned} S_2(\boldsymbol{\beta}) &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^t \boldsymbol{W}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\ &= \sum_{i=1}^{n} w_i (y_i - \boldsymbol{x}_i^t \boldsymbol{\beta})^2 \end{aligned}$$

où le vecteur colonne $\boldsymbol{x}_i$ contient les éléments de la $i$ème ligne de la matrice $\boldsymbol{X}$. Il s'agit donc de minimiser une somme pondérée des carrés des écarts entre la réponse y_i et le prédicteur linéaire $\boldsymbol{x}_i^t \boldsymbol{\beta}$. Les observations pour lesquelles le poids w_i est grand seront donc pénalisées lors de la minimisation de $S_2(\boldsymbol{\beta})$: on tolérera moins d'écart entre y_i et $\boldsymbol{x}_i^t \boldsymbol{\beta}$ pour ces indices.

Pour que $\widehat{\boldsymbol{\beta}}$ minimise S_2 il faut que ce soit un point stationnaire de cette expression. Il est donc obtenu en dérivant S_2 par rapport

à $\boldsymbol{\beta}$ et en identifiant le gradient à $\mathbf{0}$. L'estimateur du maximum de vraisemblance $\widehat{\boldsymbol{\beta}}$ de $\boldsymbol{\beta}$ est solution des équations normales

$$\boldsymbol{X}^t\boldsymbol{W}\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}^t\boldsymbol{W}\boldsymbol{Y} = 0,$$

qui donneront

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^t\boldsymbol{W}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^t\boldsymbol{W}\boldsymbol{Y} \tag{9.25}$$

qui définit l'estimateur du maximum de vraisemblance de $\boldsymbol{\beta}$.

Matrice de projection

Ayant obtenu l'estimation du vecteur $\boldsymbol{\beta}$, on peut définir un estimateur $\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ de la moyenne du vecteur $\boldsymbol{Y}$ et le vecteur des résidus $\widehat{\boldsymbol{\epsilon}} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}$. Notez qu'on peut encore écrire

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\left(\boldsymbol{X}^t\boldsymbol{W}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^t\boldsymbol{W}\boldsymbol{Y}.$$

En définissant la matrice de projection (hat matrix) comme étant la matrice $\boldsymbol{H}$ carrée $n \times n$ définie par

$$\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}^t\boldsymbol{W}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^t\boldsymbol{W},$$

on voit que cette matrice fait passer de $\boldsymbol{Y}$ à $\widehat{\boldsymbol{Y}}$ puisque $\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{H}\boldsymbol{Y}$. Notez que $\boldsymbol{H}$ ne dépend pas des $\boldsymbol{Y}$, mais uniquement des poids $\boldsymbol{W}$ et des régresseurs $\boldsymbol{X}$.

9.7 Modèles additifs

9.7.1 Principe

Dans le score η_i, on ne peut pas être sûr que les variables continues interviendront linéairement. Ainsi, si la première variable est continue (pensez par exemple à l'âge de l'assuré), on pourrait avoir avantage à considérer un modèle du type

$$Y_i \sim \mathcal{N}or(\eta_i, \sigma^2) \text{ où } \eta_i = \beta_0 + f(x_{i1}) + \sum_{j=2}^{p} \beta_j x_{ij}.$$

La fonction f, à estimer sur base des données, exprimera le lien entre le score η_i et x_{i1}, compte tenu des autres variables $x_{i2}, \ldots, x_{ip}$. C'est précisément là l'avantage de la démarche par rapport à la détermination *a priori* de classes pour une variable quantitative : lorsque cette démarche s'inscrit dans un modèle de régression, on corrige pour les autres variables explicatives.

Remarque 9.7.1. *Une autre démarche, plus classique, consiste à substituer à x_{i1} différentes transformations de la variable (par des polynômes, des sinusoïdes, etc.) afin de capturer l'influence non linéaire. Cependant, cette approche est beaucoup moins convaincante que celle consistant à estimer f directement. En effet, elle conduit à une multiplication des paramètres à estimer et impose un choix arbitraire des transformations des variables explicatives (qui peut s'avérer erroné).*

L'innovation des modèles additifs consiste à permettre au score d'être une fonction additive non nécessairement linéaire des covariables. Spécifiquement, on retiendra la forme

$$\eta = c + \sum_{j=1}^{p} f_j(x_j)$$

pour le score, où les fonctions $f_j(\cdot)$, $j = 1, \ldots, p$, supposées régulières, traduisent l'influence des variables explicatives $x_1, \ldots, x_p$ sur la réponse y. Si toutes les fonctions f_j sont affines, on est ramené au modèle linéaire.

Afin d'expliquer comment on ajuste un modèle additif, nous procèderons en deux étapes successives. Nous commencerons par traiter le modèle simple $y = f(x) + \epsilon$ où l'erreur ϵ est normale centrée. Nous verrons comment il est possible d'estimer la fonction inconnue $f(\cdot)$ traduisant l'influence de x sur y à l'aide de techniques de lissage ou d'ajustement linéaire local. Ensuite, nous passerons au cas de plusieurs variables explicatives $y = \sum_{j=1}^{p} f_j(x_j) + \epsilon$.

9.7.2 Le cas d'un seul régresseur

Commençons par considérer le modèle élémentaire suivant. Nous disposons de n observations (x_i, y_i), $i = 1, \ldots, n$, où la variable à expliquer y_i et le régresseur x_i sont continus. L'influence de x_i sur y_i est modélisée à l'aide d'une fonction $f(\cdot)$, supposée régulière, à un bruit additif près, i.e.,

$$y_i = f(x_i) + \epsilon_i \tag{9.26}$$

où les erreurs ϵ_i sont supposées indépendantes de loi $\mathcal{N}or(0, \sigma^2)$. La spécification (**??**) permet donc de s'affranchir de la contrainte de linéarité imposée dans la régression classique (où $f(x_i) = \beta_0 + \beta_1 x_i$). L'estimation de $f(\cdot)$ se fera à l'aide des techniques de lissage et nous nous intéresserons plus particulièrement à des lisseurs linéaires, en

ce sens que $\widehat{f}(x_i)$ peut s'exprimer comme une combinaison linéaire des valeurs $y_1, \ldots, y_n$, i.e.

$$\widehat{f}(x_i) = \sum_{j=1}^{n} h_{ij} y_j \tag{9.27}$$

où les poids $h_{ij} = h(x_i, x_j)$ dépendent de l'endroit x_i où la réponse $f(x_i)$ doit être estimée. Si on définit le vecteur $\boldsymbol{f} = (f(x_1), \ldots, f(x_n))^t$, (**??**) se réécrit comme $\widehat{\boldsymbol{f}} = \boldsymbol{H}\boldsymbol{y}$.

Méthode Loess

Principe : moindres carrés pondérés Cette méthode proposée par Cleveland (1979) fait partie de la famille des régressions polynomiales locales. Elle consiste à approximer localement $f(\cdot)$ par une droite[4]. Cette approche a été développée par Cleveland et al. (1988, 1991). L'idée est d'utiliser les λ plus proches voisins de x afin d'estimer $f(x)$. Le voisinage s'apprécie par rapport aux variables explicatives : on utilise les λ observations dont les variables explicatives sont les plus proches de x pour estimer la réponse $f(x)$.

La métode Loess peut se décomposer comme suit : disposant des n observations (x_i, y_i), $i = 1, 2, \ldots, n$,

(i) les λ plus proches voisins de x sont identifiés (soit $\mathcal{V}(x)$ l'ensemble de ceux-ci) ;

(ii) la distance $\Delta(x)$ séparant x du plus éloigné de ses λ plus proches voisins est calculée comme

$$\Delta(x) = \max_{i \in \mathcal{V}(x)} |x - x_i|;$$

(iii) les poids

$$w_i(x) = K\left(\frac{|x - x_i|}{\Delta(x)}\right)$$

sont assignés à chaque élément de $\mathcal{V}(x)$. La fonction $K(\cdot)$ assignant les poids $w_i(x)$ aux observations de $\mathcal{V}(x)$ doit avoir les propriétés suivantes :

(i) $K(u) \geq 0$ pour tout u

(ii) $K(u) = 0$ pour $u > 1$

4. Notez qu'on peut également approximer localement f par une constante, auquel cas, on retrouve des moyennes mobiles pondérées, ou par un polynôme de degré 2. Nous ne considérerons ici que l'ajustement linéaire local.

(iii) K est non-croissante sur (0,1).

Comme suggéré par CLEVELAND (1979), nous utiliserons la fonction $K(\cdot)$ donnée par

$$K(u) = \begin{cases} (1-u^3)^3 \text{ pour } 0 \leq u < 1 \\ 0 \text{ sinon.} \end{cases}$$

On donne donc plus de poids aux points du voisinage les plus proches de x.

(iv) $\widehat{f}(x)$ est obtenue en régressant les y_i, $i \in \mathcal{V}(x)$, sur les x_i correspondants à l'aide d'un ajustement des moindres carrés pondérés et en se servant de la droite de régression pour prédire la réponse correspondant à x.

Cette approche fournit une réponse de la forme (**??**).

Dans le cas qui nous occupe ici, à savoir celui d'un seul régresseur, la valeur ajustée $\widehat{y}_i = \widehat{\beta}_0(x_i) + \widehat{\beta}_1(x_i)x_i$ s'obtient en déterminant $\widehat{\beta}_0(x_i)$ et $\widehat{\beta}_1(x_i)$ de façon à minimiser

$$\sum_{k \in \mathcal{V}(x_i)} w_k(x_i)\big(y_k - \beta_0(x_i) - \beta_1(x_i)x_k\big)^2$$

ce qui fournit

$$\widehat{y}_i = \sum_{k=1}^{n} h_k(x_i)y_k,$$

où $h_k(x_i)$ ne dépend pas des y_j, $j = 1, \ldots, n$ ($h_k(x_i)$ ne dépend que des régresseurs).

Si on s'intéresse à la réponse pour une valeur x non observée, le modèle utilisé pour estimer $f(x)$ est donc

$$y_i = \beta_0(x) + \beta_1(x)x_i + \epsilon_i \text{ pour } i \in \mathcal{V}(x)$$

où les estimations $\widehat{\beta}_0(x)$ et $\widehat{\beta}_1(x)$ des paramètres $\beta_0(x)$ et $\beta_1(x)$ s'obtiennent en minimisant

$$\sum_{k \in \mathcal{V}(x)} w_k(x)\big(y_k - \beta_0(x) - \beta_1(x)x_k\big)^2.$$

Ceci donne finalement $\widehat{f}(x) = \widehat{\beta}_0(x) + \widehat{\beta}_1(x)x$.

Intervalles de confiance Si on note $\widehat{\boldsymbol{y}}$ le vecteur des valeurs ajustées $(\widehat{f}(x_1), \ldots, \widehat{f}(x_n))^t$ et $\widehat{\boldsymbol{\epsilon}}$ le vecteur des résidus, on a

$$\widehat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y} \text{ et } \widehat{\boldsymbol{\epsilon}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}.$$

Par conséquent, $\widehat{\boldsymbol{y}}$ et $\widehat{\boldsymbol{\epsilon}}$ obéissent tous deux aux lois normales multivariées de matrice variance covariance $\sigma^2\boldsymbol{H}\boldsymbol{H}^t$ et $\sigma^2(\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{I} - \boldsymbol{H})^t$, respectivement. Ceci nous permet d'obtenir des intervalles de confiance pour $f(x)$. Si nous posons

$$\delta_k = \text{Trace}\Big((\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{I} - \boldsymbol{H})^t\Big)^k \text{ pour } k = 1, 2,$$

on voit facilement que

$$\mathbb{E}\left[\sum_{i=1}^{n} \widehat{\epsilon}_i^2\right] = \sigma^2\delta_1$$

de sorte que

$$\widehat{\sigma^2} = \frac{1}{\delta_1}\sum_{i=1}^{n} \widehat{\epsilon}_i^2$$

est un estimateur sans biais de σ^2. De plus,

$$\frac{\delta_1^2}{\delta_2} \times \frac{\widehat{\sigma}^2}{\sigma^2}$$

suit approximativement la loi khi-carrée à δ_1^2/δ_2 degrés de liberté (où δ_1^2/δ_2 sera arrondi à l'entier le plus proche). Par conséquent,

$$\frac{\widehat{f}(x) - f(x)}{\widehat{\sigma}\sqrt{\sum_{i=1}^{n} h_i^2(x)}}$$

obéit approximativement à la loi de Student à δ_1^2/δ_2 degrés de liberté. Ceci permet d'obtenir des intervalles de confiance pour la réponse $f(\cdot)$ en différents points x.

Degré de lissage Comme le lecteur aura pu s'en rendre compte, l'approche Loess dépend donc du nombre λ de points contenu dans le voisinage $\mathcal{V}(x_0)$ du point x_0 considéré. Le nombre de plus proches voisins, le plus souvent exprimé en pourcentage de la taille du jeu de données, joue le rôle de paramètre de lissage. La sélection d'une valeur optimale pour λ est considérée plus bas.

Lors de la comparaison des modèles, il est bon de disposer d'une mesure de la complexité de ceux-ci. Ceci peut s'effectuer à l'aide d'un nombre de degrés de liberté associé aux lisseurs satisfaisant (**??**). Le nombre de degrés de liberté DF_λ est fourni par la trace de la matrice $\boldsymbol{H}_\lambda$ dont les éléments sont les h_{ij} intervenant dans (**??**). Ce choix provient du fait que dans le modèle de régression linéaire classique, nous avons vu plus haut que la trace de la matrice qui fait passer des observations y_i aux valeurs ajustées $\widehat{y}_i = \widehat{\boldsymbol{\beta}}^t \boldsymbol{x}_i$ vaut le nombre de paramètres (i.e. la dimension de $\boldsymbol{\beta}$).

Mesure de la qualité de l'ajustement La mesure de la qualité de l'ajustement est plus complexe pour les modèles où il s'agit d'estimer une fonction que pour les modèle paramétriques classiques. La plupart des critères font intervenir à la fois une mesure de la qualité de l'ajustement et une mesure de la complexité du modèle. La minimisation de tels critères permet également de sélectionner le paramètre de lissage.

Notons $\widehat{\sigma}^2$ la somme des carrés des résidus. Hurvich & Simonoff (1998) ont proposé le critère suivant :

$$AICC1 = n \ln \widehat{\sigma}^2 + n \frac{\delta_1/\delta_2 (n + DF_\lambda)}{\delta_1^2/\delta_2 - 2}$$

appelé AIC corrigé. Le premier terme intervenant dans AICC1 mesure la qualité de l'ajustement tandis que le deuxième évalue la complexité du modèle. Ce critère permet de sélectionner la valeur optimale de λ (i.e. celle minimisant AICC1).

Extension à des régresseurs multiples L'extension de la méthode Loess à plus d'un régresseur est immédiate. On considère alors un modèle de la forme

$$y_i = f(x_{i1}, x_{i2}, \ldots, x_{ip}) + \epsilon_i.$$

Notez que ce modèle n'est pas additif. Il suffit d'approcher localement la fonction f des variables explicatives par un hyperplan et ensuite de procéder comme expliqué plus haut dans le cas d'un seul régresseur. Il importe dans ce cas que les différents régresseurs aient des valeurs comparables (afin que les voisinages multidimensionnels ne soient pas déterminés sur la seule base de la variable prenant les plus grandes valeurs). Pour cela, les variables explicatives sont souvent normalisées (en les divisant par la largeur de l'intervalle interquartile par exemple). La distance utilisée est la distance euclidienne dans $\mathbb{R}^p$.

Remarque 9.7.2. *En fonction de la taille du jeu de données à traiter, on peut éventuellement recourir à la méthode des "$k-d$ trees" pour définir des voisinages multidimensionnels.*

Maximum de vraisemblance pénalisée et splines cubiques

Principe : moindres carrés pénalisés Une manière ingénieuse d'estimer la fonction $f(\cdot)$ intervenant dans (**??**) consiste à minimiser la fonction objectif

$$\mathcal{O}(f) = \sum_{i=1}^{n} \Big(y_i - f(x_i)\Big)^2 + \lambda \int_{u\in\mathbb{R}} \big(f"(u)\big)^2 du. \tag{9.28}$$

Le premier terme intervenant dans $\mathcal{O}(f)$ assure que $f(\cdot)$ ajustera au mieux les données, tandis que le second pénalise un ajustement trop irrégulier. Le recours à cette technique suppose que $f(\cdot)$ est deux fois continûment différentiable et que $f"(\cdot)$ est de carré-intégrable. L'intégrale intervenant dans (**??**) a pour but de mesurer l'irrégularité de la fonction $f(\cdot)$. Notez que deux fonctions ne différant qu'à un terme linéaire près auront mêmes dérivées secondes et, partant, présenteront le même degré d'irrégularité comme mesuré dans (**??**). Dans un contexte de régression, cette dernière propriété est particulièrement appréciée.

La fonction objectif (**??**) peut se voir comme une log-vraisemblance normale pénalisée (approche PML, pour *Penalized Maximum Log-Likelihood*) : on ajoute à la log-vraisemblance un terme pénalisant l'irrégularité de l'estimateur avant de la maximiser. Cette approche remonte aux travaux de l'actuaire E. Whittaker qui, dès 1923, utilisa une technique semblable pour lisser des tables de mortalité. Pour plus de détails, voyez par exemple Hastie & Tibshirani (1990).

Lorsque $\lambda \to +\infty$, le terme pénalisant l'irrégularité de $f(\cdot)$ domine, forçant la dérivée seconde de $f(\cdot)$ à s'annuler partout, et fournissant donc la droite de régression comme solution. Pour de grandes valeurs de λ, l'intégrale domine dans (**??**) et l'estimateur résultant de la minimisation de $\mathcal{O}(f)$ présentera une courbure très faible. Au contraire, lorsque $\lambda \to 0$, la pénalisation disparaît et on obtient une interpolation parfaite (lorsque les x_i sont distincts).

Supposons tout d'abord que $x_1 < x_2 < \ldots < x_n$. La solution $\widehat{f}_\lambda$ de la minimisation de (**??**) est un spline cubique dont les noeuds sont $x_1, x_2, \ldots, x_n$ (ce qui signifie que $\widehat{f}_\lambda$ coïncide avec un polynôme du 3ème degré sur chaque intervalle (x_i, x_{i+1}), et possède des dérivées

première et seconde continues en chacun des x_i). Ceci permet de ramener la minimisation de (**??**) à celle de

$$(\boldsymbol{y}-\boldsymbol{f})^t(\boldsymbol{y}-\boldsymbol{f})+\lambda \boldsymbol{f}^t\boldsymbol{K}\boldsymbol{f} \tag{9.29}$$

où $\boldsymbol{y}^t=(y_1,\cdots,y_n)$, $\boldsymbol{f}^t=(f(x_1),\cdots,f(x_n))$ et $\boldsymbol{K}=\boldsymbol{D}^t\boldsymbol{C}^{-1}\boldsymbol{D}$ avec $\boldsymbol{D}$ une matrice tri-diagonale de dimension $(n-2)\times n$ donnée par

$$\boldsymbol{D}=\begin{pmatrix}
\frac{1}{\Delta_1} & -\left(\frac{1}{\Delta_1}+\frac{1}{\Delta_2}\right) & \frac{1}{\Delta_2} & 0 & 0 & \cdots & 0\\
0 & \frac{1}{\Delta_2} & -\left(\frac{1}{\Delta_2}+\frac{1}{\Delta_3}\right) & \frac{1}{\Delta_3} & 0 & \cdots & 0\\
0 & 0 & \frac{1}{\Delta_3} & -\left(\frac{1}{\Delta_3}+\frac{1}{\Delta_4}\right) & \frac{1}{\Delta_4} & \cdots & 0\\
\vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots\\
0 & 0 & 0 & 0 & 0 & \cdots & \frac{1}{\Delta_{n-1}}\\
0 & 0 & 0 & 0 & 0 & \cdots & -\left(\frac{1}{\Delta_{n-2}}+\frac{1}{\Delta_{n-1}}\right)\\
0 & 0 & 0 & 0 & 0 & \cdots & \frac{1}{\Delta_{n-2}}
\end{pmatrix},$$

où $\Delta_i=x_{i+1}-x_i$, et $\boldsymbol{C}$ est une matrice tri-diagonale symétrique de dimension $(n-2)\times(n-2)$ donnée par

$$\boldsymbol{C}=\frac{1}{6}\begin{pmatrix}
2\,(\Delta_1+\Delta_2) & \Delta_2 & 0 & \cdots & 0 & 0\\
\Delta_2 & 2(\Delta_2+\Delta_3) & \Delta_3 & \cdots & 0 & 0\\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots\\
0 & 0 & 0 & \cdots & 2(\Delta_{n-3}+\Delta_{n-2}) & \Delta_{n-2}\\
0 & 0 & 0 & \cdots & \Delta_{n-2} & 2(\Delta_{n-2}+\Delta_{n-1})
\end{pmatrix}.$$

La solution $\widehat{f}_\lambda$ peut alors être obtenue en annulant le gradient $-2(\boldsymbol{y}-\boldsymbol{f})+2\lambda\boldsymbol{K}\boldsymbol{f}$ de la fonction objectif (**??**), ce qui fournit

$$\widehat{\boldsymbol{f}}_\lambda=(\boldsymbol{I}+\lambda\boldsymbol{K})^{-1}\boldsymbol{y}$$

qui a bien la forme (**??**) avec $\boldsymbol{H}=(\boldsymbol{I}+\lambda\boldsymbol{K})^{-1}$. Mentionnons pour mémoire qu'il n'est pas nécessaire d'inverser $\boldsymbol{I}+\lambda\boldsymbol{K}$ pour obtenir $\widehat{f}_\lambda$ mais qu'il est plus efficace de recourir à des techniques numériques telles que l'algorithme de Reinsch.

Remarque 9.7.3. *Dans le cas où chaque observation (x_i,y_i) est munie d'un poids w_i, on recourt à la fonction objectif*

$$\mathcal{O}_w(f)=\sum_{i=1}^{n} w_i\Big(y_i-f(x_i)\Big)^2+\lambda\int_{u\in\mathbb{R}}\big(f"(u)\big)^2du.$$

A nouveau, le minimum de $\mathcal{O}_w(f)$ est obtenu en prenant pour f un spline cublique, ce qui permet d'écrire

$$\mathcal{O}_w(f)=(\boldsymbol{y}-\boldsymbol{f})^t\boldsymbol{W}(\boldsymbol{y}-\boldsymbol{f})+\lambda\boldsymbol{f}^t\boldsymbol{K}\boldsymbol{f}$$

où la matrice diagonale $\boldsymbol{W}$ reprend les poids, dont la minimisation fournit

$$\widehat{\boldsymbol{f}}_\lambda=(\boldsymbol{W}+\lambda\boldsymbol{K})^{-1}\boldsymbol{W}\boldsymbol{y}.$$

Le poids w_i peut par exemple représenter le nombre d'observations y_i dans l'échantillon correspondant à la même valeur x_i.

Degré de lissage La mesure de la complexité du modèle s'effectue exactement comme pour Loess. Le choix de λ s'opère souvent à l'aide du critère de validation croisée

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{f}_\lambda^{-i}(x_i)\right)^2$$

où $\widehat{f}_\lambda^{-i}(x_i)$ est l'estimation de $f(x_i)$ obtenue à l'aide de l'échantillon $\{(y_j, x_j),\ \ j \neq i\}$, de taille $n-1$. Le critère de validation croisée privilégie le pouvoir de prédiction du modèle retenu (alors que la somme des carrés des résidus privilégie la qualité de l'ajustement d'un ensemble donné d'observations). En effet, $\widehat{f}_\lambda^{-i}(x_i)$ est la prévision de y_i fournie par les données à l'exclusion de la $i$ème, de sorte que la différence $y_i - \widehat{f}_\lambda^{-i}(x_i)$ mesure la qualité de la prédiction fournie par le modèle.

La valeur de $\widehat{f}_\lambda^{-i}(x_i)$ est fournie pour des splines cubiques par

$$\widehat{f}_\lambda^{-i}(x_i) = \sum_{j\neq i}\frac{h_{ij}}{1-h_{ii}}y_j$$

où les poids $\frac{h_{ij}}{1-h_{ii}}$ ont pour somme 1. L'idée est très simple : on accorde un poids nul à l'observation i et les poids sont normalisés (afin d'avoir une somme égale à 1). On a donc

$$\widehat{f}_\lambda^{-i}(x_i) = \frac{1}{1-h_{ii}}\widehat{f}_\lambda(x_i) - \frac{h_{ii}}{1-h_{ii}}y_i$$

de sorte que

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \widehat{f}_\lambda(x_i)}{1-h_{ii}}\right)^2.$$

Remarque 9.7.4. *On a aussi parfois recours au critère de validation croisée généralisé, remplaçant h_{ii} par la moyenne $\frac{1}{n}\sum_{i=1}^{n}h_{ii}$, qui donne*

$$\begin{aligned} GCV(\lambda) &= \frac{1}{n\left(1-\frac{1}{n}\sum_{i=1}^{n}h_{ii}\right)^2}\sum_{i=1}^{n}\left(y_i - \widehat{f}_\lambda(x_i)\right)^2 \\ &= \frac{1}{n\left(1-\frac{DF_\lambda}{n}\right)^2}\sum_{i=1}^{n}\left(y_i - \widehat{f}_\lambda(x_i)\right)^2. \end{aligned}$$

9.7.3 Estimation à plus d'un régresseur : backfitting

Supposons à présent que l'on dispose d'observations $(y_i, \boldsymbol{x}_i)$, $i = 1, 2, \ldots, n$, où $\boldsymbol{x}_i^t = (x_{i1}, \ldots, x_{ip})$ sont p variables explicatives continues. Nous considérons le modèle

$$y_i = c + \sum_{j=1}^{p} f_j(x_{ij}) + \epsilon_i \tag{9.30}$$

où les erreurs ϵ_i sont supposées indépendantes de loi $\mathcal{N}or(0, \sigma^2)$. Notre but est d'estimer les fonctions $f_1(\cdot), \ldots, f_p(\cdot)$ traduisant l'effet des variables explicatives sur la réponse. Pour ce faire, on recourera à l'algorithme de backfitting, qui réajustera successivement les résidus partiels à chacune des variables explicatives.

Montrons comment le modèle (**??**) peut se traiter à partir du cas simple (**??**). Le principe consiste, étant donnée une première estimation $\widehat{f}_k(\cdot)$ des $f_k(\cdot)$, $k = 1, 2, \ldots, p$, à réestimer $f_j(\cdot)$ en ajustant les résidus obtenus à partir des $f_k(\cdot)$, $k \neq j$, i.e. les $r_i^{(j)}$ donnés par

$$r_i^{(j)} = y_i - \widehat{c} - \sum_{k \neq j} \widehat{f}_k(x_{ik}), \quad i = 1, 2, \ldots, n,$$

aux valeurs du $j$ème régresseur x_{ij}. On continuera de la sorte jusqu'à ce que les résultats se stabilisent. L'idée maîtresse sous-tendant l'algorithme dit de backfitting est que quel que soit j

$$\mathbb{E}\left[Y - c - \sum_{k \neq j} f_k(X_k) \Big| X_j \right] = f_j(X_j),$$

de sorte que les résidus $r_i^{(j)}$, $i = 1, \ldots, n$ reflètent la part du comportement de la variable dépendante attribuable au $j$ème régresseur.

Dorénavant, notons $\boldsymbol{f}_j^t = (f_j(x_{j1}), \cdots, f_j(x_{jn}))$ le vecteur des évaluations de $f_j(\cdot)$ en les valeurs observées du $j$ème régresseur. Comme il est bien évidemment possible d'absorber la constante c dans l'une quelconque des fonctions $f_1, \ldots, f_p$, nous prendrons dorénavant $\widehat{c} = \overline{y}$ par souci d'identifiabilité. L'idée maîtresse de la méthode consiste à définir des résidus $r_1^{(j)}, \cdots, r_n^{(j)}$ qui seront expliqués au moyen d'un modèle additif à l'aide du $j$ème régresseur. Plus précisément, l'algorithme procède comme suit :

initialisation : $\widehat{c} \leftarrow \overline{y}$, $\widehat{\boldsymbol{f}}_j^{(0)} \leftarrow \mathbf{0}$, $j = 1, 2, \ldots, p$.

cycle : pour $r = 1, 2, \ldots,$ et $j = 1, 2, \ldots, p$, mettre à jour $\widehat{\boldsymbol{f}}_j^{(r)}$ grâce à

$$\widehat{\boldsymbol{f}}_j^{(r+1)} \leftarrow \boldsymbol{H}_{\lambda_j}\left(\boldsymbol{y} - (\overline{y}, \ldots, \overline{y})^t - \sum_{k<j} \widehat{\boldsymbol{f}}_k^{(r+1)} - \sum_{k>j} \widehat{\boldsymbol{f}}_k^{(r)}\right)$$

où $\boldsymbol{H}_{\lambda_j}$ est la matrice de lissage appliquée au résidu partiel obtenu en soustrayant de l'observation $\boldsymbol{y}$ son anticipation calculée à l'aide de tous les régresseur, à l'exception du $j$ème.

critère d'arrêt : itérer l'étape ci-dessus et s'arrêter lorsque la somme des carrés des résidus

$$\left(\boldsymbol{y} - (\overline{y}, \ldots, \overline{y})^t - \sum_{j=1}^{p} \widehat{\boldsymbol{f}}_k^{(r+1)}\right)^t \left(\boldsymbol{y} - (\overline{y}, \ldots, \overline{y})^t - \sum_{j=1}^{p} \widehat{\boldsymbol{f}}_k^{(r+1)}\right)$$

cesse de décroître.

Notez que l'influence de chacun des régresseurs peut être estimée à l'aide d'un paramètre de lissage différent.

9.7.4 Comparaison des différentes approches

Considérons des simulations de 50 observations suivant un modèle $Y_i = (X_i + 1)^2 - 1 + \varepsilon_i$, où les erreurs $\varepsilon \sim \mathcal{N}or(0, 1/2)$ sont indépendantes, et où $X_i \sim \mathcal{N}or(0, 1)$.

La Figure **??** montre l'estimation par moindres carrés de Y sur X à gauche, et de X sur Y à droite. Comme le montre la partie gauche, le principe des moindres carrés consiste à minimiser la somme des carrés des distances entre Y_i et $\widehat{Y}_i$ (représentées verticalement).

FIGURE 9.6 – Régressions de Y sur X (à gauche) et de X sur Y (à droite).

La Figure **??** montre le résultat de l'ACP du nuage $(X_1, Y_1), ..., (X_n, Y_n)$, où le premier axe est obtenu en minisant la somme des carrés des distances de (X_i, Y_i) à la droite (correspondant à la projection orthogonale de (X_i, Y_i) sur la droite). Notons que la droite obtenue est ici relativement proche de celle obtenue par régression linéaire.

La partie gauche de la Figure **??** présente la régression de Y sur X (modèle affine) et de Y sur X et X^2 (modèle quadratique). Le

FIGURE 9.7 – ACP du nuage (X_i, Y_i) simulé (à gauche) et comparaison avec la droite de régression de Y sur X (à droite).

modèle simulé étant ici quadratique, on observe que c'est effectivement une modélisation qui s'ajuste relativement bien aux données. La partie droite correspond à une approche non-paramétrique, par splines cubiques.

FIGURE 9.8 – Régressions affine et quadratique de Y sur X (à gauche), et régression par splines cubiques (à droite).

Toujours sur le même échantillon de 50 points simulés, la Figure **??** présente à gauche la méthode Loess, et à droite une régression locale polynomiale, par des polynômes de degré 1, 2 et 3 respectivement. On constate que l'approche Loess, comme celle basée sur les splines cubiques, permet d'avoir une bonne idée du modèle sous-jacent.

FIGURE 9.9 – Régression de Y sur X par la méthode Loess (à gauche) et par régression locale polynomiale, par des polynômes de degré 1, 2 et 3 (à droite).

9.8 Les modèles linéaires généralisés

9.8.1 Petit historique des applications actuarielles des modèles de régression

Longtemps, les actuaires se sont limités à utiliser le modèle linéaire gaussien lorsqu'il s'agissait de quantifier l'impact de variables explicatives sur un phénomène d'intérêt (fréquence ou coût des sinistres, probabilité d'occurrence d'événements assurés, ...). A présent que la complexité des problèmes statistiques qui se posent à l'actuaire s'est considérablement accrue, il est crucial de se tourner vers des modèles tenant mieux compte de la réalité de l'assurance que ne le fait le modèle linéaire. Ce dernier impose en effet une série de limitations peu conciliables avec la réalité des nombres ou

des coûts des sinistres : densité de probabilité (approximativement) gaussienne, linéarité du score et homoscédasticité. Même s'il est possible de s'affranchir de certaines de ces contraintes en transformant préalablement la variable réponse à l'aide de fonctions bien choisies, l'approche linéaire s'accompagne de nombreux désavantages (travail sur une échelle artificielle, difficultés de revenir aux quantités initiales, ...).

Une première étape dans l'utilisation de modèles plus appropriés à la réalité de l'assurance a été franchie lors de l'application en sciences actuarielles à la fin du 20ème siècle par les actuaires londoniens de la City University des modèles linéaires généralisés (GLM, pour *Generalized Linear Models*). Ces modèles, introduits en statistique par Nelder & Wedderburn (1972), permettent de s'affranchir de l'hypothèse de normalité, en traitant de manière unifiée des réponses dont la loi fait partie de la famille exponentielle linéaire (laquelle compte, outre la loi normale, les lois de Poisson, binomiale, Gamma et Inverse Gaussienne). Voyez notamment Gouriéroux, Monfort & Trognon (1984).

La régression de Poisson (et les modèles apparentés, tels que la régression binomiale négative) est à présent un outil de choix pour l'élaboration d'une tarification automobile, supplantant largement le modèle linéaire général et la régression logistique pour l'analyse des nombres de sinistres. La percée de cette méthode au sein des compagnies date de l'inclusion dans les logiciels statistiques les plus usités (SAS en tête) de procédures permettant d'appliquer cette technique (`GENMOD`, en l'occurrence). Outre l'approche du maximum de vraisemblance, les techniques GLM permettent l'analyse d'un grand nombre de phénomènes dans une optique de quasi-vraisemblance, en ne spécifiant que la structure moyenne-variance. Les économètres français ont à cet égard prouvé des résultats fondamentaux de convergence des estimateurs obtenus de cette façon. Voyez notamment Gouriéroux, Monfort & Trognon (1984).

Plus récemment, les techniques GLM ont été appliquées avec succès aux problématiques de l'assurance vie (établissement de tables de mortalité, estimation des indicateurs démographiques, projection de la mortalité, etc.). Voyez Delwarde & Denuit (2005) pour de nombreux exemples.

Cette section est basée sur McCullagh & Nelder (1989), Antoniadis, Berruyer & Carmona (1992), Dobson (2001) et Fahrmeir & Tutz (2002). Nous insisterons sur le fait que les estimateurs du maximum de vraisemblance peuvent être obtenus à l'aide d'ajustements des moindres carrés pondérés répétés, en

définissant des pseudo-réponses appropriées. Ceci nous permettra de lever l'hypothèse de linéarité du score en passant aux modèles généralisés additifs.

Afin de bien comprendre les généralisations du modèle linéaire gaussien dont il est question dans cette section, rappelons que dans le cadre de ce modèle, on suppose que l'on cherche à modéliser une variable Y à l'aide d'un certain nombre de variables explicatives $\boldsymbol{X} = (X_1, ..., X_p)^t$. De façon naturelle, la régression linéaire revient à supposer que

$$Y \sim \mathcal{N}or(\mu, \sigma^2) \text{ où } \mu = \boldsymbol{X}^t\boldsymbol{\beta}.$$

Ce modèle proposé par Legendre et Gauss au début du 19ème siècle, et étudié en détails par Fisher dans les années 20 s'est imposé en économétrie, mais s'avère difficilement utilisable en assurance.

Les variables que l'on cherche à modéliser en assurance sont des coûts (à valeur dans $\mathbb{R}^+$), des nombres de sinistres (à valeur dans $\mathbb{N}$) ou des indicatrices du fait d'être sinistré dans l'année (à valeur dans $\{0, 1\}$). Dans ce dernier cas, nous avions vu que les variables latentes pouvaient être une solution intéressante. Plus particulièrement, on considérait des modèles de la forme

$$Y \sim \mathcal{B}in(1, \mu) \text{ où } \mu = \mathbb{E}[Y] = F(\boldsymbol{X}^t\boldsymbol{\beta}),$$

où F désigne la fonction de répartition associée à la loi logistique (pour les modèles LOGIT) ou à la loi gaussienne centrée et réduite (pour les modèles PROBIT).

De façon générale, on souhaite garder la structure linéaire du score en $\boldsymbol{\beta}$, et considérer que l'espérance de Y est une transformation de cette combinaison linéaire. Plus précisement, on voudrait à présent passer à des modèles de régression du type

$$Y \sim \mathcal{L}oi(\mu) \text{ où } \mu = \mathbb{E}[Y] = g^{-1}(\boldsymbol{X}^t\boldsymbol{\beta}),$$

où g^{-1} est une fonction "*bien choisie*", et où $\mathcal{L}oi$ désigne une loi paramétrique permettant de modéliser correctement notre variable d'intérêt.

Ce type d'approche est à la base des modèles dits "*linéaires généralisés*", qui étendent le modèle gaussien à une famille de lois particulière, appelée famille exponentielle (naturelle).

9.8.2 Définition

Dans cette section, nous allons nous intéresser à la famille des modèles linéaires généralisés. Font partie de cette classe, en plus de

la loi normale, les lois de probabilité à deux paramètres θ et ϕ dont la densité (discrète ou continue) peut se mettre sous la forme

$$f(y|\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right), \quad y \in \mathcal{S}, \tag{9.31}$$

où le support $\mathcal{S}$ est un sous-ensemble de $\mathbb{N}$ ou de $\mathbb{R}$. Le paramètre θ est appelé paramètre naturel et ϕ est le paramètre de dispersion. Souvent, une pondération est nécessaire, et on remplace ϕ par ϕ/ω, où ω est un poids connu a priori .

Examinons quelques exemples de lois usuelles dont la densité peut se mettre sous la forme (**??**).

Exemple 9.8.1 (Loi normale)**.** *La loi normale $\mathcal{N}or(\mu,\sigma^2)$ possède une densité pouvant se mettre sous la forme* (**??**)*, avec $\mathcal{S} = \mathbb{R}$, $\theta = \mu$, $b(\theta) = \theta^2/2$, $\phi = \sigma^2$ et*

$$c(y,\phi) = -\frac{1}{2}\left(\frac{y^2}{\sigma^2} + \ln(2\pi\sigma^2)\right).$$

Exemple 9.8.2 (Loi de Poisson)**.** *Si on considère la loi de Poisson $\mathcal{P}oi(\lambda)$, on a*

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

d'où on tire $\mathcal{S} = \mathbb{N}$, $\theta = \ln\lambda$, $\phi = 1$, $b(\theta) = \exp\theta = \lambda$ et $c(y,\phi) = -\ln y!$.

