$$\Lambda_x = \sum_{t=1}^{\infty} \frac{1}{(1+i)^t} \mathbb{1}_{]t;\infty[}(T_x)$$

ressources-actuarielles.net

ISFA GRANDE ÉCOLE D'ACTUARIAT ET DE GESTION DES RISQUES
INSTITUT DE SCIENCE FINANCIÈRE ET D'ASSURANCES

Tarification IARD

Introduction aux techniques avancées

Version 1.5

Décembre 2023

Frédéric PLANCHET
frederic@planchet.net

Antoine MISERAY
antoine.miseray@primact.fr

Préambule

La réalisation d'un tarif en assurance IARD (auto, MRH, construction, *etc*.) s'appuie classiquement sur l'analyse de la prime pure dans le cadre d'un modèle fréquence x coût dans lequel l'effet des variables explicatives sur le niveau du risque est modélisé par des modèles de régression de type GLM.

Ce support a pour objectif de rappeler de manière synthétique le cadre des modèles linéaires généralisés.

Par ailleurs, l'amélioration des performances informatiques a conduit ces dernières années à un intérêt pour des approches alternatives, non paramétriques ou semi-paramétriques, qui peuvent *a priori* permettre de contourner certaines des limitations du cadre des modèles de régression paramétriques.

Quelques unes de ces approches alternatives, notamment les régressions pénalisées et les techniques d'agrégation de modèles sont ainsi présentées.

Préambule

Les étapes d'une tarification

La réalisation d'un tarif IARD nécessite plusieurs étapes :

- la constitution de la base de données ;

- la distinction des sinistres attritionnels, graves et sériels ;

- le choix des variables tarifaires ;

- la modélisation de l'effet des caractéristiques des individus (représentées par les modalités des variables tarifaires) sur les variables à expliquer (la fréquence et le coût) dans le cadre d'un modèle explicatif de la « charge espérée » ;

- le lissage du tarif brut, qui permet de prendre en compte les contraintes de la politique tarifaire ;

- le passage du tarif pur au tarif technique puis commercial.

Préambule

Les étapes d'une tarification

Il est précisé que dans le cadre de cette présentation on se limitera à l'analyse des sinistres hors « graves », « sériels » et « sans suite » et on se concentrera sur le lien entre les caractéristiques d'un individu et son risque.