Exemple 9.8.3 (Loi binomiale)**.** *La loi $\mathcal{B}in(n,p)$ possède une densité pouvant se mettre sous la forme* (**??**) *avec $\mathcal{S} = \mathbb{N}$, $\theta = \ln\{p/(1-p)\}$, $b(\theta) = n\ln(1+\exp(\theta))$, $\phi = 1$ et $c(y,\phi) = \ln\begin{pmatrix} n \\ y \end{pmatrix}$.*

Exemple 9.8.4 (Loi Gamma)**.** *La densité associée à la loi Gamma peut se réécrire*

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

qui peut se mettre sous la forme (**??**) *avec $\mathcal{S} = \mathbb{R}^+$, $\theta = -\frac{1}{\mu}$, $b(\theta) = -\ln(-\theta)$ et $\phi = \nu^{-1}$.*

Toutes les lois de probabilité dont la densité peut se mettre sous la forme (**??**) ne possèdent pas de paramètres de dispersion ϕ. Ainsi, les exemples ci-dessus nous apprennent par exemple que pour la

loi de Poisson, $\phi = 1$. Pour les lois possédant un paramètre de dispersion ϕ, celui-ci contrôle la variance, comme nous le verrons plus loin. La prime pure ne dépend quant à elle que du paramètre naturel θ. Ainsi, lorsque l'actuaire ne s'intéresse qu'à la prime pure, le paramètre θ est le paramètre d'intérêt tandis que ϕ est considéré comme un paramètre de nuisance. Toutefois, le paramètre ϕ est également fort important dans la mesure où il contrôle la dispersion (et donc le risque).

9.8.3 Moyenne et variance

Pour une variable aléatoire Y dont la densité peut se mettre sous la forme (**??**), on peut exprimer les deux premiers moments de Y à l'aide des fonctions b et c. Pour ce faire, notons

$$U = \frac{\partial}{\partial\theta} \ln f(Y|\theta,\phi).$$

et

$$U' = \frac{\partial^2}{\partial\theta^2} \ln f(Y|\theta,\phi),$$

de sorte que l'information de Fisher vaut $\mathbb{V}[U] = -\mathbb{E}[U']$ en vertu de (**??**).

Proposition 9.8.5. *Pour une variable aléatoire Y dont la densité est de la forme (**??**), on a*

$$\mathbb{E}[Y] = b'(\theta) \text{ et } \mathbb{V}[Y] = b''(\theta)\phi/\omega,$$

où $'$ et $''$ désignent les dérivées premières et secondes par rapport à θ.

Démonstration. Nous savons par la Propriété **??** que $\mathbb{E}[U] = 0$. Il suffit alors de remarquer que

$$\frac{d}{d\theta} \ln f(y|\theta,\phi) = \frac{\partial}{\partial\theta}\left(\frac{y\theta - b(\theta)}{\phi/\omega} + c(y,\phi)\right) = \frac{y - b'(\theta)}{\phi/\omega}$$

ce qui donne

$$\mathbb{E}[U] = \frac{\mathbb{E}[Y] - b'(\theta)}{\phi/\omega} = 0,$$

d'où l'expression annoncée pour la moyenne de Y. D'autre part, puisque $\mathbb{E}[U] = 0$,

$$\mathbb{V}[U] = \mathbb{E}[U^2] = \mathbb{E}\left[\left(\frac{Y - b'(\theta)}{\phi/\omega}\right)^2\right] = \frac{\mathbb{V}[Y]}{\{\phi/\omega\}^2}$$

et

$$\begin{aligned}\mathbb{E}[U^2] &= \int_{y\in\mathcal{S}} \left(\frac{\partial}{\partial\theta}\ln f(y|\theta,\phi)\right)^2 f(y|\theta,\phi)dy \\ &= \int_{y\in\mathcal{S}} \frac{\partial}{\partial\theta}\ln f(y|\theta,\phi)\frac{\partial}{\partial\theta}f(y|\theta,\phi)dy \\ &= \mathbb{E}\left[-\frac{\partial^2}{\partial\theta^2}\ln f(Y|\theta,\phi)\right] = \frac{b''(\theta)}{\phi/\omega}.\end{aligned}$$

Ainsi,

$$\mathbb{V}[U] = \mathbb{E}[-U'] = \frac{b''(\theta)}{\phi/\omega}.$$

En combinant les deux dernières égalités, on obtient le résultat annoncé. □

Dès lors, la variance de Y apparaît comme le produit de deux fonctions :

1. la première, $b''(\theta)$, qui dépend uniquement du paramètre θ est appelée fonction variance ;
2. la seconde est indépendante de θ et dépend uniquement de ϕ.

En notant $\mu = \mathbb{E}[Y]$, on voit que le paramètre θ est lié à la moyenne μ comme indiqué à la Proposition **??**. La fonction variance peut donc être définie en fonction de μ ; nous la noterons dorénavant $V(\mu)$.

La fonction variance est très importante dans les différents modèles, comme on peut le constater au Tableau **??**. Il est important de noter que, le cas de la loi normale mis à part, la variance de Y est toujours fonction de la moyenne et croît en fonction de cette dernière pour les lois de Poisson, Gamma et Inverse Gaussienne (à paramètre ϕ fixé).

Loi de probabilité	$V(\mu)$
Normale	1
Poisson	μ
Gamma	μ^2
Binomiale	$\mu(1-\mu)$

TABLE 9.1 – Fonctions variance associées aux lois de probabilité usuelles dont la densité est de la forme (**??**).

9.8.4 Modèle de régression

Considérons des variables aléatoires indépendantes mais non identiquement distribuées $Y_1, Y_2, \ldots, Y_n$ dont la densité est de la forme (**??**). Plus précisément, supposons que la densité de probabilité de Y_i est

$$f(y_i|\theta_i, \phi) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{\phi/\omega_i} + c(y_i, \phi)\right), \quad y_i \in \mathcal{S}. \tag{9.32}$$

Dès lors, la densité jointe de $Y_1, Y_2, \ldots, Y_n$ est

$$\begin{aligned} f(\boldsymbol{y}|\boldsymbol{\theta}, \phi) &= \prod_{i=1}^{n} f(y_i|\theta_i, \phi) \\ &= \exp\left(\frac{\sum_{i=1}^n y_i\theta_i - \sum_{i=1}^n b(\theta_i)}{\phi/\omega_i} + \sum_{i=1}^{n} c(y_i, \phi)\right). \end{aligned}$$

Bien entendu, la vraisemblance vaut $\mathcal{L}(\boldsymbol{\theta}, \phi|\boldsymbol{y}) = f(\boldsymbol{y}|\boldsymbol{\theta}, \phi)$. On suppose que les θ_i sont fonction d'un ensemble de $p+1$ paramètres $\beta_0, \beta_1, \ldots, \beta_p$, disons. Plus précisément, notant μ_i la moyenne de Y_i, on suppose que

$$g(\mu_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} = \boldsymbol{x}_i^t\boldsymbol{\beta} = \eta_i$$

où la fonction monotone et dérivable g est appelée fonction de lien, le vecteur $\boldsymbol{x}_i$ contient des variables explicatives relatives à l'individu i et le vecteur $\boldsymbol{\beta}$ contient les $p+1$ paramètres.

Ainsi, un modèle linéaire généralisé est composé de trois éléments, à savoir

(i) de variables à expliquer $Y_1, Y_2, \ldots, Y_n$ dont les densités sont de la forme (**??**) ;

(ii) d'un ensemble de paramètres $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^t$ appartenant à un ouvert non vide de $\mathbb{R}^{p+1}$ et des variables explicatives $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)^t$: la matrice $\boldsymbol{X}$ de dimension $n \times (p+1)$ est supposée être de rang $p+1$, i.e. la matrice carrée $\boldsymbol{X}^t\boldsymbol{X}$ de dimension $(p+1) \times (p+1)$ est inversible ;

(iii) d'une fonction de lien g telle que

$$g(\mu_i) = \boldsymbol{x}_i^t\boldsymbol{\beta} \text{ où } \mu_i = \mathbb{E}[Y_i]$$

qui lie le prédicteur linéaire $\eta_i = \boldsymbol{x}_i^t\boldsymbol{\beta}$ à la moyenne μ_i de Y_i.

La plupart du temps, les variables explicatives sont toutes catégorielles dans un tarif commercial. Considérons l'exemple donné à la Section **??** d'une compagnie segmentant selon le sexe, le caractère sportif du véhicule et l'âge de l'assuré (3 classes d'âges, à savoir moins de 30 ans, 30-65 ans et plus de 65 ans). Un assuré sera représenté par un vecteur binaire donnant les valeurs des variables ayant servi à coder les caractéristiques de l'individu.

On choisit comme niveau de référence (i.e. celui pour lequel tous les X_i valent 0) les modalités les plus représentées dans le portefeuille. Les résultats s'interprèteront ensuite comme une sur- ou sous-sinistralité par rapport à cette classe de référence. Ainsi, le vecteur (0,1,1,0) représente un assuré masculin de moins de 30 ans conduisant un vehicule sportif. Le prédicteur linéaire (ou score) sera de la forme $\beta_0+\sum_{j=1}^4 \beta_j X_j$ et le nombre ou le coût moyen de sinistre est en général une fonction non-décroissante du score. L'ordonnée à l'origine, ou intercept, β_0 représente donc le score associé à la classe de référence (i.e. celle pour laquelle $X_i = 0$ pour tout i, à savoir les hommes entre 30 et 65 ans dont le véhicule n'a pas de caractère sportif) ; si $\beta_j > 0$, cela indique que le fait de présenter la modalité traduite par X_j est un facteur agravant la sinistralité par rapport à celle de l'individu de référence, au contraire $\beta_j < 0$ indiquera les classes d'assurés moins risqués que les individus de référence.

Examinons à présent quelques exemples.

Exemple 9.8.6 (Régression gaussienne)**.** *Le modèle linéaire classique, pour lequel les $Y_i \sim \mathcal{N}or(\mu_i, \sigma^2)$ avec $\mu_i = \boldsymbol{x}_i^t\boldsymbol{\beta}$. La fonction de lien est ici l'identité. Ce modèle a été étudié en détails dans la Section* **??**.

Exemple 9.8.7 (Régression binomiale)**.** *La régression binomiale obtenue en considérant $Y_i \sim \mathcal{B}in(1, q_i)$. La quantité à expliquer q_i peut être par exemple la probabilité que la police i produise au moins un sinistre. Eu égard au fait que $q_i \in [0, 1]$, on utilisera la modélisation $q_i = F(\boldsymbol{x}_i^t\boldsymbol{\beta})$ où F est une fonction de répartition, ou de manière équivalente, $\boldsymbol{x}_i^t\boldsymbol{\beta} = F^{-1}(q_i)$. Même si théoriquement, tout fonction de répartition F pourrait être utilisée comme fonction de lien, on utilisera une des trois fonctions suivantes :*

1. *le modèle logit selon lequel*

$$logit(q_i) = \ln \frac{q_i}{1-q_i} = \boldsymbol{x}_i^t\boldsymbol{\beta}$$
$$\Leftrightarrow \quad q_i = \frac{\exp(\boldsymbol{x}_i^t\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}_i^t\boldsymbol{\beta})} = \frac{\exp\eta_i}{1+\exp\eta_i}.$$

2. le modèle probit selon lequel

$$probit(q_i) = \Phi^{-1}(q_i) = \boldsymbol{x}_i^t\boldsymbol{\beta}$$

$$\Leftrightarrow q_i = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\boldsymbol{x}_i^t\boldsymbol{\beta}} \exp\left(-\frac{z^2}{2}\right) dz.$$

3. le modèle log-log complémentaire selon lequel

$$cloglog(q_i) = \ln\Big(-\ln(1-q_i)\Big) = \boldsymbol{x}_i^t\boldsymbol{\beta}$$

$$\Leftrightarrow q_i = 1 - \exp\big(-\exp(\boldsymbol{x}_i^t\boldsymbol{\beta})\big).$$

Au contraire des fonctions logit et probit, la fonction log-log complémentaire n'est pas symétrique par rapport à 0.5. Pour de petites valeurs de q (comme c'est le plus souvent le cas en sciences actuarielles), il n'y a pratiquement pas de différence entre la transformation logistique et la transformation log-log complémentaire. L'utilisation de la transformation log-log complémentaire suppose que les probabilités de succès et d'échec doivent être traitées de manière différente.

On peut voir à la Figure **??** *les allures des 3 fonctions décrites ci-dessus. Toutefois, ce graphe ne renseigne pas parfaitement quant aux différences éventuelles entre les trois modèles de régression. En effet, il ne faut pas perdre de vue que le score est linéaire en* $\boldsymbol{\beta}$*. Donc, si on remplace la fonction de lien F par sa version standardisées*

$$F_{stand}(z) = F\left(\frac{z-\mu}{\sigma}\right)$$

où μ *et* σ^2 *désignent respectivement les moyennes et variances associées à F, alors*

$$q_i = F(\boldsymbol{x}_i^t\boldsymbol{\beta}) = F_{stand}(\boldsymbol{x}_i^t\widetilde{\boldsymbol{\beta}})$$

où $\widetilde{\beta}_0 = \mu + \sigma\beta_0$ *et* $\widetilde{\beta}_j = \sigma\beta_j$, $j = 1, 2 \ldots, p$. *Par conséquent les trois modèles présentés ci-dessus ne peuvent être comparés que pour des échelles appropriées pour le score, i.e. après centrage et réduction. Les espérances associées aux lois normale, logistique et Gumbel intervenant dans les liens probit, logit et cloglog valent 0, 0 et* $\Gamma'(1) = -0.5772$*, respectivement, où* Γ' *désigne la dérivée première de la fonction Gamma. Les variances associées valent quant à elles 1,* $\frac{\pi^2}{3}$ *et* $\frac{\pi^2}{6}$*. Après standardisation, les fonctions associées aux modèles logit et probit sont quasiment identiques. Cependant, dans le modèle cloglog, la probabilité tend plus vite vers 0 et 1 lorsque le score diverge vers* $-\infty$ *ou* $+\infty$*.*

FIGURE 9.10 – Fonctions de liens des modèles logit, probit et cloglog.

Exemple 9.8.8 (Régression de Poisson)**.** *La régression log-linéaire de Poisson est obtenue en considérant $Y_i \sim \mathcal{P}oi(\lambda_i)$, la fonction de lien étant celle induite par le paramètre naturel, i.e.*

$$\ln \lambda_i = \boldsymbol{x}_i^t \boldsymbol{\beta} \Leftrightarrow \lambda_i = \exp\left(\boldsymbol{x}_i^t \boldsymbol{\beta}\right).$$

Le plus souvent, on dispose d'une mesure de l'exposition au risque et on considère $Y_i \sim \mathcal{P}oi(d_i\lambda_i)$, où d_i est la durée de la couverture octroyée à l'assuré i (cette durée multiplie la fréquence annuelle λ_i sous l'hypothèse d'un processus de Poisson gouvernant la survenance des sinistres).

9.8.5 Fonction de lien canonique

Chacune des lois de probabilité de la famille exponentielle linéaire possède une fonction de lien spécifique, dite fonction de lien canonique, définie par $\theta = \eta$, où θ est le paramètre naturel. Le lien canonique est tel que $g(\mu_i) = \theta_i$. Or, $\mu_i = b'(\theta_i)$ d'où $g^{-1} = b'$. Les fonctions de lien canoniques sont reprises au Tableau **??**.

Loi de probabilité	Fonction de lien canonique
Normale	$\eta = \mu$
Poisson	$\eta = \ln \mu$
Gamma	$\eta = 1/\mu$
Binomiale	$\eta = \ln \mu - \ln(1-\mu)$

TABLE 9.2 – Liens canoniques associés aux lois de probabilité usuelles dont la densité est de la forme (**??**).

9.8.6 Equations de vraisemblance

En pratique, les coefficients de régression $\beta_0, \beta_1, \ldots, \beta_p$ et le paramètre de dispersion ϕ sont inconnus et doivent donc être estimés sur base des données. Dans cette section, nous nous concentrons sur l'estimation des coefficients de régression $\boldsymbol{\beta}$ par la méthode du maximum de vraisemblance. Il s'agit donc de maximiser la log-

vraisemblance

$$\begin{aligned} L\big(\boldsymbol{\theta}(\boldsymbol{\beta})|\boldsymbol{y},\phi\big) &= \sum_{i=1}^{n} \ln f(y_i|\theta_i,\phi) \\ &= \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{\phi/\omega_i} + \sum_{i=1}^{n} c(y_i,\phi) \end{aligned}$$

où $\mathbb{E}[Y_i] = b'(\theta_i) = \mu_i$ et $g(\mu_i) = \boldsymbol{x}_i^t\boldsymbol{\beta} = \eta_i$, avec g monotone et dérivable. Rechercher les estimateurs du maximum de vraisemblance revient à rechercher les $\beta_0, \beta_1, \ldots, \beta_p$ qui vérifient les équations

$$U_j = 0 \text{ pour } j = 0, 1, \ldots, p, \tag{9.33}$$

où

$$\begin{aligned} U_j &= \frac{\partial L\big(\boldsymbol{\theta}(\boldsymbol{\beta})|\boldsymbol{y},\phi\big)}{\partial\beta_j} \\ &= \sum_{i=1}^{n} \frac{\partial \ln f(y_i|\theta_i,\phi)}{\partial\beta_j} \\ &= \sum_{i=1}^{n} \frac{\partial}{\partial\beta_j}\left(\frac{y_i\theta_i - b(\theta_i)}{\phi/\omega_i} + c(y_i,\phi)\right). \end{aligned}$$

Afin d'obtenir U_j, on se sert de la formule

$$\frac{\partial \ln f(y_i|\theta_i,\phi)}{\partial\beta_j} = \frac{\partial \ln f(y_i|\theta_i,\phi)}{\partial\theta_i}\frac{\partial\theta_i}{\partial\mu_i}\frac{\partial\mu_i}{\partial\beta_j}.$$

Comme $\mu_i = b'(\theta_i)$, il vient

$$\begin{aligned} \frac{\partial \ln f(y_i|\theta_i,\phi)}{\partial\theta_i} &= \frac{y_i - b'(\theta_i)}{\phi/\omega_i} = \frac{y_i-\mu_i}{\phi/\omega_i}, \\ \frac{\partial\mu_i}{\partial\theta_i} &= b''(\theta_i), \end{aligned}$$

et

$$\frac{\partial\mu_i}{\partial\beta_j} = \frac{\partial\mu_i}{\partial\eta_i}\frac{\partial\eta_i}{\partial\beta_j} = x_{ij}\frac{\partial\mu_i}{\partial\eta_i}.$$

On obtient alors

$$\begin{aligned} \frac{\partial \ln f(y_i|\theta_i,\phi)}{\partial\beta_j} &= \frac{\frac{\partial \ln f(y_i|\theta_i,\phi)}{\partial\theta_i}\frac{\partial\mu_i}{\partial\beta_j}}{\frac{\partial\mu_i}{\partial\theta_i}} \\ &= \frac{(y_i-\mu_i)x_{ij}\frac{\partial\mu_i}{\partial\eta_i}}{\phi/\omega_i b''(\theta_i)} \\ &= \frac{(y_i-\mu_i)x_{ij}}{\mathbb{V}[Y_i]}\frac{\partial\mu_i}{\partial\eta_i}, \end{aligned}$$

et finalement

$$U_j = \sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{\mathbb{V}[Y_i]} \frac{\partial \mu_i}{\partial \eta_i} = \sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{\mathbb{V}[Y_i] g'(\mu_i)}.$$

Comme $\mathbb{V}[Y_i] = b''(\theta_i)\phi/\omega_i$,

$$U_j = 0 \Leftrightarrow \sum_{i=1}^{n} \omega_i (y_i - \mu_i) \frac{x_{ij}}{b''(\theta_i) g'(\mu_i)} = 0$$

où le paramètre ϕ n'apparaît plus. Les équations de vraisemblance relatives à $\boldsymbol{\beta}$ peuvent donc être résolues sans se préoccuper de ϕ.

Notez que si on choisit la fonction de lien canonique, les équations de vraisemblance deviennent

$$\sum_{i=1}^{n} \omega_i (y_i - \mu_i) x_{ij} = 0 \text{ pour } j = 0, 1, \ldots, p.$$

Il est intéressant de noter que cette relation traduit l'orthogonalité entre les variables explicatives et les résidus (en tout point semblable à celle obtenue dans le modèle linéaire).

Exemple 9.8.9 (Régression logistique)**.** *Supposons disposer des réalisation $y_1, \ldots, y_m$ de variables aléatoires de loi $\mathcal{B}in(n_i, q_i)$. Afin d'estimer les paramètres $\beta_0, \beta_1, \ldots, \beta_p$, on maximise la vraisemblance*

$$\mathcal{L}(\boldsymbol{\beta}|\boldsymbol{y}) \;=\; \prod_{i=1}^{m} \begin{pmatrix} n_i \\ y_i \end{pmatrix} \left(\frac{\exp \eta_i}{1 + \exp \eta_i} \right)^{y_i} \left(\frac{1}{1 + \exp \eta_i} \right)^{n_i - y_i}.$$

Les équations de vraisemblance s'écrivent donc

$$\sum_{i=1}^{n} x_{ij}(y_i - \widehat{y}_i) = 0, \;\; j = 0, 1, \ldots, p, \tag{9.34}$$

où

$$\widehat{y}_i = n_i \widehat{q}_i = n_i \frac{\exp(\widehat{\eta}_i)}{1 + \exp(\widehat{\eta}_i)}.$$

Notez que (**??**) *s'interprète comme une relation d'orthogonalité entre les résidus $y_i - \widehat{y}_i$ et les variables explicatives $\boldsymbol{x}_i$.*

Exemple 9.8.10 (Régression de Poisson)**.** *Supposons disposer des réalisations $n_1, n_2, \ldots, n_n$ de variables aléatoires de loi $\mathcal{P}oi(d_i\lambda_i)$, la log-vraisemblance est donnée par*

$$L(\boldsymbol{\beta}|\boldsymbol{n}) = \ln \mathcal{L}(\boldsymbol{\beta}|\boldsymbol{n}) = \sum_{i=1}^{n} \Big(- \ln n_i! + n_i(\eta_i + \ln d_i) - \lambda_i \Big).$$

Les équations de vraisemblance s'écrivent donc

$$\sum_{i=1}^{n} n_i = \sum_{i=1}^{n} \lambda_i \tag{9.35}$$

et pour $j = 1, 2, \ldots, p$,

$$\sum_{i=1}^{n} x_{ij}\Big(n_i - \lambda_i\Big) = 0. \tag{9.36}$$

Les équations (**??**) *s'interprètent comme une relation d'orthogonalité entre les variables explicatives* $\boldsymbol{x}_i$ *et les résidus d'estimation.*

Comme les facteurs de risque en pratique ont le plus souvent un nombre fini de niveaux et que les variables explicatives sont les indicatrices de ces niveaux, les équations de vraisemblance (**??**) *ont une signification tarifaire très importante. Elles garantissent que pour chaque sous-portefeuille correspondant à un niveau d'un des facteurs de risque, le nombre total des sinistres observés est égal à son homologue théorique. En effet, supposons par exemple que* $x_{i1} = 1$ *si l'individu* i *est un homme, et 0 sinon ;* (**??**) *pour* $j = 1$ *garantit alors que*

$$\sum_{hommes} n_i = \sum_{hommes} \widehat{\lambda}_i.$$

De plus, en vertu de (**??**) *la somme des fréquences prédites* $\widehat{\lambda}_i$ *est égale au nombre total de sinistres déclarés, puisque*

$$\sum_{i=1}^{n} \widehat{\lambda}_i = \sum_{i=1}^{n} n_i$$

pour autant qu'un intercept β_0 *soit inclus dans le score* η_i.

Bien entendu, ces interprétations ne valent que si on appliquait le modèle aux données ayant servi à son estimation. En pratique, si le portefeuille est suffisamment stable, on espère avoir des égalités approximatives lorsque le modèle sera appliqué dans le futur.

Exemple 9.8.11 (Régression Gamma)**.** *Notons* n_i *le nombre de sinistres causés par l'assuré* i*, et, lorsque* $n_i > 0$*, désignons par* $c_{i1}, c_{i2}, \ldots, c_{in_i}$ *les coûts de ceux-ci. Si nous considérons* $c_{i1}, c_{i2}, \ldots, c_{in_i}$ *comme des réalisations de variables aléatoires indépendantes et de même loi Gamma de moyenne* μ_i *et de variance*

μ_i^2/ν, *la vraisemblance s'écrit*

$$\begin{aligned}\mathcal{L}(\boldsymbol{\beta}|\boldsymbol{c}) &= \prod_{i|n_i>0}\prod_{k=1}^{n_i} f(c_{ik}|\boldsymbol{x}_i,\nu)\\ &= \prod_{i|n_i>0}\prod_{k=1}^{n_i}\left(\frac{1}{\Gamma(\nu)}\left(\frac{\nu c_{ik}}{\mu_i}\right)^{\nu}\exp\left(-\frac{\nu c_{ik}}{\mu_i}\right)\frac{1}{c_{ik}}\right).\end{aligned}$$

Les équations de vraisemblance à résoudre pour obtenir $\widehat{\boldsymbol{\beta}}$ *sont les suivantes :*

$$\frac{\partial}{\partial\beta_j}L(\boldsymbol{\beta}|\boldsymbol{c}) = \frac{\partial}{\partial\beta_j}\sum_{i|n_i>0}\sum_{k=1}^{n_i}\left(-\nu\ln\mu_i - \frac{\nu c_{ik}}{\mu_i}\right) = 0$$

ou encore

$$\sum_{i|n_i>0}\sum_{k=1}^{n_i} x_{ij}\left(1-\frac{c_{ik}}{\mu_i}\right) = 0.$$

Si on définit $\widehat{c}_i = \widehat{\mu}_i = \exp(\widehat{\boldsymbol{\beta}}^t\boldsymbol{x}_i)$ *le coût moyen d'un sinistre pour l'assuré* i*, l'estimateur du maximum de vraisemblance* $\widehat{\boldsymbol{\beta}}$ *de* $\boldsymbol{\beta}$ *est solution des équations*

$$\sum_{i|n_i>0}\underbrace{\left(n_i - \frac{c_{i\bullet}}{\widehat{c}_i}\right)}_{=\textit{coût résiduel}}\boldsymbol{x}_i = \mathbf{0}.$$

Nous avons donc également une relation d'orthogonalité entre variables explicatives et résidus.

9.8.7 Résolution des équations de vraisemblance

Les estimateurs du maximum de vraisemblance $\hat{\beta}_j$ des paramètres β_j sont solutions du système (**??**). Les équations composant ce système ne possèdent en général pas de solution explicite et doivent dès lors être résolues numériquement. On peut par exemple utiliser la méthode de Newton-Raphson, que nous rappelons brièvement ci-dessous.

Notons $\boldsymbol{U}(\boldsymbol{\beta})$ le vecteur gradient de la log-vraisemblance, dont la composante j est

$$U_j(\boldsymbol{\beta}) = \frac{\partial}{\partial\beta_j}L(\boldsymbol{\beta}|\boldsymbol{y})$$

et notons $\boldsymbol{H}(\boldsymbol{\beta})$ la matrice hessienne de $L(\boldsymbol{\beta}|\boldsymbol{y})$, i.e. celle dont l'élément (j,k) est

$$\frac{\partial^2}{\partial\beta_j\partial\beta_k}L(\boldsymbol{\beta}|\boldsymbol{y}).$$

Pour $\boldsymbol{\beta}^*$ proche de $\widehat{\boldsymbol{\beta}}$, un développement de Taylor limité donne

$$0=\boldsymbol{U}(\widehat{\boldsymbol{\beta}})\approx\boldsymbol{U}(\boldsymbol{\beta}^*)+\boldsymbol{H}(\boldsymbol{\beta}^*)\Big(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\Big)$$

qui permet d'écrire

$$\boldsymbol{U}(\boldsymbol{\beta}^*)+\boldsymbol{H}(\boldsymbol{\beta}^*)\Big(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\Big)\approx 0$$

ou encore

$$\widehat{\boldsymbol{\beta}}\approx\boldsymbol{\beta}^*-\boldsymbol{H}^{-1}(\boldsymbol{\beta}^*)\boldsymbol{U}(\boldsymbol{\beta}^*). \tag{9.37}$$

Ceci suggère une procédure itérative pour obtenir l'estimateur du maximum de vraisemblance $\widehat{\boldsymbol{\beta}}$ de $\boldsymbol{\beta}$: partant d'une valeur initiale $\widehat{\boldsymbol{\beta}}^{(0)}$ que l'on espère proche de $\widehat{\boldsymbol{\beta}}$, on définit la $(r+1)$-ème valeur approchée $\widehat{\boldsymbol{\beta}}^{(r+1)}$ de $\widehat{\boldsymbol{\beta}}$ à partir de la $r$ème $\widehat{\boldsymbol{\beta}}^{(r)}$ par

$$\widehat{\boldsymbol{\beta}}^{(r+1)}\approx\widehat{\boldsymbol{\beta}}^{(r)}-\boldsymbol{H}^{-1}(\widehat{\boldsymbol{\beta}}^{(r)})\boldsymbol{U}(\widehat{\boldsymbol{\beta}}^{(r)}). \tag{9.38}$$

Cette procédure itérative pour obtenir l'estimateur du maximum de vraisemblance correspond à la méthode de Newton-Raphson.

Remarque 9.8.12. *Signalons une méthode astucieuse de résolution itérative des équations de vraisemblance. A l'étape r, il suffit de minimiser un critère des moindres carrés pondérés du type*

$$\sum_{k=1}^{n}w_k(z_k-\boldsymbol{x}_k^t\boldsymbol{\beta})^2$$

où les pseudo-réponses z_k sont données par

$$z_k=\boldsymbol{x}_k^t\boldsymbol{\beta}^{(r)}+(y_k-\mu_k)\frac{\partial\eta_k}{\partial\mu_k}$$

et les poids

$$w_k^{-1}=\left(\frac{\partial\eta_k}{\partial\mu_k}\right)^2V(\mu_k).$$

Dans ces formules, μ_k et η_k sont calculées pour les valeurs courantes $\boldsymbol{\beta}^{(r)}$ du paramètre $\boldsymbol{\beta}$. On stoppera la procédure lorsque la différence entre $\boldsymbol{\beta}^{(r)}$ et $\boldsymbol{\beta}^{(r-1)}$ est suffisamment petite.

Exemple 9.8.13 (Régression binomiale)**.** *Si les observations y_i sont de loi $\mathcal{B}in(n_i, q_i)$, $i = 1, \ldots, n$, nous avons*

$$U_j(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{y_i - n_i q_i}{q_i(1-q_i)} \frac{x_{ij}}{g'(q_i)}.$$

Dans ce cas, $\widehat{\boldsymbol{\beta}}^{(r+1)}$ est obtenu par une régression linéaire ordinaire de la variable aléatoire dépéndante $\boldsymbol{z}_r$, dont le $i$ème élément vaut

$$\widehat{\eta}_{ir} + \frac{(y_i - n_i \widehat{q}_{ir}) g'(\widehat{q}_{ir})}{n_i}$$

en fonction des p variables explicatives, à l'aide des poids v_{ir} où

$$v_{ir} = \frac{n_i}{\widehat{q}_{ir}(1-\widehat{q}_{ir})\big(g'(\widehat{q}_{ir})\big)^2}$$

où $\widehat{q}_{ir}$ et $\widehat{\eta}_{ir}$ sont les probabilités de succès et le prédicteur linéaire relatifs à l'observation i, évalué grâce à la $r$ème itération $\widehat{\boldsymbol{\beta}}^{(r)}$.

Afin de lancer le processus itératif, on utilise les estimateurs initiaux des q_i donnés par $\widehat{q}_{i0} = \frac{y_i + 0.5}{n_i + 1}$, les poids initiaux $v_{ir} = \frac{n_i}{\widehat{q}_{i0}(1-\widehat{q}_{i0})\big(g'(\widehat{q}_{i0})\big)^2}$ et les valeurs initiales des pseudo-variables $\boldsymbol{z}$ $z_{i0} = \widehat{\eta}_{i0} = g(\widehat{q}_{i0})$. Les valeurs des z_{i0} sont alors régressées en fonction des p variables explicatives à l'aide d'une méthode des moindres carrés pondérés (les poids sont les v_{i0}). Les coefficients des variables explicatives sont les composantes de $\widehat{\boldsymbol{\beta}}^{(1)}$. On obtient alors

$$\begin{aligned}
\widehat{\eta}_{i1} &= \widehat{\boldsymbol{\beta}}^{(1)} \boldsymbol{x}_i \\
\widehat{q}_{i1} &= g^{-1}(\widehat{\eta}_{i1}) \\
v_{i1} &= \frac{n_i}{\widehat{q}_{i1}(1-\widehat{q}_{i1})\big(g'(\widehat{q}_{i1})\big)^2} \\
z_{i1} &= \widehat{\eta}_{i1} + \frac{(y_i - n_i \widehat{q}_{i1}) g'(\widehat{q}_{i1})}{n_i}.
\end{aligned}$$

Ensuite, $\widehat{\boldsymbol{\beta}}^{(2)}$ résulte d'une régression des z_{i1} en fonction des variables explicatives en tenant compte des pondérations v_{i1}, et ainsi de suite. On stoppe le processus lorsque la différence entre $\widehat{\boldsymbol{\beta}}^{(r)}$ et $\widehat{\boldsymbol{\beta}}^{(r+1)}$ est suffisamment petite.

Dans le cas particulier de la régression logistique, les pseudo-observations sont

$$z_i = \widehat{\eta}_i + \frac{y_i - n_i \widehat{q}_i}{n_i \widehat{q}_i (1-\widehat{q}_i)}$$

et les poids sont données quant à eux par

$$v_i = n_i q_i (1 - q_i).$$

Notez que les poids ne sont autres que la variance des Y_i.

Exemple 9.8.14 (Régression de Poisson)**.** *Si les observations* n_i *sont de loi* $\mathcal{P}oi(\lambda_i)$*, le vecteur gradient de* $L(\boldsymbol{\beta}|\boldsymbol{n})$*, de dimension* $p+1$*, est donné par*

$$\boldsymbol{U}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{x}_i (n_i - \lambda_i) \text{ avec } \lambda_i = d_i \exp(\boldsymbol{\beta}^t \boldsymbol{x}_i)$$

où l'on a ajouté au vecteur $\boldsymbol{x}_i$ *une composante* $x_{i0} = 1$*. La matrice hessienne, de dimension* $(p+1) \times (p+1)$ *est donnée par*

$$\boldsymbol{H}(\boldsymbol{\beta}) = -\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^t \lambda_i = -\boldsymbol{X}^t diag(\boldsymbol{\lambda}) \boldsymbol{X}$$

où $diag(\boldsymbol{\lambda})$ *désigne la matrice diagonale de dimension* $n \times n$ *dont les éléments principaux sont* $\lambda_1, \ldots, \lambda_n$*.*

La procédure itérative pour obtenir l'estimateur du maximum de vraisemblance $\widehat{\boldsymbol{\beta}}$ *de* $\boldsymbol{\beta}$ *est la suivante : partant d'une valeur initiale* $\widehat{\boldsymbol{\beta}}^{(0)}$ *que l'on espère proche de* $\widehat{\boldsymbol{\beta}}$*, on définit la* $(r+1)$*-ème valeur approchée* $\widehat{\boldsymbol{\beta}}^{(r+1)}$ *de* $\widehat{\boldsymbol{\beta}}$ *à partir de la* r*ème* $\widehat{\boldsymbol{\beta}}^{(r)}$ *par*

$$\widehat{\boldsymbol{\beta}}^{(r+1)} = \widehat{\boldsymbol{\beta}}^{(r)} + \left(\boldsymbol{X}^t diag(\widehat{\boldsymbol{\lambda}}^{(r)}) \boldsymbol{X}\right)^{-1} \boldsymbol{X}^t (\boldsymbol{n} - \widehat{\boldsymbol{\lambda}}^{(r)}). \quad (9.39)$$

Une bonne valeur initiale $\widehat{\boldsymbol{\beta}}^{(0)}$ *est obtenue en prenant* $\widehat{\beta}_0^{(0)} = \ln \overline{n}$*, où* $\overline{n}$ *est le nombre moyen de sinistres par police, et* $\widehat{\beta}_j^{(0)} = 0$ *pour* $j = 1, \ldots, p$*. Notez que* $\widehat{\boldsymbol{\beta}}^{(0)}$ *correspond en fait au modèle de Poisson homogène.*

L'algorithme itératif fournissant l'estimateur du maximum de vraisemblance de $\boldsymbol{\beta}$ *dans le modèle de Poisson peut encore s'écrire*

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}^{(r+1)} &= \widehat{\boldsymbol{\beta}}^{(r)} + \left(\sum_{i=1}^{n} \widehat{\lambda}_i^{(r)} \boldsymbol{x}_i \boldsymbol{x}_i^t\right)^{-1} \sum_{i=1}^{n} \boldsymbol{x}_i (n_i - \widehat{\lambda}_i^{(r)}) \\
&= \left(\sum_{i=1}^{n} \left(\sqrt{\widehat{\lambda}_i^{(r)}} \boldsymbol{x}_i\right) \left(\sqrt{\widehat{\lambda}_i^{(r)}} \boldsymbol{x}_i\right)^t\right)^{-1} \\
&\quad \sum_{i=1}^{n} \left(\sqrt{\widehat{\lambda}_i^{(r)}} \boldsymbol{x}_i\right) \left(\sqrt{\widehat{\lambda}_i^{(r)}} \frac{n_i - \widehat{\lambda}_i^{(r)}}{\widehat{\lambda}_i^{(r)}} + \sqrt{\widehat{\lambda}_i^{(r)}} \boldsymbol{x}_i^t \widehat{\boldsymbol{\beta}}^{(r)}\right).
\end{aligned}$$

En comparant la relation de récurrence ci-dessus à (**??**)*, on constate que* $\widehat{\boldsymbol{\beta}}^{(r+1)}$ *n'est autre que l'estimateur des moindres carrés associé au modèle de régression linéaire*

$$\sqrt{\widehat{\lambda}_i^{(r)}}\left(\frac{n_i-\widehat{\lambda}_i^{(r)}}{\widehat{\lambda}_i^{(r)}}+\boldsymbol{x}_i^t\widehat{\boldsymbol{\beta}}^{(r)}\right)=\left(\sqrt{\widehat{\lambda}_i^{(r)}}\boldsymbol{x}_i\right)^t\widehat{\boldsymbol{\beta}}^{(r+1)}+\epsilon_i$$

où ϵ_i *est un terme d'erreur gaussien centré. L'estimateur du maximum de vraisemblance du paramètre* $\boldsymbol{\beta}$ *peut donc être obtenu à l'aide d'une méthode des moindres carrés itérative.*

De manière équivalente, $\widehat{\boldsymbol{\beta}}$ *peut être obtenu grâce à un ajustement des moindres carrés pondérés des pseudo-variables*

$$z_i^{(r)}=\frac{n_i-\widehat{\lambda}_i^{(r)}}{\widehat{\lambda}_i^{(r)}}+\boldsymbol{x}_i^t\widehat{\boldsymbol{\beta}}^{(r)}$$

sur $\boldsymbol{x}_i$*, où les poids* $\sqrt{\widehat{\lambda}_i^{(r)}}$ *changent à chaque itération.*

9.8.8 Information de Fisher

L'élément (jk) de la matrice d'information de Fisher $\mathcal{I}$ est donné par $(\mathcal{I})_{jk}=\mathbb{E}[U_jU_k]$. La contribution de chaque observation Y_i à $(\mathcal{I})_{jk}$ est

$$\begin{aligned}\mathbb{E}\left[\frac{\partial\ln f_{\theta_i}(Y_i)}{\partial\beta_j}\frac{\partial\ln f_{\theta_i}(Y_i)}{\partial\beta_k}\right] &= \mathbb{E}\left[\frac{(Y_i-\mu_i)^2x_{ij}x_{ik}}{(\mathbb{V}[Y_i])^2}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2\right]\\ &= \frac{x_{ij}x_{ik}}{\mathbb{V}[Y_i]}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2.\end{aligned}$$

Dès lors,

$$(\mathcal{I})_{jk} = \sum_{i=1}^n\frac{x_{ij}x_{ik}}{\mathbb{V}[Y_i]}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2. \tag{9.40}$$

Exemple 9.8.15 (Régression binomiale)**.** *Si les observations* y_i *sont de loi* $\mathcal{B}in(n_i,q_i)$*,* $i=1,\ldots,n$*, l'élément* (j,k) *de la matrice d'information de Fisher s'écrit*

$$\mathcal{I}_{jk}(\boldsymbol{\beta}) = \mathbb{E}\left[\sum_{i=1}^n\frac{Y_i-n_iq_i}{q_i(1-q_i)}\frac{x_{ij}}{g'(q_i)}\sum_{\ell=1}^n\frac{Y_\ell-n_\ell q_\ell}{q_\ell(1-q_\ell)}\frac{x_{k\ell}}{g'(q_\ell)}\right].$$

A présent, notons que

$$\mathbb{E}\Big[(Y_i - n_i q_i)(Y_\ell - n_\ell q_\ell)\Big] = \mathbb{C}[Y_i, Y_\ell] = 0 \text{ pour } i \neq \ell$$

puisque les observations ont été supposées indépendantes, alors que lorsque $i = \ell$ on obtient

$$\mathbb{E}\Big[(Y_i - n_i q_i)^2\Big] = \mathbb{V}[Y_i] = n_i q_i (1 - q_i)$$

d'où l'on tire

$$\mathcal{I}_{jk}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{n_i}{q_i(1-q_i)\big(g'(q_i)\big)^2} x_{ij} x_{ik}.$$

Exemple 9.8.16 (Régression de Poisson)**.** *Si les observations n_i sont de loi $\mathcal{P}oi(\lambda_i)$,*

$$\begin{aligned}
\mathbb{C}[U_j, U_k] &= \mathbb{E}[U_j U_k] \\
&= \mathbb{E}\left[\sum_{i_1=1}^{n} x_{i_1 j}(N_{i_1} - \lambda_{i_1}), \sum_{i_2=1}^{n} x_{i_2 k}(N_{i_2} - \lambda_{i_2})\right] \\
&= \sum_{i_1=1}^{n}\sum_{i_2=1}^{n} x_{i_1 j} x_{i_2 k} \underbrace{\mathbb{C}[N_{i_1}, N_{i_2}]}_{=0 \text{ si } i_1 \neq i_2} \\
&= \sum_{i=1}^{n} x_{ij} x_{ik} \mathbb{V}[N_i] = \sum_{i=1}^{n} x_{ij} x_{ik} \lambda_i.
\end{aligned}$$

La matrice d'information de Fisher $\mathcal{I}$ est donc donnée par

$$\mathcal{I} = \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^t \lambda_i.$$

Si la matrice $\boldsymbol{X}$ est de rang $p+1$ alors $\boldsymbol{H}$ est non-singulière et, qui plus est, définie négative. Ceci garantit que la solution des équations de vraisemblance correspond à un maximum de $L(\boldsymbol{\beta})$.

Remarque 9.8.17. *Lorsqu'on dispose de nombreuses observations,*

$$\boldsymbol{H}(\boldsymbol{\beta}) \approx \mathbb{E}[\boldsymbol{H}(\boldsymbol{\beta})] = -\mathcal{I}(\boldsymbol{\beta}),$$

*de sorte qu'une alternative à (**??**) est donnée par*

$$\widehat{\boldsymbol{\beta}}_{r+1} \approx \widehat{\boldsymbol{\beta}}_r + \mathcal{I}^{-1}(\widehat{\boldsymbol{\beta}}_r)\boldsymbol{U}(\widehat{\boldsymbol{\beta}}_r). \tag{9.41}$$

Ce second schéma itératif est appelé méthode de Fisher ("Fisher's method of Scoring", en anglais).

Notons que $\mathbb{E}[\boldsymbol{H}] = -\mathbb{E}[\boldsymbol{U}\boldsymbol{U}^t] = -\mathcal{I}$. *Ceci nous permet de justifier l'interprétation de* $\mathcal{I}$ *en termes de quantité d'information. En effet, si* $\boldsymbol{H}$, *et donc* $\mathcal{I}$, *est petite, la vraisemblance aura une courbure faible, et la détermination de l'estimateur du maximum de vraisemblance s'en trouvera moins aisée.*

9.8.9 Intervalle de confiance pour les paramètres

Méthode du rapport de vraisemblance

La méthode du rapport de vraisemblance est basée sur le profil de vraisemblance, défini pour le paramètre β_j comme la fonction

$$\mathcal{L}_j(\beta_j|\boldsymbol{y}) = \max_{\beta_0,\ldots,\beta_{j-1},\beta_{j+1},\ldots,\beta_p} \mathcal{L}(\boldsymbol{\beta}|\boldsymbol{y}).$$

Si $\widehat{\boldsymbol{\beta}}_{\mathrm{MV}}$ est l'estimateur du maximum de vraisemblance de $\boldsymbol{\beta}$, $2\big\{L(\widehat{\boldsymbol{\beta}}_{\mathrm{MV}}|\boldsymbol{y}) - L_j(\beta_j|\boldsymbol{y})\big\}$ est approximativement de loi du khi-deux à un degré de liberté, pour autant que β_j soit la vraie valeur du paramètre, où $L_j(\beta_j|\boldsymbol{y}) = \ln \mathcal{L}_j(\beta_j|\boldsymbol{y})$. Dès lors, un intervalle de confiance au niveau $1-\alpha$ pour β_j est fourni par l'ensemble des valeurs ξ telles que la différence $L(\widehat{\boldsymbol{\beta}}_{\mathrm{MV}}|\boldsymbol{y}) - L_j(\xi|\boldsymbol{y})$ est suffisamment petite, ou encore telles que $2\big\{L(\widehat{\boldsymbol{\beta}}_{\mathrm{MV}}|\boldsymbol{y}) - L_j(\xi|\boldsymbol{y})\big\} \leq \chi^2_{1-\alpha,1}$, i.e.

$$IC = \left\{\xi \in \mathbb{R}\big| L_j(\xi|\boldsymbol{y}) \geq L(\widehat{\boldsymbol{\beta}}_{\mathrm{MV}}|\boldsymbol{y}) - \frac{1}{2}\chi^2_{1-\alpha,1}\right\}.$$

Les extrémités de cet intervalle sont obtenues numériquement en approximant la fonction de vraisemblance par une surface de degré 2. Spécifiquement, nous recourons à l'approximation

$$L(\boldsymbol{\beta}|\boldsymbol{y}) \approx L(\boldsymbol{\beta}_0|\boldsymbol{y}) + (\boldsymbol{\beta}-\boldsymbol{\beta}_0)^t\boldsymbol{U}(\boldsymbol{\beta}) + \frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^t\boldsymbol{H}(\boldsymbol{\beta})(\boldsymbol{\beta}-\boldsymbol{\beta}_0)$$

qui devrait être de bonne qualité pour $\boldsymbol{\beta}_0$ suffisamment proche de $\boldsymbol{\beta}$. En approximant $\boldsymbol{H}(\boldsymbol{\beta})$ par son espérance mathématique $-\mathcal{I}$ on obtient encore

$$L(\boldsymbol{\beta}|\boldsymbol{y}) \approx L(\boldsymbol{\beta}_0|\boldsymbol{y}) + (\boldsymbol{\beta}-\boldsymbol{\beta}_0)^t\boldsymbol{U}(\boldsymbol{\beta}) - \frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^t\mathcal{I}(\boldsymbol{\beta}-\boldsymbol{\beta}_0).$$

Méthode de Wald

Grâce à l'approximation normale (**??**) pour $\widehat{\boldsymbol{\beta}}$, un intervalle de confiance au niveau de confiance $1-\alpha$ pour β_j est donné par

$$\left[\hat{\beta}_j \pm z_{\alpha/2}\sqrt{v_{jj}}\right]$$

où v_{jj} est l'élément diagonal (jj) de $\mathcal{I}^{-1}$. Cet intervalle de confiance est souvent appelé intervalle de Wald. Les éléments de la diagonale de $\mathcal{I}^{-1}$ traduisent donc la précision des estimations ponctuelles $\hat{\beta}_j$, tandis que les éléments hors diagonale estiment les covariances existant entre les estimateurs des β_j.

9.8.10 Comparaison de modèles

L'ajustement du modèle à un jeu de données revient à substituer les moyennes μ_i aux observations initiales y_i. Il est clair qu'en général $y_i \neq \mu_i$. Il faut alors se demander si les écarts entre les y_i et les μ_i traduisent une mauvaise spécification du modèle ou peuvent être attribués au hasard. Pour ce faire, on pourrait avoir recours aux différentes statistiques suivantes.

La déviance

Nous considérons ici des modèles linéaires généralisés basés sur la même densité (**??**), possédant la même fonction de lien mais qui diffèrent par le nombre de paramètres qu'ils utilisent.

On définit la qualité d'un modèle en prenant comme référence le modèle saturé comptant autant de paramètres que d'observations et fournissant donc une description parfaite des données. Le modèle saturé est caractérisé par $\widehat{\mu}_i = y_i$ pour $i = 1, 2, \ldots, n$; la vraisemblance associée à ce modèle sera dorénavant notée $\mathcal{L}(\boldsymbol{y}|\boldsymbol{y})$. En pratique, ce modèle n'est pas intéressant puisqu'il se borne à reproduire les observations, sans les résumer. Considérons un modèle pour lequel le paramètre $\boldsymbol{\beta}$ est de dimension $p+1 < n$. Le modèle décrira bien les données lorsque $\mathcal{L}(\widehat{\boldsymbol{\mu}}|\boldsymbol{y}) \approx \mathcal{L}(\boldsymbol{y}|\boldsymbol{y})$ et mal lorsque $\mathcal{L}(\widehat{\boldsymbol{\mu}}|\boldsymbol{y}) <<< \mathcal{L}(\boldsymbol{y}|\boldsymbol{y})$. Ceci suggère la statistique du rapport de vraisemblance

$$\Lambda = \frac{\mathcal{L}(\boldsymbol{y}|\boldsymbol{y})}{\mathcal{L}(\widehat{\boldsymbol{\mu}}|\boldsymbol{y})}$$

comme mesure de la qualité de l'ajustement des données par un modèle ou, de manière équivalente,

$$\ln \Lambda = \ln \mathcal{L}(\boldsymbol{y}|\boldsymbol{y}) - \ln \mathcal{L}(\widehat{\boldsymbol{\mu}}|\boldsymbol{y}).$$

Une grande valeur de $\ln \Lambda$ laisse à penser que le modèle est de piètre qualité. Définissons la statistique $D = 2 \ln \Lambda$; celle-ci est appelée la déviance réduite dans le cadre des modèles linéaires généralisés. La déviance non réduite est elle donnée par $D^* = \phi D$.

On évalue souvent la qualité de l'ajustement fourni par un modèle par la déviance, notamment en raison des liens étroits que noue cette statistique avec celle du rapport de vraisemblance. Une petite valeur de la déviance indique un ajustement de bonne qualité, puisque la vraisemblance du modèle est proche de celle du modèle saturé. Au contraire, une grande valeur de la déviance traduira un piètre ajustement.

Si le modèle décrit bien les données observées, D est approximativement de loi χ^2_{n-p-1}. Une valeur observée D_{obs} "trop grande" suggère un modèle de mauvaise qualité. En pratique, on jugera le modèle de mauvaise qualité si

$$D_{obs} > \chi^2_{n-p-1;1-\alpha},$$

où $\chi^2_{n-p-1;1-\alpha}$ est le quantile d'ordre $1-\alpha$ de la loi khi-deux à $n-p-1$ degrés de liberté.

Exemple 9.8.18 (Régression Binomiale)**.** *Si les observations y_i sont de loi $\mathcal{B}in(n_i, q_i)$, la log-vraisemblance s'écrit*

$$\sum_{i=1}^{n} \left(\ln \left(\begin{array}{c} n_i \\ y_i \end{array} \right) + y_i \ln \widehat{q}_i + (n_i - y_i) \ln(1 - \widehat{q}_i) \right).$$

Dans le modèle saturé, les q_i sont estimés par y_i/n_i de sorte que la log-vraisemblance vaut

$$\sum_{i=1}^{n} \left(\ln \left(\begin{array}{c} n_i \\ y_i \end{array} \right) + y_i \ln \frac{y_i}{n_i} + (n_i - y_i) \ln \frac{n_i - y_i}{n_i} \right).$$

Ainsi, la déviance vaut

$$\begin{aligned} D &= 2 \sum_{i=1}^{n} \left(y_i \ln \frac{y_i/n_i}{\widehat{q}_i} + (n_i - y_i) \ln \frac{1 - y_i/n_i}{1 - \widehat{q}_i} \right) \\ &= 2 \sum_{i=1}^{n} \left(y_i \ln \frac{y_i}{\widehat{y}_i} + (n_i - y_i) \ln \frac{n_i - y_i}{n_i - \widehat{y}_i} \right) \end{aligned}$$

où $\widehat{y}_i = n_i \widehat{q}_i$ est la valeur prédite par le modèle courant pour l'observation i.

Il est important de noter que l'approximation khi-carrée pour D peut être assez mauvaise lorsque certains n_i sont petits et les probabilités ajustées $\widehat{q}_i$ proches de 0 ou de 1.

Remarque 9.8.19. *Dans le cas particulier où les observations y_i sont de loi $\mathcal{B}er(q_i)$, la déviance ne permet pas de juger de la qualité du modèle. En effet, la log-vraisemblance du modèle s'écrit dans ce cas*

$$\sum_{i=1}^{n}\Big(y_i \ln \widehat{q}_i + (1-y_i)\ln(1-\widehat{q}_i)\Big).$$

Le modèle saturé se caractérise par le fait que $\widehat{q}_i = y_i$, ce qui annule la log-vraisemblance associée puisque $y_i \ln y_i = (1-y_i)\ln(1-y_i) = 0$ car $y_i = 0$ ou 1. La déviance est alors donnée par

$$\begin{aligned} D &= -2\sum_{i=1}^{n}\Big(y_i \ln \widehat{q}_i + (1-y_i)\ln(1-\widehat{q}_i)\Big) \\ &= -2\sum_{i=1}^{n}\Big(y_i \ln \frac{\widehat{q}_i}{1-\widehat{q}_i} + \ln(1-\widehat{q}_i)\Big). \end{aligned} \tag{9.42}$$

En différenciant

$$L(\boldsymbol{\beta}|\boldsymbol{y}) = \sum_{i=1}^{n}\Big(y_i \ln q_i + (1-y_i)\ln(1-q_i)\Big)$$

par rapport à β_j, nous obtenons

$$\frac{\partial}{\partial \beta_j} L(\boldsymbol{\beta}|\boldsymbol{y}) = \sum_{i=1}^{n}\left(\frac{y_i}{q_i} - \frac{1-y_i}{1-\widehat{q}_i}\right) q_i(1-q_i)x_{ij}$$

d'où l'on tire

$$\begin{aligned} \sum_{j=1}^{p}\beta_j \frac{\partial}{\partial \beta_j} L(\boldsymbol{\beta}|\boldsymbol{y}) &= \sum_{i=1}^{n}(y_i - q_i)\sum_{j=1}^{p}\beta_j x_{ij} \\ &= \sum_{i=1}^{n}(y_i - q_i)\ln\frac{q_i}{1-q_i}. \end{aligned}$$

Le membre de gauche s'annulant lorsqu'on l'évalue en l'estimateur du maximum de vraisemblance $\widehat{\boldsymbol{\beta}}$ de $\boldsymbol{\beta}$, il vient

$$\sum_{i=1}^{n}(y_i - \widehat{q}_i)\mathit{logit}(\widehat{q}_i) = 0$$

$$\Leftrightarrow \sum_{i=1}^{n} y_i \mathit{logit}(\widehat{q}_i) = \sum_{i=1}^{n}\widehat{q}_i \mathit{logit}(\widehat{q}_i).$$

En tenant compte de cette dernière équation dans (**??**)*, il vient*

$$D = -2\sum_{i=1}^{n}\Big\{\widehat{q}_i logit(\widehat{q}_i) + \ln(1-\widehat{q}_i)\Big\}.$$

On voit que la déviance ne dépend que des valeurs ajustées $\widehat{q}_i$ des q_i, et pas des observations y_i ; dès lors, D ne donne aucune indication sur la qualité de l'ajustement des observations et ne peut pas être utilisée pour mesurer l'adéquation du modèle.

Exemple 9.8.20 (Régression de Poisson)**.** *Notons $L(\widehat{\boldsymbol{\lambda}}|\boldsymbol{n})$ la log-vraisemblance du modèle ajusté, où $\widehat{\boldsymbol{\lambda}} = (\widehat{\lambda}_1, \widehat{\lambda}_2, \ldots, \widehat{\lambda}_n)$ avec $\widehat{\lambda}_i = \exp(\widehat{\boldsymbol{\beta}}^t \boldsymbol{x}_i)$. Puisque*

$$\exp(-\lambda)\frac{\lambda^k}{k!} \leq \exp(-k)\frac{k^k}{k!}$$

quels que soient k et λ, la log-vraisemblance maximale qu'il est possible d'obtenir dans le modèle spécifiant que les N_i sont des variables indépendantes de loi de Poisson s'obtient pour $N_i \sim \mathcal{P}oi(n_i)$. Il y a alors autant de paramètres que d'observations, soit n paramètres. Notons $L(\boldsymbol{n}|\boldsymbol{n})$ la log-vraisemblance de ce modèle (qui prédira n_i pour la $i$ème observation). La déviance est alors donnée par

$$\begin{aligned} D(\boldsymbol{n}, \widehat{\boldsymbol{\lambda}}) &= 2\Big\{L(\boldsymbol{n}|\boldsymbol{n}) - L(\widehat{\boldsymbol{\lambda}}|\boldsymbol{n})\Big\} \\ &= 2\ln\left\{\prod_{i=1}^{n}\exp(-n_i)\frac{n_i^{n_i}}{n_i!}\right\} - 2\ln\left\{\prod_{i=1}^{n}\exp(-\widehat{\lambda}_i)\frac{\widehat{\lambda}_i^{n_i}}{n_i!}\right\} \\ &= \sum_{i=1}^{n}\left\{n_i \ln\frac{n_i}{\widehat{\lambda}_i} - (n_i - \widehat{\lambda}_i)\right\} \end{aligned}$$

où l'on a posé $y\ln y = 0$ lorsque $y = 0$. Puisque l'inclusion d'un intercept β_0 garantit que (**??**) *est valable, la déviance s'écrit dans ce cas*

$$D(\boldsymbol{n}, \widehat{\boldsymbol{\lambda}}) = \sum_{i=1}^{n} n_i \ln\frac{n_i}{\widehat{\lambda}_i}. \tag{9.43}$$

Remarque 9.8.21 (Pseudo-R^2 pour données de comptage)**.** *Une fois le modèle ajusté (i.e. les variables explicatives pertinentes sélectionnées et l'estimation du maximum de vraisemblance $\widehat{\boldsymbol{\beta}}$ de $\boldsymbol{\beta}$ obtenue), il est primordial d'en évaluer la qualité, c'est-à-dire son habileté à décrire le nombre des sinistres touchant les différents assurés du portefeuille. Cette évaluation peut se faire à l'aide de la déviance comme décrit ci-dessus.*

Supposons les observations N_i de loi $\mathcal{P}oi(\lambda_i)$. La généralisation de la statistique usuelle R^2 du modèle linéaire se base sur la décomposition de la déviance en

$$D(\boldsymbol{n}, \overline{\boldsymbol{n}}) = D(\boldsymbol{n}, \widehat{\boldsymbol{\lambda}}) + D(\widehat{\boldsymbol{\lambda}}, \overline{\boldsymbol{n}})$$

où

1. *$D(\boldsymbol{n}, \overline{\boldsymbol{n}})$ est la déviance du modèle contenant seulement l'intercept β_0 (i.e. celui n'utilisant pas les variables explicatives pour lequel $\widehat{\lambda}_i = \overline{n} = \frac{1}{n}\sum_{j=1}^{n} n_j$, $i = 1, 2, \ldots, n$) donnée par*

$$D(\boldsymbol{n}, \overline{\boldsymbol{n}}) = \sum_{i=1}^{n} n_i \ln \frac{n_i}{\overline{n}};$$

2. *$D(\boldsymbol{n}, \widehat{\boldsymbol{\lambda}})$ est la déviance associée au modèle considéré ;*
3. *$D(\widehat{\boldsymbol{\lambda}}, \overline{\boldsymbol{n}})$ est la déviance non expliquée par le modèle donnée par*

$$D(\widehat{\boldsymbol{\lambda}}, \overline{\boldsymbol{n}}) = \sum_{i=1}^{n} \widehat{\lambda}_i \ln \frac{\widehat{\lambda}_i}{\overline{n}} - (\widehat{\lambda}_i - \overline{n}).$$

Cette décomposition mène à la définition

$$\begin{aligned} R_D^2 &= 1 - \frac{D(\widehat{\boldsymbol{\lambda}}, \overline{\boldsymbol{n}})}{D(\boldsymbol{n}, \overline{\boldsymbol{n}})} \\ &= \frac{\sum_{i=1}^{n} n_i \ln \frac{\widehat{\lambda}_i}{\overline{n}} - (n_i - \widehat{\lambda}_i)}{\sum_{i=1}^{n} n_i \ln \frac{n_i}{\overline{n}}} \end{aligned}$$

avec la convention $y \ln y = 0$ lorsque $y = 0$. Le Pseudo-R^2 mesure la diminution de la déviance qui résulte de l'inclusion de variables explicatives au modèle. La statistique R_D^2 possède les bonnes propriétés suivantes : elle est toujours comprise entre 0 et 1, augmente lorsque de nouvelles variables sont incorporées au modèle et admet une interprétation en terme d'information comme la réduction proportionnelle du contraste de Kullback-Liebler résultant de l'inclusion de nouvelles variables. Seule cette dernière propriété requiert la spécification correcte de la loi des N_i.

Statistique de Pearson

La statistique khi-carrée de Pearson

$$X^2 = \sum_{i=1}^{n} \omega_i \frac{(y_i - \mu_i)^2}{\mathbb{V}[Y_i]}$$

permet de mesurer la qualité d'un ajustement. La statistique X^2, tout comme la déviance D, est distribuée exactement selon une loi khi-carrée dans le cas particulier du modèle linéaire gaussien. Ce résultat est asymptotiquement vrai dans les autres cas.

9.8.11 Tests d'hypothèse sur les paramètres

On désire tester l'hypothèse $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0 = (\beta_0, \beta_1, \ldots, \beta_q)^t$ contre $H_1 : \boldsymbol{\beta} = \boldsymbol{\beta}_1 = (\beta_0, \beta_1, \ldots, \beta_p)^t$ où $q < p < n$. Ceci revient donc à tester la nullité simultanée de $\beta_{q+1}, \ldots, \beta_p$. On utilise alors la statistique Δ qui vaut la différence entre les déviances des deux modèles, à savoir

$$\Delta = D_0 - D_1 = 2\left(\ln L_{\hat{\boldsymbol{\beta}}_1}(\boldsymbol{y}) - \ln L_{\hat{\boldsymbol{\beta}}_0}(\boldsymbol{y})\right) \geq 0.$$

On peut montrer que Δ est approximativement de loi χ^2_{p-q}. On rejette H_0 au profit de H_1 lorsque

$$\Delta_{obs} > \chi^2_{p-q;1-\alpha},$$

où $\chi^2_{p-q;1-\alpha}$ est le quantile d'ordre $1-\alpha$ de la loi χ^2_{p-q}.

L'intérêt de ce type de test apparaît lorsque l'actuaire se demande s'il convient de grouper certains niveaux des variables catégorielles. En effet, le test de nullité des coefficients de régression indique seulement si le niveau en question doit être fusionné avec le niveau de référence. Il se pourrait cependant que deux niveaux d'une variable catégorielle soient statistiquement équivalents, mais diffèrent tous deux du niveau de référence. On s'intéressera alors à un test de type $H_0 : \beta_1 = \beta_2$. Spécifiquement, dans l'exemple utilisé à la Section **??** pour illustrer le codage des variables catégorielles à l'aide de binaires, on pourrait tester $H_0 : \beta_3 = 0$ et $H_0 : \beta_4 = 0$, qui nous indiqueraient si les moins de 30 ans ou les plus de 60 ans diffèrent des 30-65 ans, mais aussi $H_0 : \beta_3 = \beta_4$ qui nous indiquera s'il convient de grouper les moins de 30 ans avec les plus de 65 ans.

9.8.12 Estimation du paramètre de dispersion

L'estimation du paramètre de dispersion ϕ est basée sur la déviance D. Comme $\mathbb{E}[D] \approx n - p - 1$, on pourrait estimer ϕ par

$$\widetilde{\phi} = \frac{1}{n-p-1} D.$$

Cet estimateur est toutefois peu utilisé en pratique car il est très instable. Afin d'éviter ces désagréments, on a recours à un

développement de Taylor à l'ordre 2 de la log-vraisemblance qui nous donne

$$\widehat{\phi} = \frac{1}{n-p-1}(\boldsymbol{y} - \widehat{\boldsymbol{\mu}})^t \mathcal{I}_n(\widehat{\boldsymbol{\mu}})(\boldsymbol{y} - \widehat{\boldsymbol{\mu}});$$

cette dernière estimation est souvent appelée estimation du X^2 de Pearson.

9.8.13 Analyse des résidus

Les mesures d'adéquation examinées ci-dessus (déviance et statistique de Pearson) fournissent des indications globales quant à la qualité d'un modèle. Une analyse soigneuse des résidus permet de découvrir d'où provient l'écart éventuel entre le modèle et les observations, et ainsi d'améliorer le modèle initial, si nécessaire.

Les résidus sont basés sur une distance entre l'observation y_i et la valeur prédite μ_i. Ils permettent de vérifier l'adéquation du modèle mais aussi de détecter des observations particulières (souvent appelées "outliers"). Ces dernières peuvent avoir une influence considérable sur les estimations des paramètres et il est important de recommencer la phase d'estimation après avoir enlevé ces observations, afin de s'assurer de la stabilité des résultats obtenus.

On peut distinguer deux situations où un modèle est jugé inadéquat sur base des mesures globales telles que la déviance ou la statistique de Pearson : soit un petit nombre d'observations sont mal décrites par le modèle, soit toutes les observations présentent un écart systématique par rapport au modèle.

Une représentation graphique des résidus permet de détecter les écarts par rapport au modèle pour la plupart des types de variables dépendantes. Cependant, lorsque la variable dépendante ne peut prendre que quelques valeurs (comme en régression logistique, par exemple), l'analyse des résidus ne peut cependant présenter qu'un intérêt limité.

Deux types de résidus sont couramment utilisés dans le cadre des modèles linéaires généralisés : les résidus de Pearson et les résidus de déviance.

Les résidus de Pearson

Ils sont définis par

$$r_i^P = \frac{\sqrt{w_i}(y_i - \mu_i)}{\sqrt{V(\mu_i)}}.$$

Le nom de ce premier type de résidus provient du fait que r_i^P peut se voir comme la racine carré de la contribution de le $i$ème observation à la statistique de Pearson, i.e.

$$\sum_{i=1}^{n}\{r_i^P\}^2 = X^2.$$

Exemple 9.8.22 (Régression binomiale)**.** *Si les observations y_i sont de loi $\mathcal{B}in(n_i, q_i)$, le résidu de Pearson vaut*

$$r_i^P = \frac{y_i - \widehat{y}_i}{\sqrt{n_i\widehat{q}_i(1-\widehat{q}_i)}}.$$

Exemple 9.8.23 (Régression de Poisson)**.** *Si les observations n_i sont de loi $\mathcal{P}oi(\lambda)$, le résidu de Pearson vaut*

$$r_i^P = \frac{n_i - \widehat{\lambda}_i}{\sqrt{\widehat{\lambda}_i}}.$$

Les résidus de déviance

Nous avons vu que la déviance était une mesure de la qualité de l'ajustement fourni par un modèle. On peut considérer que chaque observation y_i contribue à hauteur d'une quantité d_i à la déviance D, i.e.

$$D = \sum_{i=1}^{n} d_i.$$

Les résidus de déviance sont alors définis comme la racine carrée de la contribution d_i de la $i$ème observation à la déviance D, affectée du signe du résidu brut $y_i - \mu_i$, i.e.

$$r_i^D = \text{signe}(y_i - \mu_i)\sqrt{d_i}.$$

Ainsi,

$$\sum_{i=1}^{n}\{r_i^D\}^2 = D.$$

Exemple 9.8.24 (Régression logistique)**.** *Si les observations y_i sont de loi $\mathcal{B}in(n_i, q_i)$, le résidu de déviance vaut*

$$r_i^D \quad = \quad signe(y_i - \widehat{y}_i)\sqrt{2y_i \ln\left(\frac{y_i}{\widehat{y}_i}\right) + 2(n_i - y_i)\ln\left(\frac{n_i - y_i}{n_i - \widehat{y}_i}\right)}.$$

Exemple 9.8.25 (Régression de Poisson)**.** *Si les observations n_i sont de loi $\mathcal{P}oi(\lambda_i)$, le résidu de déviance vaut*

$$r_i^D = signe(n_i - \widehat{\lambda}_i)\sqrt{2\left\{n_i \ln \frac{n_i}{\widehat{\lambda}_i} - (n_i - \widehat{\lambda}_i)\right\}}$$

où $y \ln y = 0$ lorsque $y = 0$.

Remarque 9.8.26. *Une autre quantité fort intéressante est obtenue en faisant la différence entre la déviance obtenue à l'aide des n observations et celle obtenue en supprimant la $i$ème observation du jeu de données (donc sur base d'un jeu de données de $n-1$ observations). Ceci permet de mesurer l'influence globale de y_i sur le modèle.*

Notez que le calcul exact de ces quantités prend du temps (puisque le modèle doit être ajusté n fois à l'aide d'une jeu de données de taille $n-1$). Afin d'éviter un temps de calcul trop élevé, on a recours en pratique à différentes approximations pour la différence des déviances obtenues en omettant l'observation y_i.

Représentation des résidus

Les résidus introduits ci-dessus peuvent être portés en graphique en fonction de plusieurs statistiques, chacune fournissant une information différente sur les écarts par rapport au modèle. On peut évidemment représenter les résidus en fonction du numéro d'observation (index plot), ce qui permet d'identifier les observations conduisant à de grands résidus (et donc contribuant à l'inadéquation du modèle).

Les résidus peuvent également être représentés en fonction des valeurs prédites $\widehat{\mu}_i$ ou des prédicteurs linéaires $\widehat{\eta}_i$. On peut aussi les représenter en fonction de chacune des variables explicatives.

Observations influentes

On dit qu'une observation y_i a de l'influence sur un modèle considéré lorsqu'une petite modification de y_i ou l'omission de celle-ci conduit à des estimations fort différentes des paramètres du modèle. Une telle observation a donc un impact considérable sur les conclusions de l'étude. Cependant, une observation qui a de l'influence n'est pas nécessairement un outlier : elle peut se trouver près des autres observations et avoir un résidu petit. L'effet de levier ("*leverage effect*") d'une observation y_i sur la valeur prédite $\widehat{y}_j$ est la

dérivée de $\widehat{y}_j$ par rapport à y_i ; il indique donc comment varient les valeurs prédites des autres observations en fonction des modifications de y_i. Afin de mesurer cet effet, on se sert de la matrice de projection $\boldsymbol{H}$ qui envoie les observations y_i sur leurs prédictions $\widehat{y}_i$, i.e.

$$\widehat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y}.$$

Dans le cas des modèles linéaires généralisés, on a

$$\boldsymbol{H} = \boldsymbol{V}^{1/2}\boldsymbol{X}\left(\boldsymbol{X}^t\boldsymbol{V}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^t\boldsymbol{V}^{1/2}$$

où $\boldsymbol{V}$ est une matrice diagonale définie par

$$\boldsymbol{V} = \operatorname{diag}\left(\{V(\mu_i)\}^2\left(\frac{d\eta_i}{d\mu_i}\right)^2\frac{\phi}{\omega_i}\right).$$

Les éléments diagonaux de $\boldsymbol{H}$ donnent une information sur l'influence de chaque observation. Il est bon de noter que cette mesure dépend à la fois des variables explicatives et des estimations des paramètres $\beta_0, \beta_1, \ldots, \beta_p$. Comme la trace de $\boldsymbol{H}$ vaut $p+1$, la valeur moyenne des termes diagonaux est $(p+1)/n$. Les valeurs correspondant aux termes diagonaux qui excèdent disons deux fois la valeur moyenne de $(p+1)/n$ doivent faire l'objet d'un examen approfondi.

Distance de Cook

La distance de Cook est utilisée pour déterminer les observations qui influencent l'ensemble des paramètres $\beta_0, \beta_1, \ldots, \beta_p$. Tout comme dans le modèle linéaire, l'idée est ici de déterminer l'estimateur du maximum de vraisemblance $\widehat{\boldsymbol{\beta}}$ du paramètre $\boldsymbol{\beta}$ et ensuite le même estimateur $\widehat{\boldsymbol{\beta}}_{(i)}$ obtenu en supprimant la $i$ème observation du jeu de données. La distance de Cook vaut alors

$$C_i = \frac{1}{p+1}(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)})^t\boldsymbol{X}^t\boldsymbol{V}\boldsymbol{X}(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)})$$

qui s'interprète comme une distance entre $\widehat{\boldsymbol{\beta}}$ et $\widehat{\boldsymbol{\beta}}_{(i)}$. Si C_i est grande, cela indique que l'observation y_i a une influence importante sur l'estimation des paramètres. Evidemment, le calcul de C_i est très coûteux puisque tous les paramètres doivent être réestimés une fois que y_i a été supprimée. En statistique des assurances, ce calcul est souvent impossible en raison du grand nombre d'observations. Afin d'éviter de réajuster le modèle n fois, on a souvent recours en pratique à différentes approximations.

9.8.14 La pratique des modèles linéaires généralisés

De l'importance du choix de la sous-famille exponentielle

Les quelques exemples mentionnés en introduction suffisent souvent en pratique : modélisation des coûts des sinistres par un modèle de régression Gamma, et modélisation des nombres par un modèle de régression de Poisson. Pourtant, le choix de la sous-famille n'est pas neutre sur une tarification.

Considérons ainsi l'exemple (simpliste) suivant, basé sur trois observations,

observation i	1	2	3
variable à expliquer Y_i (coût des sinistres)	1	2	8
variable explicative X_i (puissance du véhicule)	1	2	3

On cherche à ajuster un modèle linéaire généralisé, i.e. $g(\mathbb{E}[Y]) = \alpha + \beta X$ où g est une fonction lien. La Figure **??** présente l'influence du choix de la loi de probabilité (pour la fonction de lien canonique), en considérant successivement Y_i de loi normale, de loi de Poisson et de loi Gamma.

Si les trois lois donnent des résultats proches au bord (pour des valeurs de X proches de 1 ou de 3), ailleurs, le comportement est sensiblement différent. En particulier, par rapport aux deux autres lois, la loi Gamma propose un coût des sinistres plus important (à puissance égale) pour les petites et les grandes puissances, et en contrepartie propose un coût moins important pour les puissances moyennes. Ce coût reflétant la prime, et si le vrai modèle est celui de Poisson, on peut interpréter ce graphique de la façon suivante :

- avec une modélisation normale, les véhicules de puissance moyenne payent pour le risque des autres, en ayant une prime plus importante que leur vrai risque,
- avec une modélisation Gamma, les véhicules de puissance importante, ou très faible, payent pour le risque des véhicules de puissance moyenne, sous-tarifés.

Aussi, et bien que cette analyse doit être mitigée par la prise en compte de l'impact de la fonction lien, on notera que le choix de la loi de la variable à expliquer n'est en aucun cas neutre quant aux primes pures qui en découlent.

FIGURE 9.11 – Ajustement d'un modèle linéaire généralisé à partir de trois points, pour des lois normales, Poisson et Gamma.

Modélisation du coût total, importance des contrats sans sinistre

Supposons que la variable d'intêrêt Y soit le coût (total) des polices, sur un an, pour l'ensemble des polices du portefeuille. Un très grand nombre de polices n'ayant pas de sinistre, la variable Y sera alors nulle pour la plupart des observations. Une loi Gamma (par exemple) ne permet pas de modéliser ce genre de comportement (voir le Chapitre 6 du Tome 1 sur la différence entre le modèle collectif et le modèle individuel).

La loi de *Tweedie* permet de prendre en compte ce genre de comportement, en rajoutant une mesure de Dirac en 0 à une loi de probabilité de support $\mathbb{R}^+$. La loi de Y est alors une loi Poisson composée,

$$Y \sim \mathcal{CP}oi\left(\mu^{2-\gamma}\phi(2-\gamma), \mathcal{G}am\left(-\frac{2-\gamma}{\phi(1-\gamma)}, \phi(2-\gamma)\mu^{\gamma-1}\right)\right),$$

où $1 < \gamma < 2$. On obtient alors une fonction variance de la forme $V(\mu) = \phi\mu^\gamma$. On retrouve le modèle de Poisson quand $\gamma \to 1$ et une loi Gamma quand $\gamma \to 2$. Il est en fait possible d'obtenir une classe beaucoup plus large, y compris dans le cas où $\gamma > 2$ en considérant des lois stables.

Remarque 9.8.27. *Ces modèles ont été utilisé en assurance automobile dans* Jorgensen & De Souza *(1994), où une valeur de γ proche de 1,5 a été obtenue.*

L'importance de la fonction de lien

Nous avons noté ci-dessus que le choix de la sous-famille exponentielle considérée n'était pas neutre quant à la tarification. Le même résultat reste vrai pour le choix de la fonction de lien. Toujours sur les trois observations, la Figure **??** représente l'influence de la fonction de lien (pour une loi de Poisson et une loi Gamma pour Y). On peut noter qu'à famille de lois fixée, le choix de la fonction de lien n'est, là encore, pas neutre.

Figure 9.12 – Ajustement d'un modèle linéaire généralisé à partir de trois points, pour des lois de Poisson et Gamma, avec différentes fonctions liens.

Toutefois, on notera qu'il est souvent d'usage d'utiliser la fonction de lien logarithmique puisqu'elle présente l'avantage de don-

ner un modèle multiplicatif, et les coefficients β_j ont alors une interprétation simple, en terme de multiplicateurs.

Si le choix de la fonction de lien n'est pas innocent en matière de tarification, il est toutefois possible de prendre cette fonction comme inconnue, et de chercher à l'estimer à partir des données. Pour cela, la transformation de Box-Cox permet d'avoir la forme paramétrique simple

$$g(x) = \begin{cases} \left(x^{\lambda} - 1\right)/\lambda, & \text{si } \lambda \neq 0, \\ \log(x), & \text{si } \lambda = 0. \end{cases}$$

On notera que $\lambda = 1$ correspond à une fonction lien identité (modèle additif), et $\lambda \to 0$ à une fonction lien logarithmique (modèle multiplicatif). Si $\lambda = -1$, on retrouve également la fonction lien inverse. Aussi, un grand nombre de fonctions de lien usuelles appartiennent à cette famille. Il est alors possible de chercher λ qui maximise la vraisemblance du modèle.

9.9 Les modèles additifs généralisés

9.9.1 Principe

Tout comme les modèles additifs ont permis de prendre en compte des effets non linéaires des variables explicatives, sans avoir à en spécifier a priori la forme, lorsque les variables à expliquer étaient gaussiennes, les modèles additifs généralisés (GAM pour l'anglais *Generalized Additive Models*) permettent la même extension dans les modèles de régression de Poisson, Binomial et Gamma.

Comme nous l'avons vu plus haut, l'obtention des estimations du maximum de vraisemblance des paramètres $\beta_0, \beta_1, \ldots, \beta_p$ intervenant dans le score linéaire η du GLM s'obtient à l'aide d'ajustements répétés par la méthode des moindres carrés pondérés de pseudo-observations z_i. Ceci permettra de traiter les GAM à l'aide des modèles additifs étudiés plus haut.

9.9.2 Inférence dans les modèles généralisés additifs : deux approches possibles

Modèle additif sur pseudo-variables

L'ajustement des GAM repose sur le fait que les estimations des paramètres $\boldsymbol{\beta}$ d'un GLM peuvent s'obtenir en ajustant un modèle linéaire à des pseudo-observations (à l'aide d'une technique des

moindres carrés pondérés). C'est exactement la même approche qui sera suivie ici. Précisément, nous définissons les pseudo-observations

$$z_i^{(k)} = \widehat{\eta}_i^{(k)} + \frac{y_i - \widehat{\mu}_i^{(k)}}{\widehat{D}_i^{(k)}}$$

où $\widehat{\mu}_i^{(k)} = g^{-1}(\widehat{\eta}_i^{(k)})$ avec $\widehat{\eta}_i^{(k)} = \widehat{c} + \sum_{j=1}^{p} \widehat{f}_j^{(k)}(x_{ij})$ pour $i = 1, 2, \ldots, n$, et

$$\widehat{D}_i^{(k)} = \frac{\partial}{\partial \eta} g^{-1}(\widehat{\eta}_i^{(k)}).$$

Nous associons à chacune des pseudo-observations $z_i^{(k)}$ les poids

$$\pi_i^{(k)} = \left(\frac{\widehat{D}_i^{(k)}}{\widehat{\sigma}_i^{(k)}} \right)^2 \text{ avec } \widehat{\sigma}_i^{(k)} = \sqrt{\frac{\phi V\left(g^{-1}(\widehat{\eta}_i^{(k)})\right)}{\omega_i}}.$$

La nouvelle estimation $\widehat{f}_j^{(k+1)}$ de f_j est alors obtenue en régressant $z_i^{(k)}$ sur $\boldsymbol{x}_i$ en tenant compte des poids $\pi_i^{(k)}$. La technique utilisée pour estimer les fonctions $f_1(\cdot), \cdots, f_p(\cdot)$ est celle décrite plus haut pour les modèles additifs.

Formalisée en terme d'algorithme, l'estimation dans les GAM s'effectue comme suit :

initialisation : $\widehat{c} \leftarrow g(\overline{y})$ et $\widehat{\boldsymbol{f}}_j^{(0)} \leftarrow 0$, $j = 1, \ldots, p$.

cycle : pour $k = 1, 2, \ldots$, on construit les pseudo-observations $z_i^{(k)}$ et on leur associe les poids $\pi_i^{(k)}$. Ensuite, on ajuste les $z_i^{(k)}$ à $c + \sum_{j=1}^{p} f_j(x_{ij})$ à l'aide d'un modèle additif comme décrit plus haut, i.e.

(i) on initialise $\widehat{c} \leftarrow \overline{z}^{(k)}$ et $\widehat{\boldsymbol{f}}_j^{(0)} \leftarrow \widehat{\boldsymbol{f}}_j^{(k)}$

(ii) on réévalue

$$\widehat{\boldsymbol{f}}_j^{(1)} \rightarrow \boldsymbol{H}_{\lambda_j} \left(\boldsymbol{z}^{(k)} - (\overline{z}^{(k)}, \ldots, \overline{z}^{(k)})^t - \sum_{s<j} \widehat{\boldsymbol{f}}_s^{(1)} - \sum_{s>j} \widehat{\boldsymbol{f}}_s^{(0)} \right)$$

(iii) réitérer (ii) et stopper lorsque la somme des carrés des résidus cesse de décroître.

critère d'arrêt : les variations dans les f_j deviennent négligeables.

Comme on peut le constater, chaque étape de l'algorithme itératif menant aux estimations des fonctions $f_j(\cdot)$ requiert un back-fitting complet afin de revoir l'estimation de ces fonctions sur base des pseudo-variables obtenues à cette étape.

Maximum de vraisemblance local

Une autre approche consiste à étendre directement la méthode Loess aux modèles généralisés additifs. On aura alors recours à des ajustements locaux dans des GLM, par maximum de vraisemblance (exactement comme procédait la méthode Loess à partir d'une log-vraisemblance gaussienne).

Plus précisément, étant donné un point x où on veut estimer la réponse, on détermine un voisinage $\mathcal{V}(x)$ et des poids $w_i(x)$ exactement comme dans Loess. Il s'agira de résoudre les équations de vraisemblance

$$\boldsymbol{X}^t\boldsymbol{\Omega}(x)\big(\boldsymbol{y}-\boldsymbol{\mu}(\widehat{\beta}(x))\big)=\mathbf{0}$$

où la matrice diagonale $\boldsymbol{\Omega}(x)$ reprend les poids $w_1(x),\ldots,w_n(x)$.

Exemple 9.9.1 (Régression logistique)**.** *Considérons les observations (y_i,x_i) où y_i est de loi $\mathcal{B}in(n_i,q_i)$ où q_i est une fonction de x_i. Si on veut estimer les $q_i=q(x_i)$, on déterminera le voisinage $\mathcal{V}(x)$ et on estimera en x la fonction f intervenant dans*

$$q(x)=\frac{\exp\big(f(x)\big)}{1+\exp\big(f(x)\big)}$$

en maximisant

$$L\big(\beta_0(x),\beta_1(x)\big)=\sum_{\xi\in\mathcal{V}(x)}w_\xi(x)\ln\ell_\xi\big(\beta_0(x),\beta_1(x)\big)$$

où

$$\ell_\xi\big(\beta_0(x),\beta_1(x)\big)=\left(\begin{array}{c}n_\xi\\y_\xi\end{array}\right)\big(q_\xi(x)\big)^{y_\xi}\big(1-q_\xi(x)\big)^{n_\xi-y_\xi}$$

avec

$$q_\xi(x)=\frac{\exp\big(\beta_0(x)+\beta_1(x)\xi\big)}{1+\exp\big(\beta_0(x)+\beta_1(x)\xi\big)}.$$

Finalement, la valeur ajustée de $q(x)$ s'obtiendra grâce à

$$\widehat{q(x)}=\frac{\exp\big(\widehat{f}(x)\big)}{1+\exp\big(\widehat{f}(x)\big)}$$

avec

$$\widehat{f}(x)=\widehat{\beta}_0(x)+\widehat{\beta}_1(x)x.$$

9.9.3 En pratique...

Le plus souvent, l'actuaire dispose d'un grand nombre de variables catégorielles et seulement de quelques variables continues. Supposons que les variables explicatives $\boldsymbol{x}_i$ relatives à l'assuré i soient réorganisées afin que

$$\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{if}, x_{i,f+1}, \ldots, x_{i,f+c})$$

où $x_{i1}, \ldots, x_{if}$ sont les f variables binaires utilisées pour coder les variables catégorielles décrivant l'assuré i, et $x_{i,f+1}, \ldots, x_{i,f+c}$ sont les c variables continues relatives à cet assuré. Le prédicteur linéaire η_i relatif à cet individu sera de la forme

$$\eta_i = c + \sum_{j=1}^{f} \beta_j x_{ij} + \sum_{j=1}^{c} f_j(x_{i,f+j}).$$

On peut même aller plus loin et introduire des effets d'interaction entre une variable catégorielle et une variable continue (typiquement, entre le sexe et l'âge de l'assuré dans le contexte de l'assurance automobile).

9.10 Cas pratique de tarification automobile

9.10.1 Description du portefeuille

Nous considérons dans cette section un portefeuille d'assurance automobile, que nous analysons à l'aide du logiciel SAS largement utilisé au sein des compagnies d'assurance. Les méthodes proposées peuvent bien sûr être adaptées à d'autres types de couverture (assurance dégâts matériels aux véhicules, assurance incendie des bâtiments, assurance vol, assurance annulation de voyage, etc.), en tenant compte de certaines particularités de ces contrats. Les données proviennent du portefeuille d'une grande compagnie opérant en Belgique. Nous disposons d'observations relatives à n =158 061 polices observées durant l'année 1997. Les variables reprises dans le fichier sont données au Tableau **??**.

D'emblée, précisons quelques points importants. Tout d'abord, à l'exception des variables `NSIN`, `IND` et `CTOT` décrivant la sinistralité, il s'agit toutes de variables connues *a priori* de l'assureur (c'est-à-dire qu'il peut se servir de ces variables pour personnaliser le montant de la prime réclamée à l'assuré pour la couverture du risque). Clairement, les caractéristiques du risque inconnues de l'assureur en

début de période ne peuvent pas être utilisées dans la grille de tarification *a priori*; le cas échéant, l'assureur s'en servira à d'autres fins (comme nous le verrons plus loin). Parmi les variables explicatives disponibles reprises au Tableau **??**, nous distinguons différents types :

1. celles relatives au preneur d'assurance (`AGES`, `AGGLOM` et `SEXE` dans notre exemple) ;
2. celles relatives au véhicule assuré (`CARB`, `KW`, `SPORT` et `USAGE` dans notre exemple) ;
3. celles relatives à la couverture pour laquelle le preneur a opté (`FRAC` et `GARACCESS`, dans notre exemple).

Ces variables sont décrites en détail dans la suite. Notez que les informations relatives à la sinistralité passée des assurés (telles que nombre et/ou coût des sinistres, ou encore un résumé fourni par la position occupée dans une échelle bonus-malus ou un coefficient de réduction majoration appliqué à la prime) ne peuvent normalement pas être incorporée dans le schéma de tarification *a priori*. En effet, l'assureur reverra le montant de la prime en fonction des sinistres causés par l'assuré à l'aide de la théorie de la crédibilité ou des systèmes bonus-malus (décrits dans les chapitres qui suivent). Dès lors, l'inclusion du passé sinistres dans la tarification *a priori* générerait une double pénalisation des assurés ayant déclaré des sinistres, et une sous-tarification de ceux n'en ayant pas rapporté à la compagnie.

Avant d'entamer la modélisation du nombre et du coût des sinistres, il est essentiel de bien connaître le portefeuille sur lequel on travaille. Il faut dès lors prendre le temps de décrire en détail les différentes variables tarifaires et d'examiner la composition du portefeuille à analyser.

Nous travaillons ici sur un fichier polices. Un tel fichier compte autant de lignes que de polices en portefeuille durant la période considérée. Il résume l'information disponible en début de période à propos de chacun des contrats et décrit la sinistralité relative à ceux-ci. En marge de ce fichier, l'assureur dispose également d'un fichier sinistres où sont renseignées toutes les caractéristiques des sinistres produits par les polices en portefeuille (circonstances dans lesquelles ces sinistres ont eu lieu, personnes impliquées, présence de dommages corporels, ...). Ces deux fichiers sont mis en relation grâce au numéro de police.

Remarque 9.10.1. *La phase de constitution de la base de données est cruciale pour le tarif qui découlera de l'analyse : comment arriver*

Variable	Description
AGES	Age du souscripteur
AGGLOM	Type d'agglomération où réside le souscripteur
CARB	Carburant
CTOT	Coût total des sinistres (en euros)
DUR	Durée de couverture (en jours)
FRAC	Fractionnement de la prime
GARACCESS	Etendue de la couverture
IND	Code sinistre
KW	Puissance du véhicule
NSIN	Nombre de sinistres
SEXE	Sexe du souscripteur
SPORT	Caractère sportif du véhicule
USAGE	Usage du véhicule

TABLE 9.3 – Variables comprises dans le fichier.

à un tarif correct en utilisant des données erronnées, incomplètes ou obsolètes ? Il est donc essentiel que l'actuaire se penche sur la constitution de la base de données, et ne délègue pas la tâche sans contrôle au service informatique ou à un jeune stagiaire fraîchement sorti de l'école. La phase d'extraction des données et leur nettoyage représente une part considérable du temps consacré à l'étude.

Il est important :

1. *que les données soient homogènes, c'est-à-dire concernent des polices d'un même portefeuille, dont les conditions sont semblables (si tel n'est pas le cas, il faut introduire des variables explicatives différenciant les catégories d'assurés ou identifiant la compagnie qui a émis les polices) ;*
2. *lors de la fusion de données provenant de plusieurs sources, d'identifier systématiquement la provenance de celles-ci (par exemple, lors de la fusion de bases de données relatives à des affaires vendues par courtier ou en direct, on ajoutera une variable donnant le canal de distribution par lequel la police a été produite) ;*
3. *d'examiner très soigneusement les données manquantes, et de ne les ignorer sous aucun prétexte. Souvent, l'information omise révèle certains traits de la police. On rajoutera donc systématiquement un niveau aux variables catégorielles indi-*

quant lorsque l'information est manquante. Ce n'est que lorsqu'on se sera assuré que les ommissions sont aléatoires qu'on pourra négliger les polices dont certaines informations sont manquantes.

La plupart des compagnies sont à présent conscientes de la nécessité de disposer de données aussi nombreuses et de bonne qualité que possible. La constitution et la maintenance des bases de données figurent d'ailleurs parmi les préoccupations les plus importantes des grands groupes financiers. A cet égard, la généralisation de la transmission des informations par voie électronique (le courtier ou l'agent saisit les données pour recevoir une offre de prix en ligne) permet d'éviter les erreurs d'encodage ou les valeurs manquantes (puisque l'offre n'est fournie que si tous les champs ont été dûment complétés).

9.10.2 Les variables décrivant la sinistralité

Le nombre des sinistres `NSIN`

Il s'agit du nombre de sinistres déclarés par l'assuré à la compagnie, et donc pas du nombre de sinistres causés par l'assuré sur l'année. L'assuré peut en effet estimer (à tort ou à raison) avoir intérêt à dédommager lui-même le tiers lésé en cas de préjudice mineur (comme une griffe dans la portière d'un véhicule ancien, par exemple, pour lequel le jeune propriétaire préférera peut-être un billet de 100 € , synonyme d'agapes et de réjouissances estudiantines, à une remise en état du véhicule). Il est évident que le nombre de sinistres dédommagés directement par l'assuré est fonction de la politique de tarification *a posteriori* de l'assureur (comprenez la manière dont les sinistres déclarés et ayant donné lieu à une indemnisation par la compagnie sont pénalisés). Il convient donc d'être particulièrement vigilant à cet égard lorsqu'on envisage de modifier les mécanismes de personnalisation *a posteriori* du montant des primes. Nous reviendrons sur ces questions dans les chapitres suivants.

En RC automobile, `NSIN` mérite une attention toute particulière car les coûts des sinistres ne se prêtent souvent pas à une segmentation poussée. De plus, `NSIN` jouera un rôle central dans la personnalisation *a posteriori* des montants des primes (la plupart des systèmes commerciaux, tels les mécanismes bonus-malus, n'intégrant que le nombre des sinistres dans la formule de réévaluation de la prime en cours de contrat).

Le gros avantage de NSIN est d'être généralement connu avec précision par la compagnie. Si on excepte les sinistres survenus en fin de période, qui ne seront sans doute déclarés à l'assureur qu'au début de la période suivante, les sinistres sont généralement renseignés rapidement à la compagnie, soit en application d'une clause contractuelle imposant un délai de déclaration sous peine de déchéance de la garantie, soit que l'assuré désire être dédommagé aussi vite que possible.

La fréquence moyenne des sinistres pour le portefeuille est de 12.45% par an. Le Tableau **??** nous apprend que le nombre maximum de sinistres déclarés par un assuré vaut 5. Plus précisément, 140 276 (soit 88.75%) assurés n'ont déclaré aucun sinistre, 16 085 (soit 10.18 %) en ont déclaré 1, 1 522 (soit 0.96%) en ont déclaré 2, 159 (soit 0.10%) assurés ont déclaré 3 sinistres, 17 assurés (soit 0.01%) ont déclaré 4 sinistres et 2 assurés (soit moins de 0.01%) en ont déclaré 5, au cours de l'année 1997.

Le Tableau **??** décrit l'ajustement de la distribution observée de NSIN par une loi de Poisson de paramètre λ identique pour toutes les polices. L'estimateur du maximum de vraisemblance du paramètre est $\widehat{\lambda} = 0.1245$. On constate que l'ajustement est très mauvais, et rejeté sans aucune hésitation par un test khi-carré (valeur observée de la statistique khi-carrée valant 783.75, pour une p-valeur inférieure à 10^{-4}). Si on observe la succession des signes de la séquence

$$\Pr[\widehat{\texttt{Nsin}} = k] - \Pr[\mathcal{P}oi(\widehat{\lambda}) = k], \quad k \in \mathbb{N},$$

donnée dans la dernière colonne du Tableau **??**, on voit clairement apparaître la succession de signes +,-,+,+,... Cette séquence n'est pas fortuite. Elle est due à une propriété très importante des mélanges de Poisson connue comme le "Shaked's Two Crossings Theorem" (voir la Propriété 3.7.6 du Tome I).

La succession de signes conforte donc l'hypothèse d'une loi de Poisson mélange pour NSIN au niveau du portefeuille. Ceci indique que le portefeuille est hétérogène et justifie la différenciation *a priori* des assurés.

L'occurrence des sinistres IND

Il s'agit d'une variable binaire indiquant si l'assuré a déclaré au moins un sinistre sur l'année, i.e.

$$\texttt{IND} = \mathbb{I}[\texttt{NSIN} \geq 1] = \begin{cases} 1, & \text{si } \texttt{NSIN} \geq 1, \\ 0, & \text{sinon.} \end{cases}$$

Nombre k de sinistres	Nombre de polices obs.	Nombre de polices prédit	Signe
0	140 276	139 553.33	+
1	16 085	17 379.16	-
2	1 522	1 082.15	+
3	159	44.92	+
4	17	1.40	+
5	2	0.04	+
≥ 6	0	0.00	

TABLE 9.4 – Sinistralité observée dans le portefeuille et ajustement par une loi de Poisson.

Sur les 158 061 assurés du portefeuille, 140 276 (soit 88.75%) n'ont déclaré aucun sinistre et 17 785 (soit 11.25%) ont fait jouer la garantie de la compagnie durant l'année.

Le coût des sinistres `CTOT`

Il s'agit de la charge des sinistres pour l'année 1997, c'est-à-dire du coût total (en euro) mis par l'assuré à charge de la compagnie. Il s'agit de la somme des paiements, de la réserve et des frais de gestion de sinistres survenus en 1997. Le risque monétaire a donc deux composantes : la composante occurrence `IND` qui nous renseigne si l'assuré a fait jouer la garantie de la compagnie au moins une fois sur la période considérée, et la composante coût total des sinistres déclarés (nulle si `IND`=0). On peut donc représenter `CTOT` comme

$$\texttt{CTOT} = \texttt{IND} \times \texttt{CTOT}_+$$

où $\texttt{CTOT}_+$ a même loi que `CTOT` sachant $\texttt{CTOT} > 0$ ou encore que `CTOT` sachant IND $= 1$. La variable $\texttt{CTOT}_+$ est donc strictement positive, alors que `CTOT` est le plus souvent nulle (nulle pour 88.75% des assurés dans notre exemple).

Si on distingue le coût de chacun des sinistres, on obtient

$$\texttt{CTOT} = \sum_{k=1}^{\texttt{NSIN}} C_k$$

où C_k est le coût (supposé strictement positif) du $k$ème sinistre (avec la convention $\texttt{CTOT} = 0$ si $\texttt{NSIN} = 0$). Le coût moyen $\overline{C}$ d'un sinistre

ayant touché une police est alors donné par

$$\overline{C} = \frac{\texttt{CTOT}_+}{\texttt{NSIN}}, \text{ si } \texttt{NSIN} > 0,$$

et $\overline{C} = 0$, si $\texttt{NSIN} = 0$. Lors de l'étude du coût des sinistres, il importe de tenir compte du nombre de sinistres engendrés par l'assuré. Ainsi, on analysera le plus souvent $\overline{C}$ en introduisant un poids NSIN.

Remarque 9.10.2. *Une autre variable souvent utilisée est le ratio S/P par catégories de polices (à savoir la charge totale des sinistres causés par cette catégorie d'assurés divisée par le montant des primes encaissées). Cette variable dépend donc du tarif en vigueur lors de la récolte des observations. Elle introduit donc l'ancien tarif dans l'élaboration du nouveau et cet usage est à déconseiller.*

Nous savons que seuls 17 785 assurés ont fait jouer la garantie de l'assureur durant l'année 1997. Intéressons-nous à la sinistralité de ces assurés (c'est-à-dire à la variable $\texttt{CTOT}_+$). Le coefficient d'asymétrie vaut 84.31, ce qui traduit une asymétrie gauche très importante. Plus précisément, 25% des polices qui ont donné lieu au paiement d'indemnités par la compagnie ont causé des sinistres dont le montant était inférieur à 145.02 € , 50% inférieur à 566.51 € , et 75% inférieur à 1 450.67 € . La moyenne de $\texttt{CTOT}_+$ s'élève à 1 807.46 € . La prime pure empirique vaut donc

$$\widehat{\mathbb{E}[\texttt{CTOT}_+]}\widehat{\mathbb{E}[\texttt{IND}]} = \widehat{\mathbb{E}[\texttt{CTOT}]} = 203.38 \text{ € }.$$

Le coefficient de variation de $\texttt{CTOT}_+$ vaut 982.61, pour un écart-type de 17 760.39.

Les quatre polices ayant causé les sinistres dont les montants sont les plus élevés sont 452 839 € , 499 601 € , 499 917 € , 1 989 568 € . Il est souvent nécessaire d'écréter les sinistres afin d'analyser leurs coûts. Une approche classique consiste à les plafonner à un quantile, par exemple $q_{0.99}$=19 927.4 € .La théorie des valeurs extrêmes qui sera présentée au Chapitre **??** fournit les outils adéquats pour traiter les sinistres graves.

9.10.3 La mesure de l'exposition au risque : la variable DUR

Il s'agit du nombre de jours où la police a été en vigueur durant l'année 1997. On s'en servira pour mesurer l'exposition au risque. Elle permettra de tenir compte du fait qu'un sinistre déclaré par une police en vigueur durant 1 mois est un plus mauvais signe pour l'assureur qu'un sinistre relatif à une police en vigueur toute l'année.

FIGURE 9.13 – *Durée de couverture.*

Remarque 9.10.3. *Notons au passage que l'exposition au risque devrait être mesurée par le nombre de kilomètres parcourus plutôt que par le nombre de jours où la police était en vigueur (le véhicule pourrait fort bien rester au garage pendant certaines périodes, n'étant donc pas soumis au risque). Le kilométrage annuel est cependant extrêmement difficile à mesurer, si bien que les assureurs européens ont pour la plupart renoncé à introduire le kilométrage annuel dans les critères de tarification, et ont préféré recourir à des "proxies" tels que l'usage du véhicule (les véhicules utilisés à des fins professionnelles parcourant sans doute davantage de kilomètres par an) ou le carburant (le choix du diesel étant souvent justifié par des distances couvertes plus importantes).*

Dans notre portefeuille, la durée moyenne de couverture est de 323.93 jours. Quelques polices n'ont été en vigueur que pendant une seule journée. La Figure **??** donne un idée des périodes de couverture des différentes polices en portefeuille. On constate une majorité de polices couvertes durant toute l'année, et une répartition plus ou moins uniforme des durées de couverture inférieures à un an. Typiquement, les polices dont l'exposition au risque est inférieure à l'année sont les nouvelles affaires et les résiliations. Le portefeuille que nous étudions est donc relativement stable.

9.10.4 Caractéristiques du preneur d'assurance

Variable `AGES`

Il s'agit d'une variable quantitative à valeurs entières donnant l'âge du preneur d'assurance (en années révolues) au premier janvier de l'année 1997. Plus que l'âge en tant que tel, c'est l'expérience de conduite qu'on espère récupérer par le biais de cette variable. Une variable équivalente (tant la corrélation entre les deux est grande) et souvent mieux acceptée par la clientèle est l'ancienneté du permis de conduire.

Les polices reprennent souvent la notion de conducteur habituel du véhicule, et les caractéristiques personnelles déterminant le montant de la prime sont alors celles du conducteur habituel mentionné aux conditions particulières, et pas celle du preneur d'assurance. Il est bon de noter que les caractéristiques personnelles (telles

FIGURE 9.14 – *Nombre de souscripteurs âgés de k années, $k = 18, 19, \ldots$.*

que `AGES`) se rapportent souvent au preneur d'assurance, qui n'est pas nécessairement le conducteur du véhicule. Dès lors, les conclusions obtenues sur base de ces données doivent être considérées avec précaution.

Examinons à présent la structure des âges des assurés dans notre portefeuille. Celle-ci est décrite à la Figure **??**. Calculons à présent la fréquence observée de sinistre par âge. Il s'agit d'un modèle de régression de Poisson avec une seule variable explicative, à savoir l'âge de l'assuré traité comme une variable catégorielle. Si on note λ_k la fréquence annuelle de sinistres des assurés âgés de k années, la vraisemblance associée aux données vaut

$$\mathcal{L} = \prod_{i=1}^{n} \exp(-d_i \lambda_{\texttt{AGES}_i}) \frac{(d_i \lambda_{\texttt{AGES}_i})^{n_i}}{n_i!}$$

où $\texttt{AGES}_i$ représente l'âge de l'assuré i, d_i est la durée d'exposition au risque pour la police i, et n_i le nombre de sinistres relatifs à cette police. Maximiser la log-vraisemblance revient à résoudre le système

$$\frac{\partial}{\partial \lambda_k} \left\{ \sum_{i|\texttt{AGES}_i=k} d_i \lambda_k + \sum_{i|\texttt{AGES}_i=k} n_i \ln \lambda_k \right\} = 0,$$

qui fournit les estimateurs du maximum de vraisemblance des fréquences de sinistre à chaque âge, donnés par

$$\widehat{\lambda}_k = \frac{\sum_{i|\texttt{AGES}_i=k} n_i}{\sum_{i|\texttt{AGES}_i=k} d_i}.$$

On constate donc qu'il faut diviser le nombre de sinistres par l'exposition au risque totale pour chaque âge, et non par le nombre de polices. La Figure **??** donne $\widehat{\lambda}_k$ en fonction de l'âge k. On constate clairement la sursinistralité des jeunes conducteurs, dont les fréquences de sinistre frôlent les 30%. Avec l'âge, la fréquence annuelle de sinistre a tendance à diminuer, pour légèrement augmenter chez les plus âgés.

Regardons à présent si `AGES` peut expliquer certaines variations du coût des sinistres. Pour ce faire, nous restreignons notre étude aux assurés qui ont déclaré au moins un sinistre. La moyenne de la

FIGURE 9.15 – *Fréquence observée $\widehat{\lambda}_k$ de sinistre par âge.*

FIGURE 9.16 – *Coût moyen par sinistre en fonction de l'âge de l'assuré.*

variable `CBAR` par âge est décrite à la Figure **??**. Aucune structure particulière n'est décelée à l'examen du graphique, ce qui tend à indiquer que la variable `AGES` influence peut-être la fréquence des sinistres mais peu leur coût.

Insistons sur le fait que l'analyse ci-dessus est purement marginale. L'apparente influence de l'âge sur la fréquence de sinistre pourrait ainsi être causée par une autre variable, fortement corrélée avec l'âge (nous renvoyons le lecteur aux commentaires du Tome I).

Variable `SEXE`

Cette variable qualitative binaire donne le sexe du souscripteur, i.e.

$$\texttt{SEXE} = \begin{cases} 0, \text{ si l'assuré est un homme,} \\ 1, \text{ si l'assuré est une femme.} \end{cases}$$

Le portefeuille contient 41 659 femmes (26.36%) et 116 402 hommes (73.64%). On peut voir à la Figure **??** une légère sursinistralité des femmes, et un coût moyen plus élevé pour les conductrices. Notez qu'on constate souvent le contraire en assurance RC automobile, les conductrices apparaissant beaucoup moins risquées que leurs homologues masculins. Ceci indique bien que les conclusions obtenues pour un portefeuille donné ne peuvent en général pas être extrapolées à d'autres portefeuilles. La raison en est simple : les assurés d'une compagnie ne constituent pas un échantillon aléatoire simple issu de la population des conducteurs (le choix de la compagnie révélant certaines caractéristiques des individus).

FIGURE 9.17 – *Répartition du portefeuille par sexe, fréquence moyenne de sinistre par sexe et coût moyen par sinistre par sexe (de gauche à droite).*

FIGURE 9.18 – *Répartition du portefeuille, fréquence moyenne de sinistre et coût moyen par sinistre par taille d'agglomération (de gauche à droite).*

Variable `AGGLOM`

Cette variable donne une idée de l'environnement dans lequel vit le preneur d'assurance. Précisément,

$$\texttt{AGGLOM} = \begin{cases} 0, \text{ si l'agglomération est rurale,} \\ 1, \text{ si l'agglomération est urbaine.} \end{cases}$$

Si la commune est fortement urbanisée, la plupart des voiries seront situées en agglomération, et la vitesse des véhicules sera limitée. De ce fait, la fréquence des sinistres devrait être élevée, du fait de la densité importante du trafic, mais les conséquences des sinistres devraient être faibles, du fait de la vitesse réduite. Au contraire, dans des communes rurales, la plupart des voiries seront situées hors agglomération, les sinistres seront plus rares, du fait de la faible densité du trafic, mais leurs conséquences devraient être plus graves, du fait de la vitesse plus élevée des véhicules.

Le portefeuille contient 113 189 assurés habitant dans des agglomérations rurales et 44 872 habitant dans des agglomérations urbaines, représentant respectivement 71.61 et 28.39 % du portefeuille. On peut voir à la Figure **??** que la fréquence moyenne de sinistre ainsi que le coût moyen sont plus faibles en zone rurale qu'en zone urbaine.

9.10.5 Caractéristiques du véhicule assuré

Variable `CARB`

Il s'agit d'une variable binaire prenant deux modalités, i.e.

$$\texttt{CARB} = \begin{cases} 0, \text{ si le véhicule est à essence,} \\ 1, \text{ si le véhicule est au diesel.} \end{cases}$$

Cette variable n'influence peut-être pas directement la sinistralité, mais on peut raisonnablement s'attendre à ce qu'elle soit fortement corrélée avec le kilométrage annuel (l'acquisition d'un véhicule diesel en Belgique ou en France se justifiant souvent par un usage plus fréquent, et pour de plus longues distances puisque le prix d'achat est plus élevé, de même que la taxe de circulation, mais que le prix

FIGURE 9.19 – *Répartition du portefeuille par type de carburant, fréquence moyenne de sinistre et coût moyen par sinistre par type de carburant (de gauche à droite).*

du litre de carburant est moins élevé). En pratique, `CARB` présente souvent d'autres niveaux (associés aux autres types de carburant, comme le gaz liquide ou l'électricité, par exemple).

Dans notre portefeuille, 109 139 polices concernent des véhicules à essence (soit 69.05%) et 48 922 (soit 30.95%) des véhicules diesel. On peut voir à la Figure **??** que la fréquence des sinistres est plus élevée pour les véhicules diesel, mais que le coût moyen pour ces mêmes véhicules est plus faible. La fréquence plus élevée pour les véhicules diesel pourrait s'expliquer par une exposition au risque (en kilomètres) plus importante.

Variable `KW`

Il s'agit d'une variable catégorielle donnant une idée de la puissance du véhicule, mesurée en kilowatts. L'idée est ici que plus un véhicule est puissant, plus il se révèle difficile à manier pour l'automobiliste. Nous disposons ici d'une partition des véhicules en trois classes, à savoir

$$\texttt{KW} = \begin{cases} 0, \text{ si le véhicule est une petite cylindrée} \\ \qquad\qquad\qquad\qquad \text{(i.e. moins de 40 Kw)}, \\ 1, \text{ si le véhicule est une cylindrée moyenne} \\ \qquad\qquad\qquad\qquad \text{(i.e. entre 40 et 70 Kw)}, \\ 2, \text{ si le véhicule est une grosse cylindrée} \\ \qquad\qquad\qquad\qquad \text{(i.e. plus de 70Kw)}. \end{cases}$$

Dans le portefeuille, il y a 36 417 véhicules de petite cylindrée (soit 23.04%), 94 581 véhicules de cylindrée moyenne (soit 59.84% du portefeuille) et 27 063 véhicules de grosse cylindrée (soit 17.12%). Afin d'apprécier l'influence de la puissance du véhicule sur la sinistralité, on peut voir à la Figure **??** la fréquence de sinistre observée et le coût moyen par sinistre en fonction de la puissance. On constate que la fréquence, comme le coût, semble croître avec la cylindrée du véhicule.

FIGURE 9.20 – *Répartition du portefeuille, fréquence moyenne de sinistre et coût moyen par sinistre par puissance du véhicule (de gauche à droite).*

FIGURE 9.21 – *Répartition du portefeuille, fréquence moyenne de sinistre et coût moyen par sinistre par usage du véhicule (de gauche à droite).*

Variable `USAGE`

Il s'agit d'une variable binaire décrivant l'usage du véhicule. Plus précisément,

$$\texttt{USAGE} = \begin{cases} 0, \text{ si le véhicule est utilisé exclusivement à des fins privées et pour se rendre au travail,} \\ 1, \text{ si le véhicule est utilisé à des fins professionnelles.} \end{cases}$$

Encore une fois, plus que l'influence directe de cette variable sur la sinistralité, on s'attend davantage à une forte corrélation entre cette variable et le kilométrage annuel.

Dans le portefeuille, 151 002 polices (soit 95.53%) concernent des véhicules qui sont utilisés à des fins privées (loisirs et trajet du domicile vers le lieu de travail) alors que 7 059 polices (soit 4.47%) couvrent des véhicules utilisés dans le cadre d'une activité professionnelle. On contate à la Figure **??** que l'usage du véhicule semble n'avoir que peu d'impact sur la fréquence des sinistres, mais influence sensiblement le coût moyen.

Variable `SPORT`

Cette variable identifie les véhicules à caractère sportif :

$$\texttt{SPORT} = \begin{cases} 0, \text{ si le véhicule a un caractère sportif,} \\ 1, \text{ sinon.} \end{cases}$$

La définition d'un véhicule sportif est propre à chaque compagnie. Elle est basée sur des caractéristiques techniques de l'automobile et fait généralement intervenir, outre la puissance, le poids et le nombre de sièges, notamment. Comme nous l'avons expliqué plus haut, l'ACP permet de construire un indice résumant le degré sportif du véhicule.

FIGURE 9.22 – *Répartition du portefeuille, fréquence moyenne de sinistre et coût moyen par sinistre par code sportif (de gauche à droite).*

FIGURE 9.23 – *Répartition du portefeuille, fréquence moyenne de sinistre et coût moyen par sinistre par périodicité du paiement de la prime (de gauche à droite).*

Dans notre portefeuille, 1 448 véhicules (soit 0.92%) sont considérés comme sportifs, contre 156 613 chez qui aucun caractère sportif n'est décelé (99.08%). On constate à la Figure **??** que les véhicules jugés sportifs ont une fréquence de sinistres plus élevée, mais que le coût moyen semble peu influencé par le code sportif.

9.10.6 Caractéristiques de la police

Variable `FRAC`

il s'agit d'une variable qualitative à deux niveaux spécifiant la fréquence de paiement de la prime, i.e.

$$\texttt{FRAC} = \begin{cases} 0, & \text{si la prime est payée une fois par an,} \\ 1, & \text{si le paiement est fractionné.} \end{cases}$$

Encore une fois, on voit mal comment cette variable pourrait influencer directement la sinistralité. L'idée est ici qu'un fractionnement de la prime peut traduire une position sociale fragilisée, donc un véhicule moins bien entretenu. Il faut toutefois se méfier des paiements mensuels, car certaines compagnies proposent à leurs clients de verser la prime chaque mois (par domiciliation et sans frais supplémentaires).

Dans notre portefeuille, on constate que 77 716 polices (soit 49.17% du portefeuille) prévoient une périodicité annuelle et 80 345 un paiement fractionné (50.83%). La Figure **??** illustre l'influence du fractionnement sur la sinistralité des assurés. Les assurés payant leur prime une fois par an ont une fréquence de sinistre inférieure à celle de ceux ayant opté pour le fractionnement. Les coûts moyens sont fort semblables.

Variable `GARACCESS`

Cette variable qualitative à deux niveaux décrit le degré de couverture de l'assuré :

$$\texttt{GARACCESS} = \begin{cases} 0, \text{ si seule l'assurance RC a été souscrite} \\ \qquad \text{par l'assuré} \\ 1, \text{ si en plus de la RC, l'assuré a souscrit} \\ \qquad \text{des garanties accessoires, telles que} \\ \qquad \text{dégâts matériels ou vol, par exemple;} \end{cases}$$

On peut ici s'attendre à deux types de comportement :

- Soit la théorie de l'aléa moral s'applique, et on observe plus de sinistres de la part des individus les plus couverts. De la même manière, en vertu de la théorie du signal, les assurés ayant souscrit uniquement la RC l'ont fait en connaissance de cause, se sachant moins risqués. La modalité 0 de `GARACCESS` devrait donc apparaître comme un facteur diminuant la sinistralité.
- Soit la souscription de garanties plus étendues traduit une aversion au risque plus forte de la part des individus concernés. Ceci se traduirait alors par un mode de conduite particulièrement prudent, et donc donnerait lieu à moins de sinistres. Dans ce cas, la modalité 0 de `GARACCESS` devient un facteur aggravant.

Les actuaires, par essence pragmatiques, n'essaient pas de déterminer quel mécanisme prévaut et se bornent à constater sur base des données quel est l'effet des différentes modalités de `GARACCESS`.

Dans notre portefeuille, 93 194 assurés ont souscrit exclusivement la RC automobile (soit 58.96% du portefeuille) tandis que les 64 867 autres (représentant 41.04% du portefeuille) ont souscrit des garanties accessoires. Au niveau de l'influence de l'étendue des garanties sur la sinistralité, on peut voir à la Figure **??** la fréquence de sinistre observée et le coût moyen par sinistre en fonction de l'étendue des couvertures. On constate que la fréquence et le coût moyen par sinistre sont plus élevés pour les polices ne reprenant que la RC. Ceci semble donc justifier techniquement les réductions de prime RC octroyées aux assurés souscrivant d'autres garanties optionnelles.

9.10.7 Interaction entre variables tarifaires

Souvent, on constate une interaction entre l'âge et le sexe de l'assuré. Par interaction, il faut entendre ici une infuence différente

FIGURE 9.24 – *Répartition du portefeuille, fréquence moyenne de sinistre et coût moyen par sinistre par étendue des couvertures (de gauche à droite).*

FIGURE 9.25 – *Fréquence de sinistre par âge selon le sexe du souscripteur.*

de l'âge de l'assuré sur la fréquence de sinistre, selon qu'on est un homme ou une femme. Dans notre portefeuille, si nous distinguons les hommes et les femmes, nous obtenons les fréquences de sinistre par âge décrites dans la Figure **??**. On constate une légère sursinistralité des jeunes hommes par rapport aux jeunes femmes, et ensuite des fréquences assez semblables. Nous ne poursuivons pas ici la modélisation des interactions.

9.10.8 Premier tri parmi les variables tarifaires

Test khi-carré d'indépendance

Une manière efficace d'opérer un premier tri parmi les variables à notre disposition consiste à effectuer des tests khi-carrés sur base de tables de contingence. Pour des raisons d'effectifs, il vaut mieux travailler avec `IND` qu'avec `NSIN`. En effet, si nous croisons `NSIN` avec `SEXE`, beaucoup d'effectifs attendus sont inférieurs à 5 ; on ne peut donc pas valablement effectuer de test khi-carré sur la base d'une telle table de contingence. Par contre, si nous croisons `IND` et `SEXE`, nous obtenons la table de contingence décrite au Tableau **??** (les effectifs attendus sous l'hypothèse d'indépendance sont indiqués entre parenthèses), sur laquelle nous pouvons baser le test khi-carré. La valeur observée de la statistique khi-carrée d'indépendance vaut 20.33 (pour 1 degré de liberté), ce qui donne une p-valeur inférieure à 10^{-4}. On constate donc une forte association entre le sexe et le fait de causer ou pas de sinistre.

Le rejet de l'hypothèse d'indépendance indique que le sexe influence la variable `IND` et donc également la variable `NSIN`. Le non-rejet par contre ne permet pas de conclure, d'une part pour des raisons liées à l'erreur de seconde espèce (laquelle n'est pas contrôlée) et d'autre part car le fait de ne pas influencer `IND` n'implique pas qu'on n'influence pas `NSIN` (car la variable explicative peut influencer la

IND	SEXE		
	Femme	Homme	Total
Pas de sinistre	36 722 (36 972)	103 554 (103 304)	140 276
Un ou plusieurs sinistres	4 937 (4 687.5)	12 848 (13 098)	17 785
Total	41 659	116 402	158 061

TABLE 9.5 – Table de contingence croisant IND et SEXE.

répartition dans les catégories 1,2,... sinistres).

En croisant successivement toutes les variables tarifaires avec IND, on obtient les résultats de la Table **??**. On voit ainsi que USAGE ne semble pas influencer IND. Nous pourrions donc exclure dès à présent cette variable de la suite de l'étude.

Variable	Val. observée de la stat. χ^2	# dl	p-valeur
FRAC	375.47	3	$< .0001$
CARB	139.22	1	$< .0001$
SPORT	6.73	1	0.0094
GARACCESS	22.60	1	$< .0001$
AGGLOM	202.04	2	$< .0001$
USAGE	0.07	1	0.7956
KW	23.40	2	$< .0001$

TABLE 9.6 – Résultats des tests khi-carrés sur les tables de contingence croisant les variables tarifaires et IND.

Dépendances vraie et apparente

Notez que le rejet de l'hypothèse d'indépendance entre SEXE et IND indique que le sexe de l'assuré influence la probabilité d'avoir au moins un sinistre sur la période. Il peut néanmoins s'agir d'une dépendance vraie, auquel cas la probabilité d'occurrence dépend vraiment du sexe de l'assuré, ou d'une dépendance apparente, auquel cas la probabilité d'avoir au moins un sinistre dépend d'une variable corrélée avec le sexe (qu'elle soit cachée, comme l'agressivité au volant par exemple, ou observable, comme l'âge de l'assuré, si les structures d'âge diffèrent entre hommes et femmes). Dans ce

dernier cas de figure, la dépendance entre le sexe de l'assuré et le fait d'avoir ou pas des sinistres disparaîtrait si on tenait compte de la troisième variable.

9.10.9 Analyse des fréquences de sinistre

Modèle de régression de Poisson pour le nombre des sinistres

Soit N_i le nombre de sinistres déclarés par l'assuré i durant l'année 1997, $i = 1, 2, \ldots, n$. Nous notons d_i l'exposition au risque pour l'individu i. Le modèle de Poisson suppose que la loi conditionnelle de N_i sachant $\boldsymbol{x}_i$ est de Poisson. Il suffit dès lors de spécifier sa moyenne $\mathbb{E}[N_i|\boldsymbol{x}_i]$. Comme cette dernière est strictement positive, on retient généralement une moyenne de forme exponentielle linéaire, i.e.

$$\mathbb{E}[N_i|\boldsymbol{x}_i] = d_i \exp(\boldsymbol{\beta}^t \boldsymbol{x}_i), \quad i = 1, 2, \ldots, n, \tag{9.44}$$

où $\boldsymbol{\beta}$ est un vecteur de coefficients de régression inconnus. Nous travaillons donc dans le modèle

$$N_i \sim \mathcal{P}oi\big(d_i \exp(\boldsymbol{\beta}^t \boldsymbol{x}_i)\big), \quad i = 1, 2, \ldots, n. \tag{9.45}$$

Lorsque toutes les variables sont catégorielles, chaque assuré est donc représenté par un vecteur $\boldsymbol{x}_i$ dont les composantes valent 0 ou 1. Dans ce cas, la fréquence annuelle $\lambda_i = \exp(\boldsymbol{\beta}^t \boldsymbol{x}_i)$ apparaît comme un produit de coefficients de majoration ou de réduction par rapport à la fréquence de l'individu de référence du portefeuille. Plus précisément,

$$\begin{aligned} \lambda_i &= \exp(\boldsymbol{\beta}^t \boldsymbol{x}_i) \\ &= \exp(\beta_0) \prod_{j=1}^{p} \exp(\beta_j x_{ij}) \\ &= \exp(\beta_0) \prod_{j|x_{ij}=1} \exp(\beta_j). \end{aligned}$$

Dès lors, $\exp(\beta_0)$ est la fréquence annuelle de l'individu de référence du portefeuille tandis que chacun des facteurs $\exp(\beta_j)$ traduit l'influence d'un critère de segmentation (si $\beta_j > 0$, les individus présentant cette caractéristique subiront une majoration de prime par rapport à la prime de référence $\exp(\beta_0)$, tandis que $\beta_j < 0$ indiquera une réduction de prime).

Parameter	Estimate	Error	t Value	Pr > |t|
Intercept	-1.42117	0.08396	-16.93	< .0001
CARB Diesel	0.17744	0.01582	11.21	< .0001
CARB Essence	0	.	.	.
FRAC Annuel	-0.20345	0.01475	-13.79	< .0001
FRAC Fractionné	0	.	.	.
SPORT Non sportif	-0.02708	0.06983	-0.39	0.6981
SPORT Sportif	0	.	.	.
GARACCESS RC Seule	0.09806	0.01494	6.56	< .0001
GARACCESS RC+Accessoires	0	.	.	.
AGGLOM Rural	-0.22783	0.01519	-15	< .0001
AGGLOM Urbain	0	.	.	.
SEXE Femme	0.05386	0.01646	3.27	0.0011
SEXE Homme	0	.	.	.
KW Cylindrée moyenne	0.09516	0.01852	5.14	< .0001
KW Grosse cylindrée	0.20646	0.02393	8.63	< .0001
KW Petite cylindrée	0	.	.	.
USAGE Privé	-0.03167	0.03468	-0.91	0.3612
USAGE Professionnel	0	.	.	.
Linear(AGES)	-0.01267	0.00050444	-25.11	< .0001

TABLE 9.7 – Estimation des paramètres intervenant dans la partie linéaire du modèle Poisson GAM.

Les paramètres possèdent ainsi l'interprétation suivante : si on considère le caractère codé par la $j$ème variable binaire,

$$\frac{\mathbb{E}[N_i|\text{ caractère présent}]}{\mathbb{E}[N_i|\text{ caractère absent}]} = \exp(\beta_j).$$

Sur base de cette dernière équation, $\exp(\beta_j)$ est donc le facteur par lequel il faut multiplier la fréquence de sinistre des individus ne présentant pas le caractère codé par la $j$ème variable binaire pour obtenir celle des individus présentant ce caractère.

Traitement de l'âge par un modèle additif généralisé

Nous commençons par un modèle de régression de Poisson de type GAM afin de déceler l'influence de la variable `AGES`. Plus précisément, l'effet de l'âge du souscripteur sur la fréquence annuelle de sinistre est traduite par une fonction $f_{\texttt{AGES}}$ décomposée à l'aide de splines cubliques, puis estimée sur base des statistiques disponibles. On peut voir à la Table **??** l'ajustement de la partie linéaire du modèle. On y apprend que le code sportif n'est pas pertinent (p-valeur de 0.6981). De la même manière (p-valeur de 0.3612), la variable `USAGE` pourrait être retirée du modèle. Le Tableau **??** compare les modèles avec et sans la partie non-linéaire en `AGES`. On constate que l'influence de cette variable tarifaire continue est clairement non-linéaire (p-valeur de moins de 0.0001).

L'omission des variables `SPORT` et `USAGE` ne détériore pas significativement le modèle (la déviance passe de 85 798.498 à 85 799.492). On peut voir les résultats de l'ajustement au Tableau **??**. Le Tableau

Component	Smoothing Parameter	DF	GCV	Num Unique Obs	Chi-Square	Pr > ChiSq
Spline(AGES)	0.999630	13.629696	0.6607040	78	511.6004	< .0001

TABLE 9.8 – Ajustement de la partie non-linéaire du Poisson GAM.

Parameter	Estimate	Error	t Value	Pr > |t|
Intercept	-1.47918	0.03291	-44.95	< .0001
CARB Diesel	0.17828	0.01576	11.31	< .0001
CARB Essence	0	.	.	.
FRAC Annuel	-0.20299	0.01475	-13.76	< .0001
FRAC Fractionné	0	.	.	.
GARACCESS RC Seule	0.09752	0.01492	6.54	< .0001
GARACCESS RC+Accessoires	0	.	.	.
AGGLOM Rural	-0.22785	0.01519	-15	< .0001
AGGLOM Urbain	0	.	.	.
SEXE Femme	0.05378	0.01646	3.27	0.0011
SEXE Homme	0	.	.	.
KW Cylindrée moyenne	0.0955	0.01851	5.16	< .0001
KW Grosse cylindrée	0.21003	0.02345	8.96	< .0001
KW Petite cylindrée	0	.	.	.
Linear(AGES)	-0.01267	0.00050334	-25.17	< .0001

TABLE 9.9 – Estimation des paramètres intervenant dans la partie linéaire du modèle Poisson GAM sans les variables `SPORT` et `USAGE`.

?? nous apprend que l'effet non-linéaire de `AGES` est toujours très significatif.

Examinons à présent l'influence de `AGES` sur la fréquence annuelle de sinistres. La Figure **??** nous renseigne à ce propos : on y a porté en graphique le coefficient par lequel doit être multiplié la fréquence de la classe de référence en fonction de l'âge du souscripteur, soit

$$\exp\left(\widehat{\beta}_{\texttt{AGES}}\texttt{AGES} + \widehat{f}_{\texttt{AGES}}(\texttt{AGES})\right).$$

Sur base de la Figure **??**, on décide de catégoriser la variable `AGES` comme suit :

$$\texttt{AGESGROUP} = \begin{cases} \text{Débutant, si } 18-21 \text{ ans,} \\ \text{Jeune, si } 22-30 \text{ ans,} \\ \text{Expérimenté, si } 31-55 \text{ ans,} \\ \text{Senior, si } 56-75 \text{ ans,} \\ \text{Conducteur âgé, si plus de 75 ans.} \end{cases}$$

Component	Smoothing Parameter	DF	GCV	Num Unique Obs	Chi-Square	Pr > ChiSq
Spline(AGES)	0.999646	13.440987	0.678901	78	486.5207	< .0001

TABLE 9.10 – Ajustement de la partie non-linéaire du Poisson GAM sans les variables `SPORT` et `USAGE`.

FIGURE 9.26 – *Influence de* `AGES` *sur la fréquence de sinistres.*

FIGURE 9.27 – *Répartition du portefeuille, fréquence moyenne de sinistres et coût moyen par sinistre par groupe d'âges (de gauche à droite).*

Bien sûr ce choix comporte une part d'arbitraire. Examinons a présent la variable `AGESGROUP`. Le portefeuille contient 1 212 conducteurs débutants (0.77%), 22 685 jeunes (14.35%), 89 326 expérimentés (56.51%), 40 199 seniors (25.43%) et 4 639 conducteurs âgés (2.93%). Un test khi-carré d'indépendance entre `IND` et `AGESGROUP` conduit au rejet de celle-ci (valeur observée de la statistique de test 860.15 pour 4 degrés de liberté, soit une p-valeur inférieure à 0.0001). On contate à la Figure **??** que la classe la plus représentée est celle des conducteurs expérimentés. Nous visualisons également l'allure "en cuvette" du nombre et des montants moyens de sinistre : des valeurs élevées pour les classes de jeunes, une diminution évidente au fil des ans, jusqu'à une légère augmentation pour la classe des conducteurs âgés.

Estimation des fréquences annuelles de sinistre

Maintenant que la variable `AGES` a été catégorisée, nous sommes ramenés à un modèle de régression de Poisson de type GLM. L'ajustement du modèle est donné à la Table **??**, tandis que la Table **??** donne les p-valeurs des tests consistant à exclure une à une les variables du modèle (analyse de type 3 dans le jargon de SAS). Les estimations ponctuelles des β_j sont fournies dans la troisième colonne du Tableau **??**, les deux premières permettant d'identifier le niveau auquel le coefficient de régression se rapporte. Les lignes où apparaissent des 0 correspondent aux niveaux de référence des différentes variables tarifaires. La colonne "Wald 95% Conf Limit" reprend les bornes inférieure et supérieure des intervalles de confiance pour les paramètres au niveau 95%, calculées à l'aide de la formule

$$\text{Coeff}\,\beta_j \pm 1.96\,\text{Std Error}\,\beta_j,$$

où 1.96 est le quantile d'ordre 97.5% de la loi normale centrée réduite et Std Error est la racine du $j$ème élément diagonal de $\widehat{\mathbf{\Sigma}}$ donnée dans la quatrième colonne.

Les colonnes “Chi-Sq” et “Pr>ChiSq”, qui est la p-valeur associée, permettent de tester si le coefficient β_j correspondant est significativement différent de 0. Ce test est effectué grâce à la statistique de Wald

$$\frac{(\text{Coeff}\,\beta_j)^2}{(\text{Std Error}\,\beta_j)^2},$$

qui obéit approximativement à la loi khi-carrée à 1 degré de liberté. On rejettera la nullité de β_j lorsque la p-valeur est inférieure à 5%.

L'analyse de Type 3 permet d'examiner la contribution de chacune des variables par rapport à un modèle ne la contenant pas. Dans la colonne “ChiSquare” est calculée, pour chaque variable, 2 fois la différence entre la log-vraisemblance obtenue pour le modèle contenant toutes les variables et la log-vraisemblance du modèle sans la variable en question. Cette statistique est asymptotiquement distribuée comme une Chi-carrée avec DF degrés de liberté, où DF est le nombre de paramètres associés à la variable explicative examinée. La dernière colonne nous fournit la p-valeur associée au test du rapport de vraisemblance ; cela permet d'apprécier la contribution de cette variable explicative à la modélisation du phénomène étudié.

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-1.9575	0.0176	-1.992	-1.923	12387.1	< .0001
CARB Diesel	1	0.1869	0.0159	0.1558	0.2179	138.68	< .0001
CARB Essence	0	0	0	0	0	.	.
FRAC Annuel	1	-0.277	0.0147	-0.3058	-0.2481	353.8	< .0001
FRAC Fractionné	0	0	0	0	0	.	.
SPORT Sportif	1	0.0391	0.0698	-0.0978	0.176	0.31	0.5754
SPORT Non sportif	0	0	0	0	0	.	.
GARACCESS RC+Accessoires	1	-0.1342	0.0149	-0.1635	-0.1049	80.68	< .0001
GARACCESS RC Seule	0	0	0	0	0	.	.
AGGLOM Urbain	1	0.2376	0.0152	0.2078	0.2674	244.51	< .0001
AGGLOM Rural	0	0	0	0	0	.	.
SEXE Femme	1	0.0701	0.0165	0.0378	0.1023	18.13	< .0001
SEXE Homme	0	0	0	0	0	.	.
KW Grosse cylindrée	1	0.1274	0.0203	0.0876	0.1672	39.35	< .0001
KW Petite cylindrée	1	-0.0867	0.0185	-0.1231	-0.0504	21.91	< .0001
KW Cylindrée moyenne	0	0	0	0	0	.	.
AGESGRP Débutant	1	0.7902	0.0578	0.6769	0.9034	186.92	< .0001
AGESGRP Conducteur âgé	1	-0.189	0.0482	-0.2135	-0.0244	6.09	0.1360
AGESGRP Jeune	1	0.3891	0.0184	0.353	0.4252	446.29	< .0001
AGESGRP Senior	1	-0.232	0.0192	-0.2696	-0.1943	145.77	< .0001
AGESGRP Expérimenté	0	0	0	0	0	.	.
USAGE Professionnel	1	0.0269	0.0347	-0.0411	0.0949	0.6	0.4386
USAGE Privé	0	0	0	0	0	.	.

Ajustement du modèle de régression de Poisson.

On commence par exclure la variable `SPORT`, jugée la moins significative (p-valeur de 57.75% dans le Tableau **??**). Le modèle obtenu (résultats non donnés) continue de donner une p-valeur élevée à la variable `USAGE` (44.28%). C'est pourquoi nous passons tout de suite au modèle suivant, en ôtant la variable `USAGE`. Ensuite, afin de simplifier le modèle de tarification, nous nous demandons s'il ne serait pas possible de grouper certains niveaux de la variable `AGESGROUP`.

Source	DF	Chi-Square	Pr > ChiSq
CARB	1	136.63	< .0001
FRAC	1	357.5	< .0001
SPORT	1	0.31	0.5775
GARACCESS	1	81.36	< .0001
AGGLOM	1	237.72	< .0001
SEXE	1	17.96	< .0001
KW	2	78.79	< .0001
AGESGRP	4	878.05	< .0001
USAGE	1	0.6	0.4404

TABLE 9.11 – Statistiques du rapport de vraisemblance pour analyse de type 3.

Contrast	DF	Chi-Square	Pr > ChiSq
Débutant-Jeune	1	41.21	<.0001
Conducteur âgé-Senior	1	5.06	0.2450
Jeune-Expérimenté	1	421.67	<.0001
Senior-Expérimenté	1	151.15	<.0001

TABLE 9.12 – Tests sur l'égalité des coefficients pris 2 à 2 pour la variable `AGESGRP`.

A cet effet, nous testons l'hypothèse d'égalité des coefficients pris deux par deux et dont les résultats sont repris dans la Table **??**. La p-valeur obtenue sur les niveaux "Senior" et "Conducteur âgé" indique clairement que nous pouvons regrouper ces niveaux (en un seul niveau senior) pour la suite de l'analyse. Le modèle final tenant compte de toutes ces modifications est repris aux Tables **??** et **??**.

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-1.9564	0.0176	-1.9909	-1.922	12398	< .0001
CARB Diesel	1	0.1856	0.0158	0.1547	0.2165	138.28	< .0001
CARB Essence	0	0	0	0	0	.	.
FRAC Annuel	1	-0.2751	0.0147	-0.3039	-0.2463	350.38	< .0001
FRAC Fractionné	0	0	0	0	0	.	.
GARACCESS RC+Accessoires	1	-0.134	0.0149	-0.1632	-0.1048	80.71	< .0001
GARACCESS RC Seule	0	0	0	0	0	.	.
AGGLOM Urbain	1	0.2377	0.0152	0.2079	0.2674	244.69	< .0001
AGGLOM Rural	0	0	0	0	0	.	.
SEXE Femme	1	0.0691	0.0164	0.0369	0.1014	17.67	< .0001
SEXE Homme	0	0	0	0	0	.	.
KW Grosse cylindrée	1	0.1296	0.0198	0.0907	0.1684	42.81	< .0001
KW Petite cylindrée	1	-0.086	0.0185	-0.1223	-0.0497	21.57	< .0001
KW Cylindrée moyenne	0	0	0	0	0	.	.
AGESGRP Débutant	1	0.7894	0.0578	0.6761	0.9026	186.65	< .0001
AGESGRP Jeune	1	0.3891	0.0184	0.3531	0.4252	448.41	< .0001
AGESGRP Senior	1	-0.2213	0.0185	-0.2576	-0.185	142.67	< .0001
AGESGRP Expérimenté	0	0	0	0	0	.	.

Ajustement du modèle de régression de Poisson, modèle final.

Pour l'assuré i caractérisé par un vecteur de variables explicatives $\boldsymbol{x}_i$, la fréquence annuelle prédite est $\exp(\boldsymbol{x}_i^t\widehat{\boldsymbol{\beta}})$. Ceci sera aussi le cas pour les nouveaux assurés présentant les mêmes caractéristiques que l'assuré i (l'hypothèse implicite étant que les nouvelles polices sont conclues par des individus s'identifiant parfaitement aux as-

Source	DF	Chi-Square	Pr > ChiSq
CARB	1	136.18	< .0001
FRAC	1	353.97	< .0001
GARACCESS	1	81.39	< .0001
AGGLOM	1	237.89	< .0001
SEXE	1	17.51	< .0001
KW	2	83.7	< .0001
AGESGRP	3	877.65	< .0001

TABLE 9.13 – Statistiques du rapport de vraisemblance pour analyse de type 3, modèle final.

surés qui sont à la base de la construction du tarif ; cela suppose notamment que la compagnie maîtrise parfaitement l'antisélection par une politique d'acceptation soignée).

Surdispersion

Le modèle de Poisson impose des contraintes assez fortes sur la dépendance entre la variable de comptage N_i et les facteurs de risque $\boldsymbol{x}_i$, puisque

$$\mathbb{E}[N_i|\boldsymbol{x}_i] = \mathbb{V}[N_i|\boldsymbol{x}_i] = d_i \exp(\boldsymbol{\beta}^t \boldsymbol{x}_i). \tag{9.46}$$

Ceci revient donc à supposer l'égalité entre le nombre moyen de sinistre et la variabilité de ce nombre au sein de chaque classe de risque. Remarquons cependant que la convergence des estimateurs du pseudo maximum de vraisemblance obtenus dans le modèle de Poisson permet de les utiliser, même si la loi de Poisson n'est pas adaptée (pour autant que la moyenne conditionnelle soit correctement spécifiée).

En pratique, afin de vérifier la validité de la contrainte (**??**), on calcule pour chaque classe de risque la moyenne et la variance empirique des nombres des sinistres, $\widehat{m}_k$ et $\widehat{\sigma}_k^2$, disons, et on porte le nuage de points $\{(\widehat{m}_k, \widehat{\sigma}_k^2),\ \ k = 1, 2, \ldots\}$ en graphique. Ceci permet de voir comment la variance évolue en fonction de la moyenne. Lorsque les points sont autour de la première bissectrice, on peut considérer que les deux premiers moments conditionnels sont égaux, ce qui conforte le modèle de Poisson. Dans le cas contraire, on observe souvent un phénomène de surdispersion, c'est-à-dire des classes pour lesquelles $\widehat{\sigma}_k^2 > \widehat{m}_k$. Ce phénomène est dû la plupart du temps à des variables omises.

Nous pouvons comprendre ce phénomène si nous considérons deux classes de risque C_1 et C_2 sans effet de surdispersion ($\widehat{\sigma}_1^2 = \widehat{m}_1$ et $\widehat{\sigma}_2^2 = \widehat{m}_2$), mais que l'on aurait omis de séparer. Dans la classe

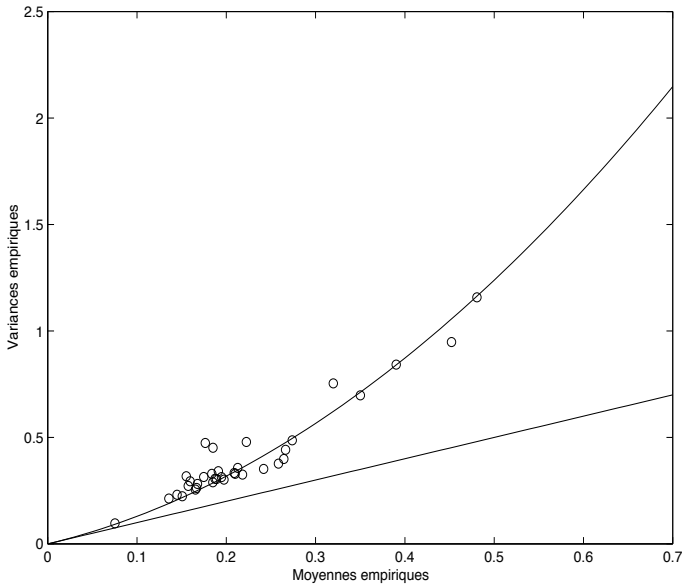


FIGURE 9.28 – *Moyennes et variances dans les classes de risque résultant de la régression de Poisson.*

$C_1 \cup C_2$, la moyenne vaut

$$\widehat{m} = p_1 \widehat{m}_1 + p_2 \widehat{m}_2$$

où p_1 et p_2 désignent les poids relatifs de C_1 et C_2, respectivement. La variance quant à elle passe à

$$\widehat{\sigma}^2 = p_1 \widehat{\sigma}_1^2 + p_2 \widehat{\sigma}_2^2 + p_1 (\widehat{m}_1 - \widehat{m})^2 + p_2 (\widehat{m}_2 - \widehat{m})^2.$$

On constate donc que dans $C_1 \cup C_2$, il y a surdispersion puisque $\widehat{\sigma}^2 > \widehat{m}$, l'égalité n'étant possible que si $\widehat{m}_1 = \widehat{m}_2$. On comprend donc aisément que l'oubli de variables explicatives importantes puisse conduire à une surdispersion des observations au sein des classes de risque.

Sur base de notre portefeuille, les points $\{(\widehat{m}_k, \widehat{\sigma}_k^2),\ \ k = 1, 2, \ldots\}$ sont représentés dans le plan à la Figure **??**. L'effet de surdispersion est clairement visible.

Conséquences d'une erreur de spécification dans le cas de la loi de Poisson

Comme nous l'avons dit plus haut, le modèle de Poisson est relativement contraignant, car il impose l'équidispersion des données.

Souvent, et c'est le cas pour notre jeu de données, cette contrainte n'est pas satisfaite. Il est donc intéressant d'examiner ce qu'il advient si le modèle de Poisson est mal spécifié. Concrètement, supposons à présent que la véritable forme de la moyenne conditionnelle soit

$$\mathbb{E}[N_i|\boldsymbol{x}_i] = d_i \exp(\boldsymbol{\beta}_0^t \boldsymbol{x}_i), \tag{9.47}$$

où $\boldsymbol{\beta}_0$ est la vraie valeur du paramètre, mais que la loi conditionnelle de N_i sachant $\boldsymbol{x}_i$ ne soit pas de Poisson. L'estimateur $\widehat{\boldsymbol{\beta}}$ obtenu en résolvant les équations de vraisemblance n'est donc plus un estimateur du maximum de vraisemblance, mais plutôt un estimateur calculé avec erreur de spécification. On parle alors d'estimateur du pseudo- (ou quasi-) maximum de vraisemblance. Néanmoins, l'estimateur du peudo-maximum de vraisemblance fondé sur le modèle de Poisson de moyenne $\exp(\boldsymbol{\beta}^t \boldsymbol{x}_i)$ est convergent pour la vraie valeur $\boldsymbol{\beta}_0$ si la moyenne est de la forme (**??**). Ceci provient des équations de vraisemblance obtenues dans le modèle de régression de Poisson, lesquelles peuvent encore s'écrire

$$\mathbb{E}\Big[\big(N_i - \mathbb{E}[N_i|\boldsymbol{x}_i]\big)\boldsymbol{x}_i\Big] = \mathbf{0}$$

$$\Leftrightarrow \mathbb{E}\Big[d_i \exp(\boldsymbol{\beta}_0^t \boldsymbol{x}_i) - d_i \exp(\widehat{\boldsymbol{\beta}}^t \boldsymbol{x}_i)\Big]\boldsymbol{x}_i = \mathbf{0}$$

ce qui garantit $\widehat{\boldsymbol{\beta}} \rightarrow_{proba} \boldsymbol{\beta}_0$ lorsque le nombre d'observations $n \rightarrow +\infty$. Seule la spécification correcte de $\mathbb{E}[N_i|\boldsymbol{x}_i]$ est nécéssaire pour obtenir des estimateurs convergents.

L'erreur de spécification n'a donc pas d'effet sur la convergence de l'estimateur, mais oblige à modifier le calcul de la variance asymptotique, qui est à présent donnée par $\boldsymbol{H}^{-1}\mathcal{I}\boldsymbol{H}$ où

$$\boldsymbol{H} = \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^t d_i \exp(\boldsymbol{\beta}_0^t \boldsymbol{x}_i) \text{ et } \mathcal{I} = \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^t \mathbb{V}[N_i|\boldsymbol{x}_i].$$

En pratique, on estimera la matrice variance-covariance asymptotique par $\widehat{\boldsymbol{H}}^{-1}\widehat{\mathcal{I}}\widehat{\boldsymbol{H}}$ où

$$\widehat{\boldsymbol{H}} = \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^t d_i \exp(\widehat{\boldsymbol{\beta}}^t \boldsymbol{x}_i) \text{ et } \widehat{\mathcal{I}} = \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^t (n_i - \lambda_i)^2.$$

La surdispersion (qui indique une erreur de spécification) affecte donc peu les estimations ponctuelles en grand échantillon. Néanmoins, la surdispersion entraîne une sous-estimation des variances des estimateurs, produisant ainsi des intervalles de confiance

trop étroits et surestimant les statistiques khi-carrées utilisées pour tester la nullité des coefficients de régression. Il se pourrait dès lors qu'une variable jugée pertinente dans le modèle de Poisson ne le soit plus après que la surdispersion ait été prise en compte.

Modèle de régression binomial négatif

Une technique simple et efficace permettant de rendre compte de cette surdispersion consiste à superposer un terme d'erreur aléatoire au prédicteur linéaire (ce qui revient à reconnaître l'hétérogénéité des assurés au sein de chaque classe de tarif : bien qu'identiques pour la compagnie, ceux-ci présentent néanmoins des profils de risque relativement différents). Si on note ϵ_i ce terme d'erreur, on supposera que

$$[N_i|\boldsymbol{x}_i, \epsilon_i] \sim \mathcal{P}oi\big(d_i \exp(\boldsymbol{\beta}^t \boldsymbol{x}_i + \epsilon_i)\big).$$

Posons $\Theta_i = \exp(\epsilon_i)$ et exigeons $\mathbb{E}[\Theta_i] = 1$. Ainsi,

$$\mathbb{E}[N_i|\boldsymbol{x}_i] = d_i \exp(\boldsymbol{\beta}^t \boldsymbol{x}_i)\mathbb{E}[\Theta_i] = d_i \exp(\boldsymbol{\beta}^t \boldsymbol{x}_i) = \lambda_i;$$

la superposition du terme d'erreur ϵ_i au prédicteur linéaire ne modifie donc pas la moyenne a priori de N_i, conditionnellement aux observables $\boldsymbol{x}_i$. Cette propriété garantit que le tarif est correct en moyenne, et donc que l'encaissement relatif à une classe de tarif suffira à dédommager les sinistres. Comme

$$\begin{aligned}\mathbb{V}[N_i|\boldsymbol{x}_i] &= \mathbb{E}\Big[\mathbb{V}[N_i|\boldsymbol{x}_i, \Theta_i]\Big] + \mathbb{V}\Big[\mathbb{E}[N_i|\boldsymbol{x}_i, \Theta_i]\Big] \\ &= \mathbb{E}[\lambda_i\Theta_i] + \mathbb{V}[\lambda_i\Theta_i] = \lambda_i\big\{1 + \sigma^2\lambda_i\big\} > \lambda_i = \mathbb{E}[N_i|\boldsymbol{x}_i]\end{aligned}$$

où l'on a posé $\sigma^2 = \mathbb{V}[\Theta_i]$, la superposition d'un terme d'erreur aléatoire au prédicteur linéaire entraîne d'office la surdispersion des données.

Donnons à présent une justification à l'introduction de l'erreur ϵ. Bon nombre de variables explicatives pertinentes ne peuvent être observées par l'assureur (pour des raisons légales ou économiques). Bien entendu, certaines variables cachées pourraient être corrélées avec les variables observables $\boldsymbol{X}$. Par exemple, le kilométrage annuel est une variable cachée pour la compagnie, mais celle-ci est vraisemblablement corrélée avec l'usage du véhicule (privé-professionnel) qui est quant à lui une variable observable par la compagnie. L'idée est donc de représenter l'effet résiduel des variables cachées par un terme d'erreur ϵ que l'on superposerait au score. Techniquement, ϵ est supposé indépendant du vecteur $\boldsymbol{X}$ des caractéristiques observables de l'assuré. Pour un individu tel que $\boldsymbol{X} = \boldsymbol{x}$, le nombre

annuel de sinistres N est donc de loi de Poisson de moyenne $\exp(\boldsymbol{\beta}^t\boldsymbol{x}+\epsilon)$ où ϵ est une variable aléatoire représentant l'influence des variables omises dans le tarif; ϵ est appelé effet aléatoire et représente l'hétérogénéité résiduelle du portefeuille. Le nombre annuel de sinistres N devient un mélange de lois de Poisson, traduisant le fait que chaque classe de risque (définie par $\boldsymbol{X}=\boldsymbol{x}$) contient un mélange d'assurés présentant des profils de risque différents sur les facteurs non-observables.

Si on note f_Θ la densité de probabilité de Θ_i, on obtient

$$\begin{aligned}\Pr[N_i=k|\boldsymbol{x}_i] &= \int_{\theta\in\mathbb{R}^+}\Pr[N_i=k|\boldsymbol{x}_i,\Theta_i=\theta]f_\Theta(\theta)d\theta\\ &= \int_{\theta\in\mathbb{R}^+}\exp(-\lambda_i\theta)\frac{(\lambda_i\theta)^k}{k!}f_\Theta(\theta)d\theta.\end{aligned}$$

Pour la plupart des choix de f_Θ, cette dernière intégrale n'admet pas d'expression explicite. Une exception notable est la densité associée à la loi Gamma, pour laquelle on obtient la loi Binomiale Négative pour $[N_i|\boldsymbol{x}_i]$. En effet, considérant que les effets aléatoires $\Theta_i=\exp\epsilon_i$, $i=1,2,\ldots,n$, sont indépendants et de même loi Gamma de moyenne 1 et de variance $1/a$, la densité de Θ_i est donnée par

$$f_\Theta(\theta)=\frac{1}{\Gamma(a)}a^a\theta^{a-1}\exp(-a\theta),\ \ \theta\in\mathbb{R}^+.$$

Le choix de la densité f_Θ se justifie par des considérations purement analytiques (la loi Gamma étant la loi conjuguée à la loi de Poisson). Conditionnellement aux variables observables $\boldsymbol{x}_i$ et à l'effet aléatoire Θ_i, la densité de probabilité discrète de N_i est donnée par

$$\Pr[N_i=n_i|\boldsymbol{x}_i,\Theta_i=\theta_i]=\exp\Big(-\theta_i d_i\exp(\eta_i)\Big)\frac{\big\{\theta_i d_i\exp(\eta_i)\big\}^{n_i}}{n_i!}.$$

Conditionnellement aux observables, N_i est donc de loi de Poisson de moyenne $d_i\exp(\eta_i)\Theta_i$. Non conditionnellement, N_i est de loi Binomiale Négative, et les probabilités sont

$$\begin{aligned}&\Pr[N_i=n_i|\boldsymbol{x}_i]\\ &= \binom{a+n_i-1}{n_i}\left(\frac{d_i\exp(\eta_i)}{a+d_i\exp(\eta_i)}\right)^{n_i}\left(\frac{a}{a+d_i\exp(\eta_i)}\right)^{a},\end{aligned}$$

pour $n_i\in\mathbb{N}$.

Bien entendu, l'estimateur du maximum de vraisemblance $\widehat{\boldsymbol{\beta}}$ dans le modèle de Poisson ne coïncide pas avec celui du modèle

avec effet aléatoire. Néanmoins, les estimations obtenues peuvent être considérées comme les résultats de la méthode des moments généralisés. En effet, l'équation (**??**) dont l'estimateur du maximum de vraisemblance $\widehat{\boldsymbol{\beta}}$ est solution peut se voir comme l'équivalent empirique de l'équation

$$\mathbb{E}\left[\sum_{i=1}^{n}\big\{N_i-\lambda_i\big\}\boldsymbol{x}_i\right]=\mathbf{0}, \tag{9.48}$$

laquelle est valable à la fois dans le modèle de Poisson simple et dans le modèle avec effet aléatoire. Dès lors, si on consent à abandonner la méthode du maximum de vraisemblance au profit de la méthode des moments généralisés, on est autorisé à conserver les estimations obtenues dans le modèle sans effet aléatoire.

Il est assez simple d'obtenir une estimation de $\sigma^2=\mathbb{V}[\Theta_i]$ à l'aide de la méthode des moments. Pour ce faire, remarquons que

$$\begin{aligned}\mathbb{V}[N_i] &= \mathbb{E}\Big[\mathbb{V}[N_i|\Theta_i]\Big]+\mathbb{V}\Big[\mathbb{E}[N_i|\Theta_i]\Big]\\ &= \mathbb{E}\Big[\Theta_i d_i\exp(\eta_i)\Big]+\mathbb{V}\Big[\Theta_i d_i\exp(\eta_i)]\Big]\\ &= d_i\exp(\eta_i)+\Big\{d_i\exp(\eta_i)\Big\}^2\sigma^2\\ &= \mathbb{E}[N_i]+\Big\{d_i\exp(\eta_i)\Big\}^2\sigma^2. \end{aligned}\tag{9.49}$$

Ecrivons une relation semblable à (**??**) pour la variance, à savoir

$$\sum_{i=1}^{n}\left\{\Big(n_i-d_i\exp(\eta_i)\Big)^2-n_i-\Big(d_i\exp(\eta_i)\Big)^2\sigma^2\right\}=0. \tag{9.50}$$

Les estimateurs $\widehat{\boldsymbol{\beta}}$ et $\widehat{\sigma}$ sont solutions du système formé par (**??**) et (**??**) ; ils sont convergents dans le modèle à effet aléatoire. Dès lors, l'estimateur du maximum de vraisemblance $\widehat{\boldsymbol{\beta}}$ de $\boldsymbol{\beta}$, solution de (**??**), est convergent dans le modèle à effet aléatoire. L'estimateur de σ^2 est quant à lui donné par

$$\widehat{\sigma}^2=\frac{\sum_{i=1}^{n}\left\{\Big(n_i-d_i\exp(\widehat{\eta}_i)\Big)^2-n_i\right\}}{\sum_{i=1}^{n}\Big\{d_i\exp(\widehat{\eta}_i)\Big\}^2}$$

où $\widehat{\eta}_i=\widehat{\boldsymbol{\beta}}^t\boldsymbol{x}_i$, $\widehat{\boldsymbol{\beta}}$ étant l'estimateur du maximum de vraisemblance de $\boldsymbol{\beta}$ dans le modèle sans effet aléatoire.

Plutôt que d'estimer d'abord les paramètres dans un modèle de régression de Poisson (**??**)-(**??**) et ensuite de superposer un effet aléatoire, sans remettre en cause l'estimation de $\boldsymbol{\beta}$, on pourrait également travailler directement dans un modèle de régression binomial négatif.

Si on retient la loi gamma pour Θ_i, la vraisemblance s'écrit

$$\mathcal{L}(\boldsymbol{\beta}|\boldsymbol{n}) = \prod_{i=1}^{n} \frac{\lambda_i^{n_i}}{n_i!} \left(\frac{a}{a+\lambda_i}\right)^a (a+\lambda_i)^{-n_i} \frac{\Gamma(a+n_i)}{\Gamma(a)}.$$

Les estimateurs du maximum de vraisemblance des paramètres $\boldsymbol{\beta}$ et de a sont solutions du système

$$\sum_{i=1}^{n} \boldsymbol{x}_i \left(n_i - \lambda_i \frac{a+n_i}{a+\lambda_i}\right) = \mathbf{0}. \tag{9.51}$$

Afin d'interpréter cette relation, remarquons que si une constante β_0 est introduite dans le score, la première équation de (**??**) garantit que

$$\sum_{i=1}^{n} n_i = \sum_{i=1}^{n} \lambda_i \frac{a+n_i}{a+\lambda_i}.$$

Ceci signifie que le modèle reproduit le nombre de sinistres en intégrant l'information fournie par ceux-ci. En effet, $\lambda_i \frac{a+n_i}{a+\lambda_i}$ est le nombre attendu de sinistres au cours de la période suivante pour un assuré de caractéristiques $\boldsymbol{x}_i$ qui aurait déclaré n_i sinistres au cours de la période d'assurance qui s'achève. Si la première variable vaut 1 si l'assuré est un homme et 0 sinon, il vient

$$\sum_{\text{hommes}} n_i = \sum_{\text{hommes}} \lambda_i \frac{a+n_i}{a+\lambda_i}.$$

On peut voir cette dernière relation comme une garantie de la non-subsidiation des hommes par les femmes, et inversément.

Les équations de vraisemblance (**??**) n'admettent pas de solution explicite ; on aura donc recours à des méthodes de résolution numérique, en utilisant les estimations fournies par la méthode des moments comme valeurs initiales.

L'ajustement du modèle de régression binomial négatif est donné aux Tableaux **??**-**??**. Comme on pouvait s'y attendre, les estimations ponctuelles sont semblables à celles obtenues par régression de Poisson, mais les intervalles de confiance sont plus larges dans le modèle binomial négatif.

Source	DF	Chi-Square	Pr > ChiSq
CARB	1	127.51	< .0001
FRAC	1	329.1	< .0001
GARACCESS	1	75.58	< .0001
AGGLOM	1	219.49	< .0001
SEXE	1	17.22	< .0001
KW	2	.	.
AGESGRP	3	813.99	< .0001

TABLE 9.14 – Statistiques du rapport de vraisemblance pour analyse de type 3, modèle binomial négatif.

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-1.9555	0.0182	-1.9913	-1.9198	11486.3	< .0001
CARB Diesel	1	0.1867	0.0164	0.1545	0.2189	129.24	< .0001
CARB Essence	0	0	0	0	0	.	.
FRAC Annuel	1	-0.2757	0.0152	-0.3055	-0.2458	326.86	< .0001
FRAC Fractionné	0	0	0	0	0	.	.
GARACCESS RC+Accessoires	1	-0.1341	0.0155	-0.1644	-0.1038	75.07	< .0001
GARACCESS RC Seule	0	0	0	0	0	.	.
AGGLOM Urbain	1	0.2375	0.0158	0.2065	0.2685	225.05	< .0001
AGGLOM Rural	0	0	0	0	0	.	.
SEXE Femme	1	0.0712	0.0171	0.0377	0.1048	17.36	< .0001
SEXE Homme	0	0	0	0	0	.	.
KW Grosse cylindrée	1	0.1297	0.0206	0.0894	0.1701	39.68	< .0001
KW Petite cylindrée	1	-0.0852	0.0192	-0.1228	-0.0475	19.66	< .0001
KW Cylindrée moyenne	0	0	0	0	0	.	.
AGESGRP Débutant	1	0.8006	0.0621	0.6788	0.9224	166.02	< .0001
AGESGRP Jeune	1	0.3906	0.0193	0.3528	0.4284	410.41	< .0001
AGESGRP Senior	1	-0.2213	0.0191	-0.2587	-0.1839	134.3	< .0001
AGESGRP Expérimenté	0	0	0	0	0	.	.
Dispersion	1	0.5431	0.0358	0.4773	0.6178		

Ajustement du modèle de régression binomial négatif, modèle final.

9.10.10 Analyse des coûts des sinistres

Difficultés

Avant de poursuivre, expliquons pourquoi l'analyse des coûts des sinistres est sensiblement plus compliquée que celle des nombres.

Si toutes les polices du portefeuille peuvent servir à estimer la fréquence annuelle de sinistres, il est clair que seules les polices sinistrées peuvent être utilisées en vue d'étudier la loi des montants des sinistres. L'actuaire dispose donc d'un nombre assez limité d'observations à exploiter pour ajuster un modèles aux montants de sinistres. De plus, les montants sont beaucoup plus difficiles à modéliser que les nombres, car le phénomène est beaucoup plus complexe.

Souvent, les sinistres de quelque gravité nécessitent des délais assez longs pour être cloturés. Pensez par exemple à un sinistre RC automobile avec dégâts corporels, où il faut attendre que l'état de la victime soit stabilisé avant de pouvoir déterminer le montant des indemnités. Pour cette raison, la compagnie ne dispose bien souvent

que de prévisions des coûts dans ses fichiers, ce qui rend l'analyse hasardeuse.

Selon certains auteurs, la prise en compte des montants des sinistres dans la tarification RC automobile est contestable. En effet, les montants payés par l'assureur sont versés aux tiers lésés au titre d'indemnisation de leurs préjudices. Ainsi, un assuré qui renverserait un piéton exposerait son assureur à des dépenses très différentes selon que ce piéton est un vieillard malade seul au monde ou un jeune cadre dynamique père de deux enfants en bas âge. On voit mal comment les caractéristiques de l'assuré pourraient expliquer ce phénomène, même si on comprend bien que le profil de l'assuré puisse expliquer la probabilité de causer un accident avec dégâts corporels. Dans le même ordre d'idées, un assuré qui accrocherait un autre véhicule obligerait son assureur à payer des indemnités très différentes selon que le véhicule tiers est une limousine ou une voiture bas de gamme. Il ne fait toutefois pas de doute que les montants des sinistres doivent être pris en considération dans les garanties connexes telles que la couverture "dégâts matériels" par exemple.

Sinistres graves

Souvent, moins de 20% des sinistres causent plus de 80% des dépenses de la compagnie. Ceci nécessite donc un traitement particulier de ces "sinistres graves", sur lesquels on ne segmente en général pas ou peu.

Différentes représentations de la charge des sinistres par police ont été examinées dans la Section 6.2.2. Nous adopterons ici le formalisme de l'Exemple 6.2.1. Plus précisément, la charge totale des sinistres produits par une police s'écrira comme suit :

$$S = \sum_{k=1}^{N} C_k + IL$$

où

N, est le nombre de sinistres standards, supposé de loi de Poisson ;

C_k, est le coût du $k$ème sinistre standard

I, indique si la police a généré au moins un sinistre grave

L, est le coût cumulé de ces sinistres graves, le cas échéant.

L'assureur segmentera son tarif sur $\mathbb{E}[N]$ et sur $\mathbb{E}[C_k]$, éventuellement sur $\mathbb{E}[I]$ mais en général pas sur $\mathbb{E}[L]$. Les sinistres graves sont trop peu nombreux pour autoriser une personnalisation des montants.

Mean	1638.510	100% Max	1.98957E+06
Median	522.584	99%	1.75898E+04
Mode	1426.379	95%	3.59446E+03
Std Deviation	17132.4456	90%	2.58231E+03
Skewness	91.7998607	75% Q3	1.42638E+03
Coeff Variation	1045.61099	50% Median	5.22584E+02
Interquartile Range	1284	25% Q1	1.42192E+02

TABLE 9.15 – Statistiques descriptives du montant de la prime pure.

CARB	AGESGROUP	Probabilité de ne pas causer de sinistre grave
Essence	Débutant-Jeune	0.9533
Essence	Senior-Expérimenté	0.962
Diesel	Débutant-Jeune	0.9435
Diesel	Senior-Expérimenté	0.954

TABLE 9.16 – Probabilités de ne pas causer de sinistres graves en fonction des caractéristiques de l'assuré.

Remarque 9.10.4. *Dans la mesure où les sinistres graves sont les plus coûteux (ils représentent plus de 80% des sommes payées par l'assureur) et se prêtent mal à la segmentation, la partie fortement segmentée de la prime ne devrait représenter que 20% de la somme payée par l'assuré. Dès lors, dans la mesure où les assurés semblent tous égaux devant les sinistres graves (à peu de choses près), la tension tarifaire constatée dans les tarifs commerciaux semble parfois discutable.*

Dans notre portefeuille, 17 785 assurés ont déclaré au moins un sinistre au cours de l'année 1997. La prime pure empirique s'élève à 203.38 € . On peut lire quelques statistiques descriptives au Tableau **??**. Le niveau au-delà duquel un sinistre est qualifié de "grave" est obtenu à l'aide de la théorie des valeurs extrêmes (présentée au Chapitre **??**).

Régression logistique pour l'analyse de l'occurrence des sinistres graves

La régression logistique permet d'expliquer l'occurrence des sinistres graves. La méthode Backward de sélection des variables explicatives conduit successivement à l'exclusion des variables suivantes : `USAGE`, `SEXE`, `KW`, `GARACCESS`, `SPORT`, `AGGLOM` et `FRAC`, ce qui donne finalement (après groupement de certaines modalités de la variable `AGESGROUP`) les résultats repris au Tableau **??**.

Analyse des coûts des sinistres standards

Modèle de régression Gamma Nous supposerons les coûts des sinistres $C_{i1}, C_{i2}, \ldots$ causés par l'assuré i indépendants et de même loi Gamma de moyenne

$$\mu_i = \mathbb{E}[C_{ik}|\boldsymbol{x}_i] = \exp(\boldsymbol{\beta}^t \boldsymbol{x}_i)$$

et de variance

$$\mathbb{V}[C_{ik}|\boldsymbol{x}_i] = \frac{\{\exp(\boldsymbol{\beta}^t \boldsymbol{x}_i)\}^2}{\nu}.$$

Remarque 9.10.5. *Souvent, seul le coût total $C_{i\bullet} = \sum_{k=1}^{n_i} C_{ik}$ est disponible, et pas le détail des différentes composantes de la somme C_{ik}. Dans ce cas, on travaillera avec le coût moyen $\overline{C}_{i\bullet} = C_{i\bullet}/n_i$. Il suffira alors de laisser le paramètre ν varier d'assuré à assuré en spécifiant $\nu_i = \nu\omega_i$, où le poids ω_i est le nombre de sinistre n_i.*

En négligeant les sinistres graves, la prime pure pour l'assuré i s'écrit

$$\mathbb{E}\left[\sum_{k=1}^{N_i} C_{ik}\right] = \mathbb{E}[N_i]\mathbb{E}[C_{i1}] = \exp\Big((\boldsymbol{\beta}_{\text{freq}} + \boldsymbol{\beta}_{\text{cost}})^t \boldsymbol{x}_i\Big),$$

qui fournit un tarif multiplicatif lorsque toutes les variables explicatives sont binaires. La prime pure s'élève alors à

$$\exp\Big((\boldsymbol{\beta}_{\text{freq}} + \boldsymbol{\beta}_{\text{cost}})^t \boldsymbol{x}_i\Big) + q_i \mathbb{E}[L],$$

où $q_i = \mathbb{E}[I_i]$ est la probabilité que l'assuré i cause au moins un sinistre grave.

Ajustement du modèle La méthode Backward de sélection des variables explicatives conduit à l'exclusion de plusieurs variables explicatives. Le Tableau **??** décrit le modèle final.

Remarque 9.10.6. *La régression Gamma n'est bien évidemment pas la seule manière de modéliser les coûts des sinistres. Dans le modèle log-normal, on suppose que le logarithme népérien des montants des sinistres suit une loi normale dont la moyenne est un prédicteur linéaire $\boldsymbol{\beta}^t \boldsymbol{x}_{it}$ et la variance une constante σ^2, i.e.*

$$\ln C_{ik} \sim \mathcal{N}or(\boldsymbol{\beta}^t \boldsymbol{x}_i, \sigma^2).$$

Parameter	Estimate	Wald 95% Confidence Limits		Pr > ChiSq
Intercept	6.6389	6.5908	6.6873	<.0001
FRAC	-0.0402	-0.0764	-0.0039	0.0296
FRAC	0	0	0	.
GARACCESS	0.0702	0.0342	0.1062	0.0001
GARACCESS	0	0	0	.
AGESGRP	-0.0657	-0.1095	-0.0222	0.0032
AGESGRP	0	0	0	.
AGGLOM	0.0826	0.0314	0.1345	0.0017
AGGLOM	0	0	0	.
Scale	0.7294	0.7155	0.7435	.

TABLE 9.17 – Estimation des paramètres du modèle de régression gamma pour les coûts.

L'équation de vraisemblance donnant l'estimateur $\widehat{\boldsymbol{\beta}}$ *s'écrit*

$$\sum_{i \ t.q. \ n_i>0} \sum_{k=1}^{n_i} \Big(\ln c_{ik} - \boldsymbol{\beta}^t \boldsymbol{x}_i \Big) \boldsymbol{x}_i = \mathbf{0} \Leftrightarrow \sum_{i \ t.q. \ n_i>0} lcres_i \boldsymbol{x}_i = \mathbf{0}$$

où $lcres_i$ *est le résidu d'estimation défini pour les assurés ayant déclaré au moins un sinistre par*

$$lcres_i = \sum_{k=1}^{n_i} \Big\{ \ln c_{ik} - \boldsymbol{\beta}^t \boldsymbol{x}_i \Big\} = \sum_{k=1}^{n_i} \ln c_{ik} - n_i \boldsymbol{\beta}^t \boldsymbol{x}_i.$$

Cette équation exprime donc également une relation d'orthogonalité entre variables explicatives et résidus d'estimation.

Cette manière de procéder n'est pas sans inconvénient : puisque la variable dépendante est le logarithme des montants des sinistres, les conclusions inférentielles se rapporteront aux montants des sinistres ainsi transformés, et il n'est pas toujours aisé de repasser aux données initiales.

Remarque 9.10.7 (Autre application de la régression Gamma : un modèle de provisionnement individuel)**.** *Une application fort intéressante de l'analyse des coûts des sinistres consiste à déterminer des règles de provisionnement individuelles. Dès qu'un sinistre est déclaré, la compagnie doit en effet mettre en réserve un montant correspondant au coût probable de cet événement, afin de donner à ses actionnaires, au marché et aux autorités de contrôle une image fidèle de sa situation financière.*

Une approche automatique consiste à expliquer le montant à provisionner en fonction des caractéristiques des sinistres déclarés dans le passé (outre les variables a priori, l'assureur utilisera également des informations relatives aux circonstances des sinistres, comme la période de la journée où il s'est produit, la présence de dégâts coporels, ...).

9.11 Tarification sur données de panel

Cette section est basée sur Denuit, Pitrebois & Walhin (2003), et n'aborde que les nombres des sinistres.

9.11.1 Tarification sur base de données en panel

Souvent, les actuaires utilisent plusieurs années d'observation afin de construire leur tarif (dans le but d'augmenter la taille de la base de données, mais aussi pour éviter d'accorder trop d'importance à des événements relatifs à une année particulière). Ceci a notamment pour conséquence que certaines des données ne seront plus indépendantes. En effet, les observations réalisées sur un même assuré au cours des différentes périodes considérées sont sans doute corrélées (ce qui est la raison d'être de la tarification *a posteriori* qui sera examinée dans les chapitres suivants). Nous sommes donc en présence de données en panel.

Dans le cadre de la tarification *a priori*, la dépendance existant entre les observations relatives à la même police est considérée comme une nuisance : l'actuaire veut à ce stade déterminer l'impact des facteurs observables sur le risque assuré, et les corrélations existant entre les données l'empêchent de recourir aux techniques statistiques classiques (pour la plupart fondées sur l'hypothèse d'indépendance). Nous montrerons ici comment prendre cette dépendance en compte afin d'améliorer la qualité des estimations à l'aide des techniques proposées par Liang & Zeger (1986) et Zeger et al. (1988).

Les estimateurs des fréquences de sinistres obtenus sous l'hypothèse d'indépendance des données individuelles relatives à différentes périodes sont convergents (c'est-à-dire qu'ils tendront en probabilité vers les valeurs population si la taille de l'échantillon croît). Dès lors, on peut raisonnablement espérer que pour des portefeuilles automobiles de grande taille, l'impact de l'hypothèse simplificatrice d'indépendance sur les estimations ponctuelles soit minime. C'est en effet ce que nous mettrons en évidence dans la partie empirique de notre étude.

9.11.2 Notations

Comme nous l'avons expliqué plus haut, les compagnies d'assurance utilisent souvent plusieurs périodes d'observation pour construire leur tarif. Les observations individuelles sont donc doublement indicées, par la police i et la période t. Dorénavant, N_{it}

représente le nombre de sinistres déclarés par l'assuré i durant la période t, $i = 1, 2, \ldots, n$, $t = 1, 2, \ldots, T_i$, où T_i désigne le nombre de périodes d'observation pour l'assuré i. Nous noterons d_{it} la durée de la $t$ème période d'observation pour l'individu i. Lors de chaque modification des variables observables, un nouvel intervalle commence, de sorte que d_{it} peut être différent de 1. Nous supposons que nous disposons par ailleurs d'autres variables $\boldsymbol{x}_{it}$, connues au début de la période t, et pouvant servir de facteurs explicatifs pour la sinistralité de l'assuré i. En plus des variables explicatives, on peut introduire le temps calendaire en composante de régression afin de prendre en compte certains événements ponctuels ou d'éventuelles tendances dans la sinistralité, dans l'esprit de Besson & Partrat (1992).

Typiquement, nous sommes en présence de données de panel : une même variable est mesurée sur un grand nombre n d'individus au cours du temps, à un nombre $\max_{1\leq i\leq n} T_i$ relativement faible de reprises. L'asymptotique se fera ici en faisant tendre n vers l'infini, et non pas le nombre d'observations effectuées sur un même individu (comme c'est typiquement le cas dans le cadre de l'analyse des séries chronologiques).

9.11.3 Présentation du jeu de données

Nous illustrons nos propos sur un portefeuille d'assurance belge comprenant 20 354 polices, observées durant une période de 3 ans. La Figure **??** donne une idée de la durée d'exposition au risque des polices en portefeuille. Un peu plus de 34% des assurés sont restés en portefeuille durant les trois ans. Pour chaque police et pour chaque année sont renseignés le nombre de sinistres et certaines caractéristiques de l'assuré : le sexe du conducteur (homme-femme), l'âge du conducteur (trois classes d'âge : $18-22$ ans, $23-30$ ans et > 30 ans), la puissance du véhicule (trois classes de puissance : < 66kW, $66-110$kW et > 110kW), la taille de la ville de résidence du conducteur (grande, moyenne ou petite, en fonction du nombre d'habitants) et la couleur du véhicule (rouge ou autre). Sur l'ensemble du portefeuille la fréquence annuelle moyenne est de 18.4% (ce qui est largement supérieur à la moyenne européenne).

Les Figures **??** à **??** montrent des histogrammes décrivant, pour chaque variable explicative, la répartition du portefeuille entre les différents niveaux de la variable et, pour chacun de ces niveaux, la fréquence moyenne (en %) de sinistres.

Ces histogrammes appellent les quelques commentaires suivants. On constate à la Figure **??** une légère sous-sinistralité pour les

FIGURE 9.29 – Durée d'exposition au risque (en mois).

FIGURE 9.30 – Répartition et fréquence de sinistres selon le sexe du conducteur.

FIGURE 9.31 – Répartition et fréquence de sinistres selon l'âge du conducteur.

FIGURE 9.32 – Répartition et fréquence de sinistres selon la puissance du véhicule.

FIGURE 9.33 – Répartition et fréquence de sinistres selon la taille de l'agglomération où réside l'assuré.

FIGURE 9.34 – Répartition et fréquence de sinistres selon la couleur du véhicule.

femmes (17.7% contre 18.8%), qui représentent 36% des assurés du portefeuille. La sur-sinistralité des jeunes conducteurs ressort clairement de la Figure **??** (mais ils sont sous-représentés dans le portefeuille). Les fréquences de sinistres semblent décroître avec l'âge, passant de 30.8% à 20.8% et enfin à 16.3%. En ce qui concerne la puissance du véhicule, on constate à la Figure **??** une sous-sinistralité pour les grosses cylindrées. L'examen de la Figure **??** révèle que la fréquence des sinistres est plus élevée dans les grandes agglomérations. La sinistralité semble décroître avec la taille de l'agglomération. Enfin, on constate à la Figure **??** que la couleur rouge du véhicule ne semble pas être un facteur aggravant.

9.11.4 Régression de Poisson en supposant l'indépendance temporelle

Modèle

En première approximation, on supposera les N_{it} indépendantes pour différentes valeurs de i et de t. Il s'agit bien entendu d'une hypothèse simplificatrice forte dont nous évaluerons l'impact en comparant les résultats obtenus à ceux fournis par différentes méthodes permettant de tenir compte de la dépendance sérielle existant entre les N_{it} à i fixé.

Nous supposons que la loi conditionnelle de N_{it} sachant $\boldsymbol{x}_{it}$ est de Poisson et nous spécifions une moyenne de forme exponentielle linéaire, i.e.

$$N_{it} \sim \mathcal{P}oi\big(d_{it}\exp(\eta_{it})\big),\ \ i=1,2,\ldots,n,\ \ t=1,2,\ldots,T_i. \quad (9.52)$$

La fréquence de sinistre relative à l'individu i durant la période t est $\lambda_{it} = d_{it}\exp(\eta_{it})$.

Estimation des paramètres

Notons n_{it} le nombre de sinistres déclarés par l'assuré i durant la période t. La vraisemblance associée à ces observations vaut alors

$$\mathcal{L}(\boldsymbol{\beta}|\boldsymbol{n}) = \prod_{i=1}^{n}\prod_{t=1}^{T_i}\exp\{-\lambda_{it}\}\frac{\{\lambda_{it}\}^{n_{it}}}{n_{it}!};$$

il s'agit de la probabilité d'obtenir les observations réalisées au sein du portefeuille dans le modèle considéré (notez que $\mathcal{L}$ est une fonction des paramètres $\boldsymbol{\beta}$, les observations étant supposées connues).

L'estimation de $\boldsymbol{\beta}$ par la méthode du maximum de vraisemblance consiste à déterminer $\widehat{\boldsymbol{\beta}}$ en maximisant $\mathcal{L}(\boldsymbol{\beta}|\boldsymbol{n})$: $\widehat{\boldsymbol{\beta}}$ est donc la valeur du paramètre rendant les observations recueillies par l'actuaire les plus probables. Afin de faciliter l'obtention du maximum, on passe souvent à la log-vraisemblance, laquelle est donnée par

$$L(\boldsymbol{\beta}|\boldsymbol{n}) = \ln\mathcal{L}(\boldsymbol{\beta}|\boldsymbol{n}) = \sum_{i=1}^{n}\sum_{t=1}^{T_i}\Big\{-\ln n_{it}! + n_{it}(\eta_{it}+\ln d_{it}) - \lambda_{it}\Big\}.$$

Dès lors, $\widehat{\boldsymbol{\beta}}$ est solution du système

$$\frac{\partial}{\partial\beta_0}L(\boldsymbol{\beta}|\boldsymbol{n}) = 0 \Leftrightarrow \sum_{i=1}^{n}\sum_{t=1}^{T_i}n_{it} = \sum_{i=1}^{n}\sum_{t=1}^{T_i}\lambda_{it} \quad (9.53)$$

et pour $j = 1, 2, \ldots, p$,

$$\frac{\partial}{\partial \beta_j} L(\boldsymbol{\beta}|\boldsymbol{n}) = 0 \quad \Leftrightarrow \quad \sum_{i=1}^{n} \sum_{t=1}^{T_i} x_{itj} \Big\{ n_{it} - \lambda_{it} \Big\} = 0. \qquad (9.54)$$

Sans surprise, on peut interpréter les équation de vraisemblance (**??**) comme une relation d'orthogonalité entre les variables explicatives $\boldsymbol{x}_{it}$ et les résidus d'estimation.

La matrice variance-covariance $\boldsymbol{\Sigma}$ de l'estimateur du maximum de vraisemblance $\widehat{\boldsymbol{\beta}}$ du paramètre $\boldsymbol{\beta}$ est l'inverse de la matrice d'information de Fisher $\mathcal{I}$. Elle peut être estimée par

$$\widehat{\boldsymbol{\Sigma}} = \left\{ \sum_{i=1}^{n} \sum_{t=1}^{T_i} \boldsymbol{x}_{it} \boldsymbol{x}_{it}^t \widehat{\lambda_{it}} \right\}^{-1}.$$

En vertu de la théorie asymptotique de la méthode du maximum de vraisemblance, $\widehat{\boldsymbol{\beta}}$ est approximativement de loi normale de moyenne la vraie valeur du paramètre et de matrice variance-covariance $\widehat{\boldsymbol{\Sigma}}$. Ceci permet d'obtenir des intervalles et des zones de confiance pour les paramètres.

Illustration numérique

La procédure `GENMOD` de SAS permet de réaliser la régression de Poisson du nombre de sinistres sur les 5 variables explicatives présentées à la Section 1.7. La variable explicative "couleur du véhicule" n'est pas significative. Après l'avoir éliminée du modèle, nous regroupons les niveaux de puissance "66 − 110kW" et "> 110kW" en une seule classe. Nous en arrivons alors au modèle retenu, lequel est décrit au Tableau **??**.

[hp]

Variable	Level	Coeff β	Std Error	Wald 95% Conf Limit		Chi-Sq	Pr>ChiSq
Intercept		−1.9277	0.0299	−1.9862	−1.8692	4165.69	< .0001
Sexe	Femme	−0.0575	0.0265	−0.1093	−0.0056	4.72	0.0299
Sexe	Homme	0	0	0	0	.	.
Age	17 − 22	0.6668	0.0582	0.5526	0.7809	131.02	< .0001
Age	23 − 30	0.2547	0.0260	0.2038	0.3056	96.09	< .0001
Age	> 30	0	0	0	0	.	.
Puissance	> 66kW	0.0508	0.0269	−0.0019	0.1034	3.57	0.0587
Puissance	< 66kW	0	0	0	0	.	.
Ville	Grande	0.2545	0.0306	0.1944	0.3145	69.03	< .0001
Ville	Moyenne	0.0757	0.0311	0.0148	0.1365	5.93	0.0148
Ville	Petite	0	0	0	0	.	.

Résultats de la régression de Poisson pour le modèle final en supposant l'indépendance sérielle.

La log-vraisemblance vaut -19 283.2 et l'analyse de Type 3 fournit les résultats présentés au Tableau **??**. A l'exception de la variable

puissance, toutes les variables sont statistiquement significatives et l'omission d'une quelconque d'entre elles détériore significativement le modèle (au seuil de 5%). Nous décidons cependant de garder la variable puissance en raison de son importance dans les tarifs pratiqués par les compagnies d'assurances et du faible dépassement du seuil (0.93%, seulement). La log-vraisemblance du modèle final est à peine moins bonne que celle du modèle non contraint (à savoir, -19 282.6).

Source	DF	ChiSquare	Pr > ChiSq
Sexe	1	4.74	0.0294
Age	2	176.07	< .0001
Puissance	1	3.56	0.0593
Ville	2	73.82	< .0001

TABLE 9.18 – Résultats de l'analyse de Type 3 pour le modèle final en supposant l'indépendance sérielle.

Analyse des résidus

La Figure **??** décrit les résidus de déviance individuels. On peut y constater la structure reflétant les quelques valeurs observées pour les N_{it}. On ne peut donc juger de la qualité du modèle sur base de la Figure **??**. Si on recalcule les résidus par classes, on obtient la Figure **??**. On n'y constate aucune structure particulière, mais des valeurs assez élevées de certains résidus, qui mettent en question la justesse du modèle.

FIGURE 9.35 – Graphe des résidus individuels en fonction des valeurs prédites

FIGURE 9.36 – Graphe des résidus par classe en fonction des valeurs prédites en supposant l'indépendance sérielle.

Surdispersion

Afin de détecter une éventuelle surdispersion, on calcule pour chaque classe de risque la moyenne et la variance empirique des nombres des sinistres, $\widehat{m}_k$ et $\widehat{\sigma}_k^2$, disons, et on porte le nuage de points

$\{(\widehat{m}_k, \widehat{\sigma}_k^2),\ k = 1, 2, \ldots\}$ en graphique. Le résultat est visible à la Figure **??**. On y constate une forte surdispersion et ce pour toutes les catégories d'assurés. Les points $(\widehat{m}_k, \widehat{\sigma}_k^2)$ sont en effet situés au-dessus de la première bissectrice du quadrillage. Ceci nous conduit également à considérer que le modèle de Poisson avec indépendance temporelle n'est pas adapté.

FIGURE 9.37 – Vérification de la validité de (**??**) sur les données

Remarque 9.11.1. *Il est possible de tenir compte de la surdispersion constatée dans les données, sans reconnaître l'éventuelle dépendance sérielle. A cette fin, on recourt soit à un modèle de Poisson mélange, soit à une approche de quasi-vraisemblance en spécifiant*

$$\mathbb{V}[N_{it}|\boldsymbol{x}_{it}] = \phi\mathbb{E}[N_{it}|\boldsymbol{x}_{it}] = \phi\lambda_{it}. \tag{9.55}$$

Afin d'éprouver graphiquement la validité de cette dernière relation, nous avons ajusté le nuage de points de la Figure **??** *à l'aide d'une droite passant par l'origine (donc d'équation $y = \phi x$). Ceci donne un paramètre de dispersion ϕ estimé à 1.9122 et un coefficient de détermination $R^2 = 86.17\%$ (ce qui signifie que la droite explique plus de 86% de la variabilité du nuage de points). A titre de comparaison, si nous avions tenté un ajustement à l'aide d'une courbe du second degré (du type $y = x + \gamma x^2$, caractéristique du lien moyenne-variance dans un modèle de Poisson mélange), on aurait obtenu $y = x + 2.9545x^2$ avec $R^2 = 90.90\%$. Un mélange de Poisson (la loi binomiale négative, par exemple) aurait donc pu également être considéré. Nous privilégions cependant dans cette section une approche de quasi-vraisemblance. Cela consiste à déterminer $\widehat{\boldsymbol{\beta}}$ en résolvant le système* (**??**)-(**??**)*. Ensuite, $\widehat{\phi}$ est obtenu en divisant soit la déviance, soit la statistique de Pearson par le nombre de degrés de liberté. La valeur estimée de ϕ sur nos données est 1.35, ce qui traduit bien la surdispersion des données.*

L'introduction du paramètre de surdispersion ϕ gonfle les variances et les covariances des $\widehat{\beta}_j$ (lesquelles sont multipliées par $\widehat{\phi}$). Ceci a pour effet de réduire la valeur des statistiques de test utilisées pour éprouver la nullité des β_j ou la pertinence de l'inclusion de certaines variables dans le modèle. La prise en compte de la surdispersion peut donc mener à l'exclusion de variables tarifaires qui auraient été conservées dans le modèle de Poisson pur. On observe

un phénomène de ce type sur notre jeu de données, la p-valeur de la variable puissance dans l'analyse de type 3 passant à 10.44%.

9.11.5 Prise en compte de la dépendance temporelle

Détection de l'aspect sériel

Afin d'avoir une première idée du type de dépendance existant entre les N_{it}, on peut par exemple considérer les observations N_{it}, $t = 2, \ldots, T_i$, $i = 1, \ldots, n$, et effectuer une régression de celles-ci sur les variables explicatives $\boldsymbol{x}_{it}$ correspondantes ainsi que le nombre $N_{i,t-1}$ de sinistres observés au cours de la période de couverture précédente. Ceci permettra également de voir l'effet de l'inclusion de valeurs passées de la variable d'intérêt sur les variables explicatives.

Afin de mettre cette dépendance en évidence, nous travaillons avec les observations des deux dernières années à notre disposition. Nous considérons donc les observations N_{it}, $t = 2, 3$, $i = 1, \ldots, n$, et nous effectuons une régression de celles-ci sur les variables explicatives $\boldsymbol{x}_{it}$ correspondantes auxquelles nous ajoutons la variable $N_{i,t-1}$, i.e. le nombre de sinistres observés au cours de la période précédente. Nous partons d'un modèle contenant les 5 variables explicatives déjà présentées et nous l'affinons par étapes successives, grâce à l'analyse de Type 3. Nous commençons par éliminer la variable "couleur du véhicule" qui a une p-valeur de de 27.37% et dans une deuxième étape nous éliminons la variable "sexe du conducteur" dont la p-valeur est devenue 21.10%. Nous obtenons alors le modèle dont les résultats sont présentés dans les Tableaux **??** et **??**. Le coefficient de régression obtenu pour le nombre passé de sinistres est hautement significatif, ce qui indique une dépendance sérielle.

[hp]

Variable	Level	Coeff β	Std Error	Wald 95% Conf Limit		Chi-Sq	Pr>ChiSq
Intercept		−2.0405	0.0370	−2.1131	−1.9680	3041.80	< .0001
Age	17 − 22	0.5841	0.0983	0.3914	0.7767	35.31	< .0001
Age	23 − 30	0.1822	0.0348	0.1140	0.2503	27.41	< .0001
Age	> 30	0	0	0	0	.	.
Puissance	> 110kW	−0.0745	0.1035	−2.2773	0.1283	0.52	0.4716
Puissance	66 − 110kW	0.0933	0.0357	0.0233	0.1633	6.83	0.0090
Puissance	< 66kW	0	0	0	0	.	.
Ville	Grande	0.2201	0.0412	0.1394	0.3009	28.54	< .0001
Ville	Moyenne	0.1050	0.0413	0.0242	0.1859	6.48	0.0109
Ville	Petite	0	0	0	0	.	.
N_{t-1}		0.3113	0.0371	0.2387	0.3839	70.59	< .0001

Résultats de la régression pour le modèle tenant compte de la sinistralité passée.

Dans une deuxième approche nous repartons de la fréquence obtenue sous l'hypothèse d'indépendance et sans ajout du nombre de sinistres de l'année précédente comme variable explicative. Cette

Source	DF	ChiSquare	Pr > ChiSq
Age	2	50.58	< .0001
Puissance	2	7.94	0.0188
Ville	2	28.68	< .0001
N_{t-1}	1	63.38	< .0001

TABLE 9.19 – Résultats de l'analyse de Type 3 pour le modèle tenant compte de la sinistralité passée.

prime est alors corrigée par un facteur multiplicatif, obtenu par une régression de Poisson sur la seule variable "nombre de sinistres de l'année précédente" (en mettant la prime fréquence obtenue sous l'hypothèse d'indépendance en offset [5]). Les résultats de cette régression se trouvent dans les Tableaux **??** et **??**.

Variable	Coeff β	Std Error	Wald 95% Conf Limit		Chi-Sq	Pr>ChiSq
Intercept	−0.1147	0.0180	−0.1500	−0.0793	40.42	< .0001
N_{t-1}	0.3040	0.0370	0.2316	0.3765	67.65	< .0001

TABLE 9.20 – Résultats de la régression pour le modèle tenant compte de la sinistralité passée en figeant l'influence des variables explicatives.

Source	DF	ChiSquare	Pr > ChiSq
N_{t-1}	1	60.84	< .0001

TABLE 9.21 – Résultats de l'analyse de Type 3 pour le modèle tenant compte de la sinistralité passée en figeant l'influence des variables explicatives.

Remarque 9.11.2. *Il est intéressant de noter au passage que cette manière de procéder fournit immédiatement des coefficients bonus-malus "à la française". En effet, le Tableau* **??** *nous apprend que les assurés qui n'ont déclaré aucun sinistre sur l'année verront leur prime multipliée par* $\exp(-0.1147) = 0.8916$ *alors que ceux ayant déclaré* k *sinistres subiront une majoration de prime valant* $\exp(-0.1147 + k \times 0.3040) = 0.8916 \times (1.3553)^k$. *Il est toujours intéressant de comparer ces coefficients à ceux produits par un modèle plus orthodoxe formulé en termes de variables latentes.*

Les données suggèrent donc une dépendance sérielle. Cela invalide les résultats obtenus à la section précédente, lesquels se fondent

5. Le terme offset désigne une composante du score qui n'est pas multipliée par un coefficient de régression à estimer.

notamment sur l'hypothèse que les N_{it} sont indépendantes pour différentes valeurs de i et de t. Théoriquement, on peut cependant montrer que l'estimateur du maximum de vraisemblance $\widehat{\boldsymbol{\beta}}$ calculé sous l'hypothèse d'indépendance sérielle (donc avec erreur de spécification) est convergent. Si la taille du portefeuille est suffisamment grande, on s'attend donc à peu d'impact sur les estimations ponctuelles des différents β_j. Par contre, la variance de $\widehat{\boldsymbol{\beta}}$ ne peut plus être calculée comme décrit plus haut, et est quant à elle affectée par la dépendance sérielle.

Estimation des paramètres à l'aide de la technique GEE

En présence de dépendance sérielle, on pourrait songer à garder l'estimateur du maximum de vraisemblance dans le modèle de Poisson avec indépendance temporelle (donc solution de (**??**)-(**??**)), choix qui se justifie par le caractère convergent de celui-ci. Comme l'ont montré Liang & Zeger (1986), il est possible d'améliorer cette approche (i.e. d'obtenir des estimateurs dont la variance asymptotique sera plus faible que celle de ceux que nous venons de décrire). Il s'agit de la méthode des GEE (pour l'anglais "Generalized Estimating Equation") proposée par Liang & Zeger (1986). Les estimateurs fournis par cette méthode sont convergents ; on espère donc que les estimations ainsi obtenues seront de bonne qualité vu le grand nombre d'observations dont dispose en général l'actuaire.

L'idée est simple : retenir l'estimateur du maximum de vraisemblance $\widehat{\boldsymbol{\beta}}$ solution de (**??**)-(**??**) pour estimer $\boldsymbol{\beta}$ dans le modèle avec effet aléatoire n'est certainement pas optimal puisqu'on ne tient pas compte de la structure de corrélation des N_{it}. Réécrivons le système (**??**)-(**??**) sous forme vectorielle :

$$\sum_{i=1}^{n} \boldsymbol{X}_i^t(\boldsymbol{n}_i - \mathbb{E}[\boldsymbol{N}_i]) = \boldsymbol{0} \text{ où } \boldsymbol{X}_i = (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iT_i})^t. \tag{9.56}$$

La matrice de covariance des N_{it} dans le modèle de Poisson avec indépendance sérielle est

$$\boldsymbol{A}_i = \begin{pmatrix} \lambda_{i1} & 0 & \cdots & 0 \\ 0 & \lambda_{i2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_{iT_i} \end{pmatrix}.$$

Cette matrice ne rend donc compte ni de la surdispersion, ni de la dépendance sérielle présente dans les données. Si on fait apparaître

explicitement la matrice $\boldsymbol{A}_i$ dans (**??**), on obtient

$$\sum_{i=1}^{n} \left(\frac{\partial}{\partial \boldsymbol{\beta}} \mathbb{E}[\boldsymbol{N}_i] \right)^t \boldsymbol{A}_i^{-1} (\boldsymbol{n}_i - \mathbb{E}[\boldsymbol{N}_i]) = \boldsymbol{0} \tag{9.57}$$

puisque

$$\frac{\partial}{\partial \boldsymbol{\beta}} \mathbb{E}[\boldsymbol{N}_i] = \boldsymbol{A}_i \boldsymbol{X}_i.$$

Le principe des GEE consiste à substituer à $\boldsymbol{A}_i$ dans (**??**) un candidat plus raisonnable pour la matrice variance-covariance de $\boldsymbol{N}_i$, plus raisonnable signifiant ici rendant compte de la surdispersion et de la corrélation temporelle. Spécifions à présent une forme plausible pour la matrice de covariance de $\boldsymbol{N}_i$: on pourrait penser à

$$\boldsymbol{V}_i = \phi \boldsymbol{A}_i^{1/2} \boldsymbol{R}_i(\boldsymbol{\alpha}) \boldsymbol{A}_i^{1/2}$$

où la matrice de corrélation $\boldsymbol{R}_i(\boldsymbol{\alpha})$ rend compte de la dépendance sérielle existant entre les composantes de $\boldsymbol{N}_i$ et dépend d'un certain nombre de paramètres $\boldsymbol{\alpha}$. La matrice $\boldsymbol{R}_i$ est une sous-matrice carrée de dimension $T_i \times T_i$ d'une matrice $\boldsymbol{R}$ de dimension $T_{\max} \times T_{\max}$ dont les éléments ne dépendent pas des caractéristiques $\boldsymbol{x}_{it}$ de l'individu i. La surdispersion est quant à elle prise en compte puisque $\mathbb{V}[N_{it}] = \phi \lambda_{it}$. Notez que la matrice $\boldsymbol{V}_i$ ainsi définie n'est la matrice de covariance de $\boldsymbol{N}_i$ que si $\boldsymbol{R}_i$ est la matrice de corrélation de $\boldsymbol{N}_i$, ce qui n'est pas nécessairement le cas.

Comme annoncé ci-dessus, l'idée consiste alors à substituer la matrice $\boldsymbol{V}_i$ à $\boldsymbol{A}_i$ dans (**??**), et de retenir comme estimation de $\boldsymbol{\beta}$ la solution de

$$\sum_{i=1}^{n} \left(\frac{\partial}{\partial \boldsymbol{\beta}} \mathbb{E}[\boldsymbol{N}_i] \right)^t \boldsymbol{V}_i^{-1} (\boldsymbol{n}_i - \mathbb{E}[\boldsymbol{N}_i]) = \boldsymbol{0} \tag{9.58}$$

Cette dernière relation exprime également une orthogonalité entre les résidus de régression et les variables explicatives. Les estimateurs ainsi obtenus sont convergents quel que soit le choix de la matrice $\boldsymbol{R}_i(\boldsymbol{\alpha})$. On sent évidemment bien qu'ils seront d'autant plus précis que $\boldsymbol{R}_i(\boldsymbol{\alpha})$ est proche de la véritable matrice de corrélation de $\boldsymbol{N}_i$.

Modélisation de la dépendance à l'aide de la "working correlation matrix"

Comme nous l'avons compris à la lecture de ce qui précède, c'est la matrice de corrélation $\boldsymbol{R}_i$ qui tient compte de la dépendance entre

les observations relatives à un même assuré. Cette matrice de dimension $T_i \times T_i$ est appelée "working correlation matrix". Il s'agit d'une matrice de corrélation de forme spécifiée dépendant d'un certain nombre de paramètres repris dans le vecteur $\boldsymbol{\alpha}$.

Si $\boldsymbol{R}_i(\boldsymbol{\alpha})$ =Identité, (**??**) donne exactement les équations de vraisemblance (**??**) sous l'hypothèse d'indépendance.

En général, on spécifie dans le cadre de la tarification a priori une matrice $\boldsymbol{R}_i(\boldsymbol{\alpha})$ traduisant une structure de type autorégressive. Ainsi, les éléments diagonaux de $\boldsymbol{R}_i$ valent 1 et hors diagonale, l'élément jk vaut $\alpha_{|j-k|}$ pour $|j-k| \leq m$ et 0 pour $|j-k| > m$. On prendra $m = T_{\max} - 1$. Les composantes du vecteur $\boldsymbol{\alpha}$ paramétrant la matrice $\boldsymbol{R}_i(\alpha)$ décrivant le type de dépendance entre les données sont à estimer sur base des observations.

Obtention des estimations

L'équation (**??**) est généralement résolue à l'aide d'une méthode du score de Fisher modifiée pour $\boldsymbol{\beta}$ et une estimation des moments pour $\boldsymbol{\alpha}$ (nous renvoyons le lecteur à Liang & Zeger (1986) pour une description complète de la méthode). Spécifiquement, partant d'une valeur initiale $\widehat{\boldsymbol{\beta}}^{(0)}$ solution du système (**??**)-(**??**), nous calculons

$$\begin{aligned}\widehat{\boldsymbol{\beta}}^{(j+1)} &= \widehat{\boldsymbol{\beta}}^{(j)} + \left\{\sum_{i=1}^{n} \boldsymbol{D}_i^t(\widehat{\boldsymbol{\beta}}^{(j)})\boldsymbol{V}_i^{-1}(\widehat{\boldsymbol{\beta}}^{(j)}, \boldsymbol{\alpha}(\widehat{\boldsymbol{\beta}}^{(j)}))\boldsymbol{D}_i(\widehat{\boldsymbol{\beta}}^{(j)})\right\}^{-1} \\ &\qquad \left\{\sum_{i=1}^{n} \boldsymbol{D}_i^t(\widehat{\boldsymbol{\beta}}^{(j)})\boldsymbol{V}_i^{-1}(\widehat{\boldsymbol{\beta}}^{(j)}, \boldsymbol{\alpha}(\widehat{\boldsymbol{\beta}}^{(j)}))\boldsymbol{S}_i(\widehat{\boldsymbol{\beta}}^{(j)})\right\}\end{aligned}$$

où $\boldsymbol{D}_i(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}}\mathbb{E}[\boldsymbol{N}_i]$ et $\boldsymbol{S}_i(\boldsymbol{\beta}) = \boldsymbol{N}_i - \mathbb{E}[\boldsymbol{N}_i]$. A chaque étape, ϕ et $\boldsymbol{\alpha}$ sont réestimés à partir des résidus de Pearson $r_{it}^P = \frac{n_{it}-\lambda_{it}}{\sqrt{\lambda_{it}}}$ grâce aux formules

$$\widehat{\phi} = \frac{1}{\sum_{i=1}^{n} T_i - p} \sum_{i=1}^{n}\sum_{t=1}^{T_i} \{r_{it}^P\}^2$$

et

$$\widehat{\alpha}_\tau = \frac{1}{\widehat{\phi}\left(\sum_{i|T_i>\tau}(T_i-\tau) - p\right)} \sum_{i|T_i>\tau}\sum_{t=1}^{T_i-\tau} r_{it}^P r_{it+\tau}^P.$$

Illustration numérique

La dépendance sérielle des N_{it} à i fixé ayant clairement été mise en évidence à la Section 3.1, il importe de mesurer l'impact de

l'hypothèse d'indépendance sur l'estimation des fréquences. L'approche GEE peut être réalisée par la procédure `GENMOD` de SAS. Une sélection des variables, basée comme précédemment sur l'analyse de Type 3, nous conduit à retenir les mêmes variables que pour le modèle où l'on supposait l'indépendance. Les résultats se trouvent dans les Tableaux **??** et **??**. L'estimation de la "working correlation matrix" de structure autorégressive d'ordre 2 (i.e. d'ordre $T_{\max}-1$) donne

$$\begin{pmatrix} 1 & 0.0493 & 0.0462 \\ 0.0493 & 1 & 0.0493 \\ 0.0462 & 0.0493 & 1 \end{pmatrix}$$

et $\widehat{\phi} = 1.3437$.

[hp]

Variable	Level	Coeff β	Std Error	95% Conf Limit		Z	Pr> $\lvert Z \rvert$
Intercept		−1.9233	0.0319	−1.9858	−1.8608	−60.32	< .0001
Sexe	Femme	−0.0581	0.0289	−0.1148	−0.0014	−2.01	0.0446
Sexe	Homme	0	0	0	0	.	.
Age	17 − 22	0.6586	0.0617	0.5376	0.7797	10.67	< .0001
Age	23 − 30	0.2557	0.0281	0.2006	0.3107	9.10	< .0001
Age	> 30	0	0	0	0	.	.
Puissance	> 66kW	0.0532	0.0292	−0.0041	0.1105	1.82	0.0686
Puissance	< 66kW	0	0	0	0	.	.
Ville	Grande	0.2542	0.0332	0.1892	0.3192	7.67	< .0001
Ville	Moyenne	0.0719	0.0336	0.0060	0.1379	2.14	0.0326
Ville	Petite	0	0	0	0	.	.

Résultats de la régression de Poisson avec approche GEE et structure de dépendance AR(2).

Source	DF	ChiSquare	Pr > ChiSq
Sexe	1	4.09	0.0431
Age	2	128.79	< .0001
Puissance	1	3.28	0.0701
Ville	2	60.71	< .0001

TABLE 9.22 – Résultats de l'analyse de Type 3 pour le modèle avec approche GEE et structure de dépendance AR(2).

Si on compare les $\widehat{\beta}_j$ des Tableaux **??** (sous l'hypothèse d'indépendance) et **??** (reconnaissant la dépendance sérielle), on constate des différences modestes. Les erreurs-standards sont systématiquement plus élevées dans l'approche GEE (la dépendance sérielle augmentant la surdispersion).

Impact sur les fréquences

Pour terminer, comparons les fréquences obtenues en supposant l'indépendance sérielle ou en reconnaissant explicitement la

Classes de risque				Fréquences		
Sexe	Age	Puissance	Ville	Inf	Prime	Sup
Homme	17-22	< 66	Petite	0.25251	0.28339	0.31084
			Moyenne	0.27221	0.30566	0.34322
			Grande	0.32535	0.3655	0.41061
		≥ 66	Petite	0.26395	0.29815	0.33678
			Moyenne	0.28464	0.32158	0.36332
			Grande	0.34027	0.38454	0.43457
	23-30	< 66	Petite	0.17708	0.18768	0.19892
			Moyenne	0.19135	0.20243	0.21416
			Grande	0.22878	0.24206	0.25612
		≥ 66	Petite	0.1848	0.19746	0.21099
			Moyenne	0.19976	0.21298	0.22707
			Grande	0.23893	0.25467	0.27146
	> 30	< 66	Petite	0.13721	0.14548	0.15425
			Moyenne	0.1483	0.15692	0.16604
			Grande	0.17754	0.18764	0.19831
		≥ 66	Petite	0.14413	0.15306	0.16254
			Moyenne	0.15588	0.16509	0.17485
			Grande	0.18668	0.19741	0.20876
Femme	17-22	< 66	Petite	0.23779	0.26756	0.30106
			Moyenne	0.25631	0.28859	0.32494
			Grande	0.30646	0.34509	0.38859
		≥ 66	Petite	0.24766	0.2815	0.31996
			Moyenne	0.26704	0.30362	0.34523
			Grande	0.31935	0.36307	0.41277
	23-30	< 66	Petite	0.16643	0.1772	0.18867
			Moyenne	0.17975	0.19113	0.20322
			Grande	0.21509	0.22855	0.24285
		≥ 66	Petite	0.17265	0.18643	0.20131
			Moyenne	0.18652	0.20108	0.21679
			Grande	0.22322	0.24045	0.25901
	> 30	< 66	Petite	0.12906	0.13736	0.14619
			Moyenne	0.13942	0.14816	0.15743
			Grande	0.16703	0.17716	0.1879
		≥ 66	Petite	0.13462	0.14451	0.15513
			Moyenne	0.14549	0.15587	0.1670
			Grande	0.17431	0.18639	0.1993

TABLE 9.23 – Estimations des fréquences des différentes classes de risque sous l'hypothèse d'indépendance.

dépendance temporelle ; celles-ci sont fournies aux Tableaux **??** et **??**. On constate des différences au niveau des estimations des fréquences annuelles de sinistre associées aux classes de risque, mais ces différences restent limitées (elles seront néanmoins exacerbées par la multiplication par le coût moyen d'un sinistre et par les chargements de sécurité et commerciaux). La prise en considération de la dépendance sérielle a également un impact sur les intervalles de confiance pour les fréquences, lesquels sont plus larges dans l'approche GEE.

9.12 Justifications techniques de la segmentation

9.12.1 Tarif technique et tarif commercial

Une des tâches fondamentales de l'actuaire consiste d'une part à évaluer la prime pure, c'est-à-dire la contribution de chacun per-

Classes de risque				Fréquences		
Sexe	Age	Puissance	Ville	Inf	Prime	Sup
Homme	17-22	< 66	Petite	0.24954	0.28233	0.31942
			Moyenne	0.26804	0.30348	0.34338
			Grande	0.32137	0.36405	0.41240
		≥ 66	Petite	0.26135	0.29776	0.33924
			Moyenne	0.28104	0.31996	0.36428
			Grande	0.33671	0.38395	0.43781
	23-30	< 66	Petite	0.17725	0.18869	0.20086
			Moyenne	0.19059	0.20276	0.21571
			Grande	0.22865	0.24331	0.25890
		≥ 66	Petite	0.18531	0.19899	0.21369
			Moyenne	0.19965	0.21384	0.22904
			Grande	0.23918	0.25660	0.27529
	> 30	< 66	Petite	0.13727	0.14612	0.15554
			Moyenne	0.14759	0.15701	0.16704
			Grande	0.17748	0.18842	0.20003
		≥ 66	Petite	0.14455	0.15410	0.16429
			Moyenne	0.15578	0.16560	0.17603
			Grande	0.18701	0.19871	0.21114
Femme	17-22	< 66	Petite	0.23532	0.26634	0.30156
			Moyenne	0.25294	0.28625	0.32396
			Grande	0.30359	0.34350	0.38865
		≥ 66	Petite	0.24518	0.28095	0.32193
			Moyenne	0.26381	0.30190	0.34550
			Grande	0.31640	0.36227	0.41479
	23-30	< 66	Petite	0.16643	0.17804	0.19045
			Moyenne	0.17918	0.19131	0.20427
			Grande	0.21541	0.22957	0.24466
		≥ 66	Petite	0.17254	0.18776	0.20427
			Moyenne	0.18606	0.20177	0.21880
			Grande	0.22333	0.24210	0.26248
	> 30	< 66	Petite	0.12867	0.13787	0.14773
			Moyenne	0.13851	0.14815	0.15847
			Grande	0.16687	0.17778	0.18940
		≥ 66	Petite	0.13426	0.14540	0.15747
			Moyenne	0.14477	0.15625	0.16864
			Grande	0.17409	0.18750	0.20193

TABLE 9.24 – Estimations des fréquences des différentes classes de risque dans l'approche GEE.

mettant à l'assureur de dédommager les sinistres, et d'autre part, à évaluer le chargement de sécurité garantissant la stabilité des résultats de la compagnie. Le concept de prime pure a fait l'objet du Chapitre 3 tandis que la détermination des chargements de sécurité a été étudiée dans le Chapitre 4. A cet égard, l'actuaire procédera à une analyse technique aussi fine que possible du risque à couvrir, et construira le tarif technique. Ce dernier donne donc le prix coûtant de la couverture octroyée par l'assureur, en fonction du profil de risque de l'assuré.

Le tarif technique est à usage purement interne. La compagnie appliquera aux assurés le tarif commercial, qui peut considérablement différer du tarif technique. Si le tarif commercial doit se baser sur le tarif technique, des considérations d'ordre réglementaire ou relatives au positionnement de la compagnie sur le marché, de même qu'un souci de simplification de la grille tarifaire peuvent mener à des primes commerciales parfois très éloignées du tarif technique. Deux grilles tarifaires co-existent donc : l'une purement technique et à usage exclusivement interne, fruit d'une analyse fine réalisée par l'actuaire, et l'autre commerciale, décrivant les montants qui seront effectivement réclamés aux assurés.

9.12.2 Segmentation des tarifs technique et commercial

Au niveau technique, l'actuaire peut soit segmenter finement le portefeuille, soit recourir à des modèles de mélange afin de tenir compte de l'hétérogénéité résultant de l'absence de segmentation. Intéressons-nous à présent au tarif commercial, où certaines imperfections du marché peuvent obliger l'actuaire à segmenter.

Au cours de la dernière décennie, les tarifs commerciaux autrefois quasi-uniformes ont été progressivement différenciés en fonction du profil des assurés. On parle souvent de segmentation pour désigner ce phénomène (voyez la Section 3.8). Même si la segmentation ne se limite pas à la différenciation tarifaire, mais recouvre aussi la sélection du risque à laquelle procède l'assureur lors de la conclusion du contrat (acceptation) ou en cours du contrat (résiliation).

Le but de cette section n'est ni de justifier, ni de critiquer le principe de segmentation au niveau commercial. En cette matière, la liberté contractuelle doit jouer et c'est à l'Etat de réguler le marché si nécessaire. Nous voulons seulement montrer que la plupart des choix tarifaires opérés par les compagnies résultent de considérations commerciales ou concurrentielles, et ne sont nullement imposés par la

science actuarielle. Seule l'antisélection émanant des preneurs d'assurance peut obliger un assureur à répercuter sur les assurés un critère de différenciation tarifaire. Ce critère se retrouve alors à la fois dans le tarif technique et dans le tarif commercial (mais peut-être pas de la même manière).

Le lecteur doit garder à l'esprit que la différenciation tarifaire n'a aucun impact sur la sinistralité. Donc, quelle que soit la politique de tarification de la compagnie, l'encaissement pur relatif au portefeuille doit être le même (puisqu'il correspond au coût attendu des sinistres). Dès lors, toute diminution de prime pour une catégorie d'assurés entraîne nécessairement une augmentation corollaire des primes payées par d'autres catégories d'assurés.

Une différenciation tarifaire trop poussée peut donc avoir des effets désastreux sur un marché d'assurance. En effet, une telle différenciation restreint considérablement le champ de l'assurabilité : en diminuant (souvent très modérément) le montant de la prime de certaines catégories d'assurés, on augmente celui supporté par d'autres, parfois au point de les exclure du marché de l'assurance. L'exemple de la RC automobile est édifiant à ce propos.

Le degré de différenciation des primes toujours plus poussé constaté sur le marché s'explique surtout par la spirale de la segmentation : c'est le jeu de la concurrence qui pousse les compagnies à différencier sans cesse davantage le montant des primes. En effet, un assureur peut difficilement maintenir un tarif uniforme sur un marché ou des concurrents différencient les risques. Théoriquement, dès qu'un acteur du marché introduit un nouveau critère de différenciation tarifaire, ses concurrents sont obligés de le reconnaître implicitement ou explicitement dans leurs tarifs comme nous l'avons expliqué en détails dans le Tome I. C'est pourquoi on assiste depuis plusieurs années à une différenciation tarifaire de plus en plus poussée, notamment en RC automobile. C'est donc la spirale de la segmentation induite par la concurrence effrénée que se livrent les acteurs du marché de l'assurance qui engendre la segmentation à outrance, et non un quelconque prescrit de technique actuarielle.

9.12.3 Antisélection et segmentation

Une caractéristique des assurances privées réside dans la liberté de contracter des deux parties concernées, preneur d'assurance et assureur. Le candidat-preneur choisira donc le contrat d'assurance qu'il considère le plus attrayant. En outre, il dispose d'un avantage certain en matière d'information par rapport à un assureur

qui segmente peu. Le candidat connaît en effet sa propre situation de manière très précise tandis que l'assureur concerné n'a aucune connaissance de certains facteurs d'aggravation du risque qui seraient présents. Cette information asymétrique mène à l'antisélection émanant des preneurs d'assurance : les assurés dont le profil est plus risqué souscrivent en masse aux couvertures offertes par la compagnie, détériorant la statistique sinistre de l'assureur.

L'antisélection émanant des preneurs d'assurance a pour effet de limiter l'accès à l'assurance. Pour s'en convaincre, considérons par exemple une couverture contre une maladie redoutée M. Si dans la population se côtoient des individus prédisposés à cette maladie et d'autres qui ne le sont pas, sans que les uns ni les autres ne connaissent leur éventuelle prédisposition, l'assureur pourra tous les couvrir sur base d'une prime moyenne. La police souscrite peut alors se voir comme un contrat multigaranties, qui couvre tout d'abord le risque d'être prédisposé à la maladie, et ensuite le coût du traitement de celle-ci lorsqu'elle se déclare.

Supposons par exemple que 10% de la population soit prédisposée à la maladie redoutée M. Plus précisément, en cas de prédisposition, la probabilité de développer la maladie M est de 2%, alors qu'elle n'est que de 0.1% en l'absence de prédisposition. Cette prédisposition peut être détectée à l'aide d'un test génétique, mais cet usage est interdit à l'assureur, et trop onéreux pour l'assuré. Le coût du traitement de la maladie M est de 10 000 € . La prime pure pour une couverture contre cette maladie est de

$$10\%\times 2\%\times 10\ 000\ € + 90\%\times 0.1\%\times 10\ 000\ € = 20\ € + 9\ € = 29\ € .$$

En examinant cet exemple de plus près, on voit que la tarification uniforme couvre en fait deux risques distincts : tout d'abord celui d'être prédisposé à la maladie M et ensuite celui de développer cette maladie. Formellement, chaque assuré s'acquitte d'une prime de 10 € , comme s'il n'était pas prédisposé, et ajoute à cela une prime

$$10\% \times (2\% - 0.1\%) \times 10\ 000\ € = 19\ €$$

qui le couvre contre le risque d'être prédisposé à la maladie M.

On constate aussi que cette approche permet de couvrir le risque à des conditions financières acceptables pour tous les individus, qu'ils soient prédisposés ou non. Ce point de vue, assimilant les polices offertes par un assureur qui ne segmente pas à des contrats multigaranties, a été développé par les économistes de l'assurance, dont Chiappori (1996,1997). Il justifie le montant plus élevé de la

prime pour les "bons" risques qui s'ignorent par une étendue plus large des garanties.

Considérons un portefeuille de 10 000 assurés dont la composition est semblable à celle de la population. Les individus ignorent donc leur éventuelle prédisposition. Ce portefeuille comprend donc en moyenne 1 000 assurés prédisposés à la maladie M, et 9 000 qui ne le sont pas. Parmi les premiers, 20 en moyenne développeront la maladie M, ce qui engendrera un coût de 200 000 € pour l'assureur. Parmi les seconds, 9 en moyenne développeront la maladie M, ce qui engendrera un coût de 90 000 € pour l'assureur. Le coût total de 290 000 € en moyenne sera compensé par l'encaissement, soit 10 000 primes de 29 € chacune.

Les problèmes surviennent lorsque l'assuré acquiert de l'information à propos de son niveau de risque individuel. En effet, l'avantage informationnel de l'assuré induit alors l'antisélection et oblige l'assureur à différencier le montant des primes, implicitement ou explicitement. Revenant à l'exemple de la maladie redoutée, si les individus prédisposés peuvent découvrir leur niveau de risque plus élevé (grâce à un test génétique par exemple, dont l'usage est interdit à l'assureur), ils vont s'assurer en masse, détériorant la statistique sinistre de l'assureur qui relèvera ses primes. Les individus dont le niveau de risque est plus faible finiront par renoncer à s'assurer et l'antisélection aura restreint le marché de l'assurance contre cette maladie.

Afin de s'en convaincre, supposons que les assurés se sachant prédisposés cherchent à s'assurer, et que 1 000 d'entre eux rejoignent les 10 000 assurés dont il a été question plus haut. L'encaissement de la compagnie passe à 11 000 x 29 € alors que la sinistralité vaut en moyenne

$$290\ \ 000 \in\ + 20 \times 10\ \ 000 \in ,$$

d'où un perte moyenne de 171 000 € . A la fin de la période, la compagnie est obligée de réévaluer la prime à la hausse, passant de 29 € à

$$29 \in\ + \frac{171\ \ 000 \in}{11\ \ 000} = 29 \in\ + 15.55 \in\ = 44.55 \in .$$

Chacun des 11 000 assurés doit donc augmenter sa contribution de 15.55 € pour avoir droit à la couverture.

Si cette tendance se confirme, la prime continuera à augmenter (pour tendre vers 200 €), jusqu'à décourager les assurés non prédisposés à se couvrir. Les assurés au profil le moins risqué sont

donc exclus du marché de l'assurance. L'assureur n'a d'autre alternative que de faire usage lui aussi du test génétique, de dépister la prédisposition à M et d'en tenir compte dans son tarif. La compagnie réclamera alors 10 € aux assurés non prédisposés à la maladie M, et 200 € aux autres. L'importance de la prime risque alors de décourager les assurés les plus risqués, les excluant du marché de l'assurance. A cet égard, les procédures de classification a priori peuvent être vues comme un moyen de diminuer les effets d'antisélection. Ce n'est que dans cette mesure que la différenciation des primes commerciales s'impose : si l'omission d'un facteur de segmentation induit de l'antisélection de la part des preneurs d'assurance, ce facteur doit techniquement être incorporé dans le tarif commercial.

Cet exemple simple se généralise à toutes les couvertures d'assurance, même si en pratique de nombreux effets viennent compliquer les choses : connaissance imparfaite des preneurs quant à leur niveau risque, multitudes de profils de risque (plutôt que deux dans l'exemple), conditions différentes de couverture, etc.

Il s'agit ici de bien saisir le fait que le problème réside dans la connaissance qu'a l'assuré de son profil de risque. Au contraire, si l'ignorance est symétrique, c'est-à-dire si ni l'assuré, ni l'assureur n'ont connaissance de facteurs influençant le niveau de risque, l'assureur couvrira le péril sur base du risque moyen collectif. Comme évoqué plus haut, lorsque les contractants n'ont pas connaissance d'informations de nature à influencer le niveau de risque, ils concluent en fait un contrat d'assurances multi-garanties : l'assureur couvre d'une part, le niveau de risque inconnu de l'assuré, et ensuite la sinistralité qui en découle.

L'ignorance symétrique a donc un aspect particulièrement avantageux en élargissant au maximum les possibilités d'assurance (c'est d'ailleurs la portée du paradoxe de Hirschleifer examiné dans le Tome I). Ainsi, dans l'exemple de la maladie redoutée, la généralisation des tests génétiques rendrait difficile la couverture de certaines populations à risque, alors que l'ignorance de la prédisposition de l'assuré à certaines maladies rend l'assurance possible. On comprend donc bien qu'il soit interdit aux assureurs d'utiliser des tests génétiques, mais il faut alors aussi que cet usage soit interdit aux assurés.

9.12.4 L'iniquité de la tarification a priori

L'assureur qui désire faire usage d'un critère de segmentation doit pouvoir démontrer, statistiques à l'appui, le lien causal entre ce critère et les variations de la sinistralité qu'il est supposé induire. Outre le fait que la mise en évidence d'un lien causal s'apparente à la quête du Graal pour un statisticien (hormis le cas d'un plan d'expérience dont les paramètres sont contrôlés), cette obligation discrédite beaucoup de critères de segmentation utilisés à l'heure actuelle : qui pourrait croire que le fait de cohabiter ou d'avoir plusieurs enfants améliore la qualité de la conduite automobile ? On s'attend plutôt à ce que la vie maritale ou la charge de famille incite à plus de prudence et au refus de toute prise de risque inconsidérée, donc à moins de témérité au volant, ce qui à son tour diminue le risque en RC automobile. Il n'y a donc pas de lien de causalité entre ces variables tarifaires et la sinistralité automobile. Les facteurs de risque pertinents sont cachés : agressivité au volant, respect du code de la route, consommation d'alcool, kilométrage annuel, etc., et ne peuvent donc être incorporés dans le tarif commercial.

Il y a cependant une solution envisageable en assurance automobile : le système PAYD (pour l'anglais "Pay As You Drive") développé de part et d'autre de l'Atlantique par Norwich Union en Grande Bretagne et par Progressive aux USA. Il s'agit d'une vision résolument novatrice de l'assurance automobile, rendue possible grâce aux technologies les plus récentes. Une "boîte noire" est placée sur le véhicule assuré, enregistrant la vitesse, les accélérations et décélérations, les kilomètres parcourus et le type de voirie empruntée. La vie privée est parfaitement protégée car les données sont filtrées avant de parvenir à l'assureur, ce dernier étant dès lors incapable sur base du résumé qui lui est fourni de reconstituer les trajets effectués. Par contre, ce résumé lui permet d'évaluer très précisément la qualité de conduite du véhicule : en comparant les vitesses enregistrées aux vitesses maximales autorisées selon le type de voirie, l'assureur peut apprécier la mesure dans laquelle les assurés respectent les limitations imposées par le code de la route, les accélérations et décélérations indiquent quant à elles la nervosité de la conduite, etc. L'assureur dispose ainsi de l'information pertinente pour tarifer le contrat.

Cette manière de procéder favorise également la sécurité routière, et permettrait sans aucun doute d'alléger la facture sociale très lourde des accidents de la route. Outre le calcul de la prime, le système présente des avantages accessoires, comme la localisation

possible du véhicule par les forces de police en cas de vol, l'avertissement des secours en cas d'accident, etc.

Le coût du système est raisonnable (on parle de 500 € la première année, couvrant l'acquisition du matériel et son placement, et d'un coût de maintenance limité les années suivantes). Le système est testé actuellement par 5 000 conducteurs en Grande Bretagne, et 5 000 autres aux USA. D'autres expériences sont en cours, notamment en Italie. Notez qu'à Singapour, un tel système est utilisé pour dresser les factures mensuelles de péage, en fonction des heures durant lesquelles l'automobiliste a utilisé les voiries, et le type des voiries empruntées.

9.13 Notes bibliographiques

Sur les scores, le lecteur pourra consulter Gouriéroux (1992,1999) ainsi que Bardos (2001). Les approches usuelles d'analyse des données (multivariées) sont présentées dans Saporta (1990) ou, plus récemment, dans Lebart, Morineau & Piron (2000). Les méthodes de classification sont présentées de manière claire et pédagogique dans Nakache & Confais (2004). La théorie des modèles linéaires généralisés remonte à Nelder & Wedderburn (1972). Elle a été très complètement exposée par McCullagh & Nelder (1989). Une excellente introduction est proposée par Dobson (2001).

Nous conseillons également aux lecteurs de consulter Bailey (1963) et Bailey & Leroy (1960,1964) à propos des origines de la classification des risques, ainsi que Anderson et al. (2004) à propos de la pratique des modèles linéaires généralisés en tarification. Parmi les précurseurs des modèles de régression en tarification, on notera Ter Berg (1980a,b), Albrecht (1983a,b,c) et surtout Renshaw (1994).

Le traitement des sinistres graves poura se faire à l'aide de la théorie des valeurs extrêmes, étudiée dans la suite de cet ouvrage. On pourra lire Cebrian, Denuit & Lambert (2003) pour un cas pratique.

Nous n'avons pas abordé le problème de la tarification par zone géographique. Une introduction abordable, avec un cas pratique, est donnée par Brouhns, Denuit, Masuy & Verrall (2002). Denuit & Lang (2004) passent en revue les différentes approches utilisées dans la tarification a priori, et proposent un modèle de tarification intégré basé sur les modèles additifs généralisés traités

dans le paradigme bayésien. Tous les facteurs tarifaires, qu'ils soient catégoriels, continus, spatiaux ou temporels, sont traités de manière unifiée.

Les études empiriques sont relativement rares dans la littérature. Mentionnons, en plus de celles citées plus haut, Ramlau-Hansen (1988) et Beirlant, Derveaux, De Meyer, Goovaerts, Labies & Maenhoudt (1991). Pour d'autres modèles de régression que ceux utilisés dans ce chapitre, voyez notamment Beirlant, Goegebeur, Verlaak & Vynckier (1998), Cummins, Dionne, McDonnald & Pritchett (1990), Ter Berg (1996) et Keiding, Andersen & Fledelius (1998).

Pour terminer, notez que les modèles décrits dans ce chapitre sont également fort utiles en assurance vie, pour dresser des tables de mortalité. Voyez Delwarde & Denuit (2005) pour plus de détails.

9.14 Exercices

Exercice 9.14.1. *Disposant des observations $X_1, X_2, \ldots, X_n$ indépendantes et de même loi, notons $X_{(1)} \leq \ldots \leq X_{(n)}$ les observations rangées par ordre croissant.*

(i) Définissons l'événement

$$A(i_1, i_2) = \left[X_{(i_1)} < q_p < X_{(i_2)}\right]$$

qui signifie que pas moins de i_1 observations sont inférieures au quantile d'ordre p, q_p, et pas moins de $n - i_2 + 1$ lui sont supérieures. Montrez que

$$\Pr[A(i_1, i_2)] \quad = \quad \sum_{j=i_1}^{i_2-1} \left(\begin{array}{c} n \\ j \end{array} \right) p^j (1-p)^{n-j}.$$

(ii) Montrez qu'un intervalle de confiance au niveau $1-\alpha$ pour q_p est $]x_{(i_1)}, x_{(i_2)}[$, où les entiers i_1 et i_2 satisfont la relation

$$1 - \alpha = \Pr[i_1 \leq \mathcal{B}in(n, p) < i_2].$$

(iii) Lorsque n est grand, montrez que i_1 et i_2 sont approximativement donnés par

$$i_1 = \lfloor -z_{\alpha/2} \sqrt{np(1-p)} + np \rfloor$$

et

$$i_2 = \lceil z_{\alpha/2} \sqrt{np(1-p)} + np \rceil$$

où $z_{\alpha/2}$ est tel que $\Phi(z_{\alpha/2}) = 1 - \alpha/2$ et où $\lfloor x \rfloor$ (resp. $\lceil x \rceil$) représente la partie entière du réel x (resp. le plus petit entier supérieur à x).

Exercice 9.14.2. *Montrez que pour la loi $\mathcal{B}in(m, p)$,*

$$\mathcal{I}(\widetilde{p}|p) = m\widetilde{p}\ln\left(\frac{\frac{\widetilde{p}}{1-\widetilde{p}}}{\frac{p}{1-p}}\right) + m\ln\frac{1-\widetilde{p}}{1-p}.$$

Exercice 9.14.3. *Montrez que pour la loi $\mathcal{P}oi(\lambda)$,*

$$\mathcal{I}(\widetilde{\lambda}|\lambda) = \widetilde{\lambda}\left(\ln\frac{\widetilde{\lambda}}{\lambda} - 1\right) + \lambda.$$

Exercice 9.14.4. *Montrez que pour la loi $\mathcal{N}or(\mu, \sigma^2)$,*

$$\mathcal{I}(\widetilde{\mu}|\mu) = \frac{(\widetilde{\mu} - \mu)^2}{2\sigma^2}.$$

Notez que, dans ce cas, $\mathcal{I}(\widetilde{\mu}|\mu)$ est une véritable distance.

Exercice 9.14.5. *Considérons le modèle linéaire simple $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, noté (1), où les ε_i sont indépendantes et de même loi $\mathcal{N}or(0, \sigma^2)$.*

(i) On pose $Z_i = Y_i - X_i$, $\delta_0 = \beta_0$ et $\delta_1 = \beta_1 - 1$, de sorte $Z_i = \delta_0 + \delta_1 X_i + \varepsilon_i$, noté (2) est équivalent au modèle (1). Comparez le R^2 obtenu dans les deux modèles.

(ii) En déduire qu'en régressant Y non pas sur X mais sur $Y - X$, il est parfois possible d'augmenter (artificiellement) la valeur du R^2.

Exercice 9.14.6. *Les résidus d'Anscombe sont basés sur une transformation A qui rapproche la loi de probabilité en question de la loi normale. Dans le cas de la famille exponentielle naturelle, A est donnée par*

$$A(\mu) = \int \frac{d\mu}{V^{1/3}(\mu)}.$$

Cependant, la transformation A qui "normalise" la loi de probabilité ne stabilise pas la variance. Il faut dès lors diviser par la racine carré de la variance de $A(Y)$ qui vaut au premier ordre $A'(\mu)\sqrt{V(\mu)}$.

(i) Montrez que le résidu d'Anscombe pour la loi de Poisson est donné par

$$r_i^A = \frac{3(y_i^{3/2} - \mu_i^{3/2})}{\mu_i^{1/6}}.$$

(ii) Montrez que le résidu d'Anscombe pour la loi Gamma est donné par

$$r_i^A = \frac{3(y_i^{1/3} - \mu_i^{1/3})}{\mu_i^{1/3}}.$$

Exercice 9.14.7. *(i) La compagnie A a en portefeuille 8 000 conducteurs expérimentés et 2 000 conducteurs novices. Ceux-ci ont déclaré respectivement 400 sinistres et 200 sinistres. Estimez (par maximum de vraisemblance dans un modèle de régression de Poisson) les scores associés à ces deux catégories d'assurés.*

(ii) Deux compagnies se partagent le marché : la compagnie A et un concurrent, la compagnie B, qui ne segmente pas son tarif a priori. Expliquez ce qui devrait se produire sur le marché en supposant les assurés rationnels.

Exercice 9.14.8. *Afin d'expliquer la fréquence annuelle de sinistres des assurés en fonction de leurs caractéristiques observables, l'actuaire en charge de la tarification RC Automobile d'une grande compagnie d'assurances a effectué une régression de Poisson : le nombre N_i de sinistres causés par l'assuré i est supposé de loi de Poisson de moyenne $d_i \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3})$, où d_i est la durée de couverture et*

$$x_{i1} = \begin{cases} 1, & \textit{si l'assuré } i \textit{ est une femme,} \\ 0, & \textit{sinon,} \end{cases}$$

$$x_{i2} = \begin{cases} 1, & \textit{si l'assuré } i \textit{ fractionne le paiement de sa prime,} \\ 0, & \textit{s'il la paie une fois par an,} \end{cases}$$

et

$$x_{i3} = \begin{cases} 1, & \textit{si l'assuré } i \textit{ conduit un véhicule diesel,} \\ 0, & \textit{s'il conduit un véhicule à essence.} \end{cases}$$

L'ajustement obtenu par maximum de vraisemblance est décrit dans le tableau suivant :

j	$\widehat{\beta}_j$	$\widehat{\sigma}(\widehat{\beta}_j)$
0	*-1.9575*	*0.0176*
1	*0.0701*	*0.0165*
2	*0.2770*	*0.0147*
3	*0.1869*	*0.0159*

On peut lire dans ce tableau les estimations ponctuelles des coefficients de régression et les écarts-type correspondants.

a) Quelles sont les équations dont $\widehat{\beta}_0$, $\widehat{\beta}_1$, $\widehat{\beta}_2$ et $\widehat{\beta}_3$ sont solutions ? Que garantissent-elles en termes de répartition de la sinistralité entre les différentes catégories d'assurés ?

b) Tous les facteurs de risque influencent-ils significativement la sinistralité en fréquence ?

c) Quelle est la fréquence annuelle de sinistres de l'assuré de référence ? Donnez-en une estimation ponctuelle et un intervalle de confiance à 95%.

d) Quelle est la fréquence annuelle de sinistres d'une femme payant sa prime une fois l'an et conduisant un véhicule à essence ? Donnez-en un intervalle de confiance à 95% si le coefficient de corrélation linéaire entre $\widehat{\beta}_0$ et $\widehat{\beta}_1$ est estimé à 0.1.

e) Quelle est la fréquence annuelle estimée de sinistres d'une femme fractionnant le paiement de sa prime et conduisant un véhicule à essence ?

f) Afin de détecter une éventuelle surdispersion, on porte en graphique la moyenne et la variance empiriques des nombres des sinistres observés au sein de chacune des classes. A quelle courbe s'agirait-il d'ajuster le nuage de points pour conforter l'hypothèse de loi de Poisson mélange. Expliquez et justifiez.

Exercice 9.14.9. *Afin d'expliquer la fréquence annuelle de sinistre des assurés en fonction de leurs caractéristiques observables, l'actuaire en charge de la tarification "RC Auto" d'une grande compagnie d'assurances a effectué une régression de Poisson, sur un fichier de 100 000 polices observées durant un an, en utilisant l'expérience de conduite (novice/expérimenté) et la puissance du véhicule (petite/moyenne/grosse cylindrée) : le nombre N_i de sinistres causés par l'assuré i est supposé de loi de Poisson de moyenne $d_i \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3})$, où d_i est la durée de couverture et*

$$x_{i1} = \begin{cases} 1, & \textit{si l'assuré } i \textit{ est un conducteur novice,} \\ 0, & \textit{sinon,} \end{cases}$$

$$x_{i2} = \begin{cases} 1, & \textit{si l'assuré } i \textit{ conduit un véhicule de petite cylindrée,} \\ 0, & \textit{sinon,} \end{cases}$$

et

$$x_{i3} = \begin{cases} 1, & \textit{si l'assuré } i \textit{ conduit un véhicule de grosse cylindrée,} \\ 0, & \textit{sinon.} \end{cases}$$

L'ajustement obtenu par maximum de vraisemblance fournit

$$\widehat{\beta}_0 = -2.10,\ \widehat{\beta}_1 = 0.06,\ \widehat{\beta}_2 = -0.03,\ \textit{et}\ \widehat{\beta}_3 = 0.05;$$

on donne également les erreurs standards :

$$\widehat{\sigma}\big(\widehat{\beta}_0\big) = 0.03,\ \widehat{\sigma}\big(\widehat{\beta}_1\big) = 0.02,\ \widehat{\sigma}\big(\widehat{\beta}_2\big) = 0.01,\ \textit{et}\ \widehat{\sigma}\big(\widehat{\beta}_3\big) = 0.01.$$

a) Quelles sont les équations dont les estimations ponctuelles données plus haut sont solutions ? Donnez-en une interprétation actuarielle.

b) Testez les hypothèses $\beta_j = 0$, $j = 0, 1, 2, 3$, et interprétez les résultats.

c) Estimez la fréquence annuelle de sinistre de l'assuré de référence, et donnez-en un intervalle de confiance.

d) Faites de même pour la fréquence annuelle d'un conducteur novice conduisant un véhicule de cylindrée moyenne si $\widehat{\mathbb{C}[\widehat{\beta}_0, \widehat{\beta}_1]} = 0.0001$.

Exercice 9.14.10. *Le coût des sinistres causés par l'assuré i se met sous la forme*

$$S_i = \sum_{k=1}^{N_i} C_{ik},$$

où N_i est de loi de Poisson de moyenne λ_i et les C_{ik} sont de loi Gamma, de moyenne μ_i et de variance μ_i^2/ν. Toutes les variables aléatoires en présence sont indépendantes.

L'actuaire en charge de la tarification utilise le sexe (homme/femme) et le domicile de l'assuré (rural/urbain), codés à l'aide de deux variables binaires, les niveaux de référence étant "homme" et "rural". Il intègre l'information via

$$\lambda_i = d_i \exp(\beta_0^{freq} + \beta_1^{freq} x_{i1} + \beta_2^{freq} x_{i2}),$$

avec d_i la durée d'exposition au risque, et

$$\mu_i = \exp(\beta_0^{coût} + \beta_1^{coût} x_{i1} + \beta_2^{coût} x_{i2}).$$

Les estimations ponctuelles de ces paramètres par maximum de vraisemblance sur base des observations relatives à un grand portefeuille donnent les résultats suivants :

$$\widehat{\beta}_0^{freq} = -1.95,\ \widehat{\beta}_1^{freq} = 0.03,\ \textit{et}\ \widehat{\beta}_2^{freq} = 0.05,$$

pour les fréquences, et

$$\widehat{\beta}_0^{coût} = 6.64, \;\; \widehat{\beta}_1^{coût} = -0.05, \;\; et \; \widehat{\beta}_2^{coût} = -0.07,$$

pour les coûts, avec $\widehat{\nu} = 0.75$. *On donne également les erreurs standards :*

$$\widehat{\sigma}\big(\widehat{\beta}_0^{freq}\big) = 0.01, \;\; \widehat{\sigma}\big(\widehat{\beta}_1^{freq}\big) = 0.01, \;\; et \; \widehat{\sigma}\big(\widehat{\beta}_2^{(freq)}\big) = 0.02,$$

pour les fréquences, et

$$\widehat{\sigma}\big(\widehat{\beta}_0^{coût}\big) = 0.02, \;\; \widehat{\sigma}\big(\widehat{\beta}_1^{coût}\big) = 0.01, \;\; et \; \widehat{\sigma}\big(\widehat{\beta}_2^{coût}\big) = 0.02,$$

pour les coûts.

a) Décrivez l'algorithme itératif qui a permis d'obtenir les estimations ponctuelles $\widehat{\beta}_0^{freq}$, $\widehat{\beta}_1^{freq}$ *et* $\widehat{\beta}_2^{freq}$ *données plus haut.*

b) Quelles sont les équations dont les estimations ponctuelles $\widehat{\beta}_0^{coût}$, $\widehat{\beta}_1^{coût}$ *et* $\widehat{\beta}_2^{coût}$ *sont solutions ? Donnez-en une interprétation actuarielle.*

c) Estimez la prime pure $\mathbb{E}[S_i]$ *et la variance* $\mathbb{V}[S_i]$ *pour les différentes classes tarifaires.*

Bibliographie

[1] Albrecht, P. (1983a). Parametric multiple regression risk models : Connections with tariffication, especially in motor insurance. *Insurance : Mathematics & Economics* **2**, 113-117.

[2] Albrecht, P. (1983b). Parametric multiple regression risk models : Theory and statistical analysis. *Insurance : Mathematics & Economics* **2**, 49-66.

[3] Albrecht, P. (1983c). Parametric multiple regression risk models : Some connections with IBNR. *Insurance : Mathematics & Economics* **2**, 69-73.

[4] Allais, M. (1953). Le comportement de l'homme rationnel devant le risque, critiques des postulats et axiomes de l'Ecole américaine. *Econometrica* **21**, 503-546.

[5] Anderson, D., Feldblum, S., Modlin, C., Schirmacher, D., Schirmacher, E., & Thandi, N. (2004). A practitioner's guide to generalized linear models. *Discussion Papers of the Casualty Actuarial Society.*

[6] Antoniadis, A., Berruyer, J., & Carmona, R. (1992). *Régression Non-Linéaire et Applications.* Economica, Paris.

[7] Arrow, K.J. (1971). *Essays in the Theory of Risk Bearing.* Chicago, Markham Publishing Compagny.

[8] Astesan, E. (1938). *Les Réserves Techniques des Sociétés d'Assurances contre les Accidents d'Automobiles.* Collection d'études sur le droit des assurances.

[9] Bailey, R.A. (1963). Insurance rates with minimum bias. *Proceedings of the Casualty Actuarial Society* **L**, 4-14.

[10] Bailey, R.A., & LeRoy, J.S. (1960). Two studies in automobile insurance ratemaking : A. Effectiveness of merit rating and class rating. *Proceedings of the Casualty Actuarial Society* **47**, 1–19.

[11] Bailey, R.A., & Leroy, S.J. (1964). Two studies in automobile insurance ratemaking : B. Improved methods for determining classification rate relativities. *Proceedings of the Casualty Actuarial Society* **L**, 192-217.

[12] Bardos, M. (2001). *Analyse Discriminante, Application au Risque et Scoring Financier*. Dunod, Paris.

[13] Barjonnet, Khlifi & Mignot (1997) De la représentation du risque en zone urbaine à sa prévention : décalages et contradictions. Rapport de recherche INRETS.

[14] Beck, U. (2001). *La Société du Risque : Sur la Voie d'une autre Modernité*. Aubier, Alto.

[15] Beirlant, J., Derveaux, V., De Meyer, A.M., Goovaerts, M.J., Labies, E., & Maenhoudt, B. (1991). Statistical risk evaluation applied to (Belgian) car insurance. *Insurance : Mathematics & Economics* **10**, 289-302.

[16] Beirlant, J., Goegebeur, Y., Segers, J., & Teugels, J.L. (2004). *Practical Analysis of Extreme Values*. Leuven University Press, Belgium.

[17] Beirlant, J., Goegebeur, Y., Verlaak, R., & Vynckier, P. (1998). Burr regression and portfolio segmentation. *Insurance : Mathematics & Economics* **23**, 231-250.

[18] Beirlant, J., Teugels, J.L., & Vynckier, P. (1996). *Practical Analysis of Extreme Values*. Leuven University Press, Belgium.

[19] Bénéplanc, G., & Rochet, J.C. (2005). *Risk Management & Corporate Financing. à paraître.*

[20] Bermúdez, L, Denuit, M., & Dhaene, J. (2001). Exponential bonus-malus systems integrating a priori risk classification. *Journal of Actuarial Practice* **9**, 84-112.

[21] Besson, J.-L., & Partrat, Ch. (1992). Trend et systèmes de Bonus-Malus. *ASTIN Bulletin* **22**, 11-31.

[22] Bingham, N.H., Goldie, C.M., & Teugels, J.L. (1987). *Regular Variation*. Cambridge University Press.

[23] Borch, K. (1962). Equilibrium in a reinsurance market. *Econometrica* **30**, 424-444.

[24] Borgan, Ø, Hoem, J.M., & Norberg, R. (1981). A nonasymptotic criterion for the evaluation of automobile bonus systems. *Scandinavian Actuarial Journal*, 265-178.

[25] Boyer, R., Chavance, B., & Godard, O. (1991). La dialectique réversibilité-irréversibilité. In *Les Figures de l'Irréversibilité en Economie*, Boyer, Chavance & Godard eds, Ecole des Hautes Etudes en Sciences Sociales.

[26] Briys, E., & de Varenne, F. (1998). Assurance et marchés financiers : concurrence ou complémentarité. In *Encyclopédie de l'Assurance*, Ewald & Lorenzi eds, Economica.

[27] Brouhns, N., & Denuit, M. (2003). Actuarial modelling of longitudinal claims data through GAMM's : some methodological results. *German Actuarial Bulletin* **26**, 25-39.

[28] Brouhns, N., Denuit, M., Masuy, B., & Verrall, R. (2002). Ratemaking by geographical area in the Boskov and Verrall model : A case study using Belgian car insurance data. *actu-L* **2**, 3-28.

[29] Brouhns, N., Guillén, M., Denuit, M., & Pinquet, J. (2003). Bonus-malus scales in segmented tariffs with stochastic migration between segments. *Journal of Risk and Insurance* **70**, 577-599.

[30] Bühlmann, H. (1964). Optimale Prämienstufensysteme. *Bulletin of the Swiss Association of Actuaries*, 193-213.

[31] Bühlmann H. (1967). Experience rating and credibility I. *ASTIN Bulletin* **4**, 199–207.

[32] Bühlmann H. (1969). Experience rating and credibility II. *ASTIN Bulletin* **5**, 157–165.

[33] Bühlmann, H. (1970). *Mathematical Methods in Risk Theory*. Springer Verlag, New York.

[34] Cebrian, A., Denuit, M., & Lambert, Ph. (2003). Generalized Paréto fit to the Society of Actuaries' large claims database. *North American Actuarial Journal* **7**, 18-36.

[35] Celeux, G., & Nakache, J.P. (1994). *Analyse Discriminante sur Variables Qualitatives*. Polytechnica, Paris.

[36] Centeno, M., & Silva, J.M.A. (2001). Bonus systems in an open portfolio. *Insurance : Mathematics & Economics* **28**, 341-350.

[37] Chiappori, P.A. (1997). *Risque et Assurance*. Flammarion.

[38] Chiappori, P.A. (1999). Tests génétiques et assurance : une analyse économique. *Risques* **40**, 107–109.

[39] Cleveland, W.S. (1979). Robust locally-weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* **74**, 829-836.

[40] Cleveland, W.S., & Devlin, S.J. (1988). Locally weighted regression : An approach to regression by local fitting. *Journal of the American Statistical Association* **83**, 596-610.

[41] Cleveland, W.S., Devlin, S.J., & Grosse, E. (1988). Regression by local fitting. *Journal of Econometrics* **37**, 87-114.

[42] Cleveland, W.S., & Grosse, E. (1991). Computational methods for local regression. *Statistics and Computing* **1**, 47-62.

[43] Cohen, M., & Tallon, J. M. (2000). Décision dans le risque et l'incertain : l'apport des modèles non-additifs. *Revue d'Economie Politique, Bilans et Essais* **5**, 631-681.

[44] Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer Verlag, New York.

[45] Cummins, D.J., Dionne, G., McDonnald, J.B., & Pritchett, M.B. (1990). Application of the GB2 family of distribution in modelling insurance loss processes. *Insurance : Mathematics and Economics* **9**, 257-272.

[46] Dab, W. (1997). Précaution et santé publique : le cas des champs électriques et magnétiques de basse fréquence. In *Le Principe de Précaution dans la Conduite des Affaires humaines*, Godard éd., Maison des sciences de l'homme.

[47] Dannenburg D.R., Kaas R. & Goovaerts M.J. (1996). *Practical actuarial credibility models*. Institute of Actuarial Science, Amsterdam.

[48] David, H.A. (1981). *Order Statistics*. Wiley, New York.

[49] Davison, A.C. & Hinkley, D. V. (1997). *Bootstrap Methods and Their Application*. Cambridge University Press.

[50] Delaporte, P. (1959). Quelques problèmes de statistique mathématique posés par l'assurance automobile et le bonus pour non sinistre. *Bulletin Trimestriel de l'Institut des Actuaires Français* **70**, 87-102.

[51] Delwarde, A., & Denuit, M. (2005). *Construction de Tables de Mortalité Périodiques et Prospectives*. Economica, Paris.

[52] Denuit, M., & Dhaene, J. (2001). Bonus-Malus scales using exponential loss functions. *German Actuarial Bulletin* **25**, 13-27.

[53] Denuit, M., Dhaene, J., Goovaerts, M. & Kaas, R. (2005). *Actuarial Theory for Dependent Risks*. Wiley, New York.

[54] Denuit, M., Dhaene, J., & Van Wouwe, M. (1999). The economics of insurance : a review and some recent developments. *Bulletin of the Swiss Association of Actuaries*, 137-175.

[55] Denuit, M., & Lambert, Ph. (2001). Smoothed NPML estimation of the risk distribution underlying Bonus-Malus systems. *Proceedings of the Casualty Actuarial Society* **88**, 142-174.

[56] Denuit, M., & Lang, S. (2004). Nonlife ratemaking with Bayesian GAM's. *Insurance : Mathematics and Economics* **35**, 627-647.

[57] Denuit, M., Maréchal, X., & Closon, J.-P. (2005). Etude relative aux coûts potentiels liés à une éventuelle modification des règles du droit de la responsabilité médicale. Phase II : Développement d'un modèle actuariel et premières estimations. KCE Reports - Centre Fédéral d'Expertise des Soins de Santé, Volume 16B. Disponible sur http ://www.centredexpertise.fgov.be/fr/Publications.html

[58] Denuit, M., Pitrebois, S., & Walhin, J.-F. (2001). Méthodes de construction de systèmes bonus-malus en RC auto. *actu-L* **1**, 7-38.

[59] Denuit, M., Pitrebois, S. & Walhin, J.-F. (2003). Tarification automobile sur données de panel. *Bulletin of the Swiss Association of Actuaries*, 51-81.

[60] Denuit, M., Purcaru, O., & Van Keilegom, I. (2004). Bivariate archimedean copula modelling for Loss-ALAE data in nonlife insurance. Working Paper 04-03, Institut des Sciences Actuarielles, Université Catholique de Louvain, Louvain-la-Neuve, Belgium.

[61] Desjardins, D., G., Dionne, & Pinquet, J. (2001). Experience rating schemes for fleets of vehicles. *ASTIN Bulletin* **31**, 81-105

[62] Devroye, L. (1992). *Non Uniform Random Variate Generation*. Springer-Verlgag, New York.

[63] Dionne, G., & Vanasse, C. (1989). A generalization of actuarial automobile insurance rating models : the Negative Binomial distribution with a regression component. *ASTIN Bulletin* **19**, 199-212.

[64] Dionne, G., & Vanasse, C. (1992). Automobile insurance ratemaking in the presence of asymmetrical information. *Journal of Applied Econometrics* **7**, 149-165.

[65] Dobson, A.J. (2001). *An Introduction to Generalized Linear Models*. Chapman & Hall/CRC, London.

[66] Dodge, Y., & Rousson, V. (2004). *Analyse de Régression Appliquée*. Dunod, Paris.

[67] Doherty, N., & Schlesinger, H. (1983). Optimal insurance in incomplete markets. *Journal of Political Economy* **91**, 1045-1054.

[68] Duché, P. (1998). Origine et développement des "captives". In *Encyclopédie de l'Assurance*, Ewald & Lorenzi eds, Economica.

[69] Dupuy, J.P. (2001). *Pour un Catastrophisme Eclairé. Quand l'Impossible est Certain*. Seuil.

[70] Eeckhoudt, L., & Kimball, M. (1992). Background risk, prudence and the demand for insurance. In *Contributions to Insurance Economics*, Dionne éd., Kluwer Academic Publishers, 239-254.

[71] Efron, B. & Tibshirani,R.J. (1994). *An Introduction to the Bootstrap*. Chapman and Hall, London.

[72] Ehrlich, I., & Becker, G. (1972). Market insurance, self-insurance and self-protection. *Journal of Political Economy*, 623-648.

[73] Embrechts, E., Klüpelberg, C., & Mikosch, T. (1997). *Modeling Extremal Events for Insurance and Finance*. Springer Verlag, New York.

[74] England, P. (2002). Addendum to "Analytic and bootstrap estimates of prediction errors in claims reserving". *Insurance : Mathematics & Economics* **31**, 461-466.

[75] England, P., & Verrall, R. (1999). Analytic and bootstrap estimates of prediction errors in claims reserving. *Insurance : Mathematics & Economics* **25**, 281-293.

[76] England, P., & Verrall, R.J. (2002). Stochastic claims reserving in general insurance. *British Actuarial Journal*.

[77] Epstein, L.G. (1980). Decision making and the temporal resolution of uncertainty. *International Economic Review*. **21**, 269–283.

[78] Ewald, F. (1986). *L'Etat-Providence*. Grasset.

[79] Ewald, F., Gollier, C., & de Sadeleer, N. (2001). *Le Principe de Précaution*. Presses Universitaires de France, Que sais-je ?

[80] Fahrmeir, L., & Tutz, G. (2002). *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer–Verlag, New York.

[81] Fang, K.T., Kotz, S., & Ng, K.W. (1990). *Symmetric Multivariate and Related Distributions*. Chapman & Hall, London.

[82] Fishburn, P.C. (1970). *Decision and Value Theory*. John Wiley, New York.

[83] Fisher, R.A., & Tippett, L.H.C. (1928). On the estimation of the frequency distribution of the largest or smallest member of a sample. *Proceedings of the Cambridge Philosophical Society* **24**, 180-190.

[84] Fishman, G.S. (1997). *Monte Carlo : Concepts, Algorithms and Applications*. Springer-Verlag, New York.

[85] Franckx, E.. (1960). Théorie du bonus, conséquence de l'étude de Mr le professeur Fréchet. *ASTIN Bulletin* **1**, 113-122.

[86] Fréchet, M. (1959). Essai d'une étude des successions de sinistres considérés comme processus stochastique. *Bulletin Trimestriel de l'Institut des Actuaires Français* **70**, 67-85.

[87] Frees, E.W., & Valdez, E.A. (1998). Understanding relationships using copulas. *North American Actuarial Journal* **2**, 1-25.

[88] Galambos, J. (1987). *The Asymptotic Theory of Extreme Order Statistics*. Krieger.

[89] Gayant, J.P. (2001). *Risque et Décision*. Vuibert.

[90] Gentle, J.E. (1998). *Random Number Generation and Monte Carlo Methods*. Springer-Verlag, New York.

[91] Gerber, H.U. (1987). Actuarial applications of utility functions. In "*Advances in the Statistical Sciences, vol. 6, Actuarial Sciences*" 53-61, MacNeill, I., and Umphrey, G., Editors. Reidel, Dordrecht.

[92] Gerber, H.U., & Jones, D. (1975). Credibility formulas of the updating type. *Transactions of the Society of Actuaries* **27**, 31-52.

[93] Gerber, H.U., & Pafumi, G. (1998). Utility functions : From risk theory to finance. *North American Actuarial Journal* **2**, 74-91.

[94] Gilde, V., & Sundt, B. (1989). On bonus systems with credibility scales. *Scandinavian Actuarial Journal*, 13-22.

[95] Gnedenko, B. (1943). Sur la distribution limite du terme maximum d'une série aléatoire. *The Annals of Mathematics* **44**, 423-453.

[96] Godard, O. (1997). L'ambivalence de la précaution et la transformation des rapports entre science et décision. In *Le Principe de Précaution dans la Conduite des Affaires Humaines*, Godard éd., Maison des sciences de l'homme.

[97] Godard, O. (1997). *Le Principe de Précaution dans la Conduite des Affaires Humaines*. Maison des sciences de l'homme.

[98] Godard, O., Henry, C., Lagadec, P., & Michel-Kerjan, E. (2002). *Traité des Nouveaux Risques*. Gallimard.

[99] Godard, O. & Salles, J.M. (1991). Entre nature et société, les jeux de l'irréversibilité dans la construction économique et sociale du champ de l'environnement. In *Les Figures de l'Irréversibilité en Economie*, Boyer, Chavance & Godard éds., Ecole des Hautes Etudes en Sciences Sociales.

[100] Gouriéroux, C. (1992). Courbes de performances et de discrimination. *Annales d'Economie et de Statistique* **28**, 107–123.

[101] Gouriéroux, C. (1999). *Statistique de l'Assurance*. Economica, Paris.

[102] Gouriéroux, C., Monfort, A., & Trognon, A. (1984). Pseudo maximum likelihood methods : Theory. *Econometrica* **52**, 681-720.

[103] Gumbel, E. (1958). *Statistics of Extremes*. Columbia University Press.

[104] Hastie, T., & Tibshirani, R. (1990). *Generalized Additive Models*. Chapman and Hall, London.

[105] Henriet, D., & Rochet, J.C. (1991). *Microéconomie de l'Assurance*. Economica, Paris.

[106] Henry, M. (1937). Etude sur le coût moyen des sinistres en responsabilité civile automobile. *Bulletin Trimestriel de l'Institut des Actuaires Français* **169**.

[107] Hermitte, M.A. (1997). Le principe de précaution à la lumière du drame de la transfusion sanguine. In *Le Principe de Précaution dans la Conduite des Affaires Humaines*, Godard éd., Maison des sciences de l'homme.

[108] Hosking, J., & Wallis, J. (1987). Parameter and quantile estimation for the generalized Pareto distribution. *Technometrics* **29**, 339-349.

[109] Hurwich, C.M., & Simonoff, J.S. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society - Series B* **60**, 271-293.

[110] Ingenbleek, J.F., & Lemaire, J. (1988). What is a Sports Car ? *ASTIN Bulletin* **18**, 175-187.

[111] IPCC - Intergovernmental Panel on Climate Change (2001). Climate change, synthesis report.

[112] Jeleva, M., & Villeneuve, B. (2003). Insurance contracts with imprecise probabilities and adverse selection. *Economic Theory*.

[113] Jonas, H. (1990). *Le Principe de Responsabilité*. Le Cerf.

[114] Jorgensen, B., & Paes de Souza, M.C. (1994). Fitting Tweedie's compound Poisson model to insurance claims data. *Scandinavian Actuarial Journal*, 69-93.

[115] Kaas, R., Goovaerts, M.J., Dhaene, J., & Denuit, M. (2001). *Modern Actuarial Risk Theory*. Kluwer Academic Publishers, Dordrecht.

[116] Kant, E. (1781). *Critique de la Raison Pure*. Flammarion.

[117] Keiding, N., Andersen C., & Fledelius, P. (1998). The Cox regression model for claims data in non-life insurance. *ASTIN Bulletin* **28**, 95-118.

[118] Kelle, M. (2000). Modélisation du système de bonus malus français. *Bulletin Français d'Actuariat* **4**, 45-64.

[119] Kimball, M. (1990). Precautionary saving in the small and in the large. *Econometrica* **58**, 53-73.

[120] Klugman, S.A., Panjer, H.H. & Willmot, G.E. (1998). *Loss Models : From Data to Decisions*. Wiley-Interscience.

[121] Klugman, S.A., & Parsa, R. (1999) Fitting bivariate loss distributions with copulas. *Insurance : Mathematics and Economics* **24**, 139-148.

[122] Kotz, S., & Nadarajah, S. (2000). *Extreme Value Distributions : Theory and Applications*. Imperial College Press.

[123] Kourilsky, K. (2002). *Du Bon Usage du Principe de Précaution*. Odile Jacob.

[124] Kunreuther, H., Ginsberg, R., Miller, L., Sagi, P., Slovic, P., Borkan, B., & Katz, N. (1978). *Disaster Insurance Protection : Public Policy Lessons*. Wiley, New York.

[125] Laffont, J.J. (1991). *Economie de l'Incertain et de l'Information – Volume 2 du Cours de Théorie Microéconomique*. Economica, Paris.

[126] Lagadec, P. (1981). *La Civilisation du Risque. Catastrophes Technologiques et Responsabilité Sociale*. Seuil.

[127] Landsberger, M., & Meilijson, I. (1994). Monopoly insurance under adverse selection when agents differ in risk aversion. *Journal of Economic Theory* **63**, 392-407.

[128] Latour, B. (2001). Du principe de précaution au principe du bon gouvernement. *Etudes* **3394**, 339–346.

[129] Lebaron, F. (2000). *La Croyance Economique : Les Economistes entre Science et Politique*. Seuil.

[130] Lebart, L., Morineau, A., & Piron, M. (2000). *Statistique Exploratoire Multidimensionnelle*. Dunod, Paris.

[131] L'Ecuyer, P. (1994). Uniform random number generation. *Annals of Operations Research* **53**, 77–120.

[132] Lemaire, J. (1995). *Bonus-Malus Systems in Automobile Insurance*. Kluwer Academic Publishers, Boston/Dordrecht/London.

[133] Liang, K.Y., & Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13-22.

[134] Mack, T. (1993). Distribution-free calculation of the standard error of chain ladder reserve estimates. *ASTIN Bulletin* **29**, 361-366.

[135] Mack, T. (1994). Which stochastic model is underlying the chain ladder method ? *Insurance : Mathematics and Economics* **15**, 133-138.

[136] Magis, C., Denuit, M., & Walhin, J.-F. (2003). Au-delà de Chain-Ladder : Les méthodes stochastiques de réservation. *actu-L* **3**, 31-58.

[137] Malinvaud, E. (1999). *Leçons de Théorie Microéconomique*. Dunod.

[138] Matheu, M. (2002). *La Décision Publique Face aux Risques : Rapport du Séminaire "Risques"*. La Documentation française.

[139] Mc Cullagh, P., & Nelder, J.A. (1989). *Generalized Linear Models*. Chapman & Hall, New York.

[140] Michel-Kerjan, E. (2003). Les Etats-Unis face à la menace terroriste. *Risques* **54**, 81–90.

[141] Michel-Kerjan, E. (2004). Terrorisme à grande échelle : partage des risques et politiques publiques. *Cahiers du Laboratoire d'Econométrie de l'Ecole Polytechnique*, **2004-06**.

[142] Monfort, A. (1996). *Cours de Statistique Mathématique*. Economica, Paris.

[143] Moss, D. (2002). *When All Else Fails : Government as the Ultimate Risk Manager*. Harvard University Press.

[144] Mossin, J. (1968). Aspects of rational insurance purchasing. *Journal of Political Economy* **76**, 553-568.

[145] Mowbray, A.H. (1914). How extensive a payroll exposure is necessary to give a dependable pure premium. *Proceedings of the Casualty Actuarial Society* **1**, 24–30.

[146] Nakache, J.P. & Confais, J. (2004). *Approche Pragmatique de la Classification*. Technip, Paris.

[147] Nelder, J.A., & Wedderburn, R.W.M. (1972). Generalized linear models. *Journal of the Royal Statistical Society - Series A* **135**, 370–384.

[148] Norberg, R. (1976). A credibility theory for automobile bonus systems. *Scandinavian Actuarial Journal*, 92-107.

[149] Palm, R. (1998). Urban earthquake hazards : The impacts of culture on perceived risk and response in the USA and Japan. *Applied Geography*, **18**, 35–46.

[150] Peretti-Watel, P. (2000). *Sociologie du Risque*. Armand Colin.

[151] Pinquet, J. (1997). Allowance for cost of claims in Bonus-Malus systems *ASTIN Bulletin* **27**, 33-57.

[152] Pinquet, J. (1998). Designing optimal Bonus-Malus systems from different types of claims. *ASTIN Bulletin* **28**, 202-220.

[153] Pinquet, J. (1999). Experience rating through heterogeneous models. In *Handbook of Insurance*, edited by G. Dionne. Kluwer Academic Publishers.

[154] Pinquet, J., Guillen, M., & Bolancé, C. (2001). Allowance for the age of claims in bonus-malus systems. *ASTIN Bulletin* **31**.

[155] Pitrebois, S., De Longueville, Ph., Denuit, M., & Walhin, J.-F. (2002). Etude de techniques IBNR modernes. *actu-L* **2**, 29-62.

[156] Pitrebois, S., Denuit, M., & Walhin, J.-F. (2003a). Fitting the Belgian Bonus-Malus system. *Belgian Actuarial Bulletin* **3**, 58-62.

[157] Pitrebois, S., Denuit, M., & Walhin, J.-F. (2003b). Setting a bonus-malus scale in the presence of other rating factors : Taylor's work revisited. *ASTIN Bulletin* **33**, 419-436.

[158] Pitrebois, S., Denuit, M., & Walhin, J.-F. (2003c). Marketing et systèmes bonus-malus. *actu-L* **3**, 89-105.

[159] Pitrebois, S., Denuit, M., & Walhin, J.-F. (2004). Bonus-malus scales in segmented tariffs : Gilde & Sundt's work revisited. *Australian Actuarial Journal* **10**, 107-125.

[160] Pitrebois, S., Denuit, M., & Walhin, J.-F. (2005a). Multi-event bonus-malus scales. *Journal of Risk and Insurance*.

[161] Pitrebois, S., Denuit, M., & Walhin, J.-F. (2005b). Bonus-malus systems with varying deductibles. *ASTIN Bulletin*.

[162] Popper, K. (1987). *La Logique de la Découverte Scientifique*. Payot.

[163] Pratt, J. W. (1964). Risk aversion in the small and in the large. *Econometrica* **32**, 122-136.

[164] Purcaru, O., & Denuit, M. (2002a). On the dependence induced by frequency credibility models. *Belgian Actuarial Bulletin* **2**, 73-79.

[165] Purcaru, O., & Denuit, M. (2002b). On the stochastic increasingness of future claims in the Bühlmann linear credibility premium. *German Actuarial Bulletin* **25**, 781-793.

[166] Purcaru, O., & Denuit, M. (2003). Dependence in dynamic claim frequency credibility models. *ASTIN Bulletin* **33**, 23-40.

[167] Purcaru, O., Guillén, M., & Denuit, M. (2004). Linear credibility models based on time series for claim counts. *Belgian Actuarial Bulletin* **4**, 62-74.

[168] Quiggin, J.P. (1982). A theory of anticipated utility. *Journal of Economic Behavior and Organization* **3**, 323-343.

[169] Quiggin, J. (1993). *Generalized Expected Utility Theory : The Rank-Dependent Model*. Kluwer, Boston.

[170] Ramlau-Hansen, H. (1988). A solvency study in non-life insurance. *Scandinavian Actuarial Journal*, 3-34.

[171] Rasmussen, N.C. (1975). Reactor Safety Study : An Assessment of Accident Risks in US Commercial Nuclear Power Plants. *U.S. Nuclear Regulatory Commission Report* **WASH-1400**.

[172] Reid, D.H. (1986). Discussion of methods of claim reserving in non-life insurance. *Insurance : Mathematics and Economics* **5**, 45-56.

[173] Reiss, R.D. & Thomas, M. (2001). *Statistical Analysis of Extreme Values with Application to Insurance, Finance, Hydrology and Other Fields*. Birküser.

[174] Renshaw, A.E. (1994). Modelling the claim process in the presence of covariates. *ASTIN Bulletin* **24**, 265-285.

[175] Ripley, B.D. (1987). *Stochastic Simulation*. Wiley, New York.

[176] Robert, C. (1996). *Méthodes de Monte Carlo par Chaînes de Markov*. Economica, Paris.

[177] Rochet, J.C. (1998). Assurabilité et financement des risques. In *Encyclopédie de l'assurance*, Ewald & Lorenzi eds, Economica.

[178] Rolski, T., Schmidli, H., Schmidt, V., & Teugels, J. (1999). *Stochastic Processes for Insurance and Finance*. Wiley, New York.

[179] Rosenblatt, M. (1952). Remarks on a multivariate transformation. *Annals of Mathematical Statistics* **23**, 470-472.

[180] Rothschild, M., & Stiglitz, J.E. (1976). Equilibrium in competitive insurance markets : An essay on the economics of

imperfect information. *Quarterly Journal of Economics* **90**, 629-649.

[181] Rubinstein, R.Y. (1981). *Simulation and Monte Carlo Method*. Wiley, New York.

[182] Saporta, G. (1990). *Probabilités, Analyse des Données et Statistique*. Technip, Paris.

[183] Savage, L.J. (1954). *The Foundation of Statistics*. Dover Publications.

[184] Schmidt, K.D. (1999). A bibliography on loss reserving. Available from www.math.tu-dresden.de/sto/schmidt/dsvm/dsvm99-4.pdf

[185] Sherman, R.E. (1984). Extrapolating, smoothing, and interpolating development factors. *Proceedings of the Casualty Actuarial Society* **LXXI**, 122-155.

[186] Stiglitz, J. E. (1977). Monopoly, non-linear pricing and imperfect information : The insurance market. *Review of Economic Studies* **44**, 407-430.

[187] Sundt, B. (1988). Credibility estimators with geometric weights. *Insurance : Mathematics and Economics* **7**, 113-122.

[188] Taylor, G. (1997). Setting a Bonus-Malus scale in the presence of other rating factors. *ASTIN Bulletin* **27**, 319-327.

[189] Taylor, G.C. (2000). *Loss Reserving : An Actuarial Perspective*. Kluwer Academic Publishers.

[190] Ter Berg, P. (1980a). Two pragmatic approaches to loglinear claim cost analysis. *Astin Bulletin* **11**, 77-90

[191] Ter Berg, P. (1980b). On the loglinear Poisson and gamma model. *ASTIN Bulletin* **11**, 35-40.

[192] Ter Berg, P. (1996). A loglinear lagrangian Poisson Model. *ASTIN Bulletin* **26**, 123-129.

[193] Thépaut, A. (1959). Aspect politique et aspect administratif du Bonus pour non sinistre. *Bulletin Trimestriel de l'Institut des Actuaires Français* **70**, 125-39.

[194] Tosetti, A., Béhar, T., Fromenteau, M., & Ménart (2000). *Assurance : Comptabilité, Réglementation, Actuariat*. Economica, Paris.

[195] Treich, N. (1997). Vers une théorie économique de la précaution ? *Risques* **32**.

[196] Von Neumann, J., & Morgenstern, O. (1947). *Theory of Games and Economic Behavior*. Princeton University Press.

[197] Walhin, J.-F., & Paris, J. (2000). The true claim amount and frequency distributions within a bonus-malus system, *ASTIN Bulletin* **30**, 391-403

[198] Walhin, J.-F., & Paris, J. (2001). The practical replacement of a bonus-malus system, *ASTIN Bulletin* **31**, 317-335

[199] Whitney, A.W. (1918). The theory of experience rating. *Proceedings of the Casualty Actuarial Society* **4**, 274–292.

[200] Wigley, T.M.L., Richels, R. & Edmonds, J. (1996). Economic and environmental choices in the stabilization of atmospheric $C0_2$ concentrations. *Nature*. **379**, 240–243.

[201] Ycart, B. (2002). *Simulation de Variables Aléatoires*. Cahiers de Mathématiques Appliquées, Presses Universitaire de Tunis.

[202] Zajdenweber, D. (1996). Extreme values in business interruption insurance. *The Journal of Risk and Insurance* **63**, 95–110.

[203] Zajdenweber, D. (2001). *Economie des Extrêmes*. Flammarion.

[204] Zeger, S.L., Liang, K.Y., & Albert, P.S. (1988). Models for longitudinal data : A generalized estimating equation approach. *Biometrics* **44**, 1049-1060.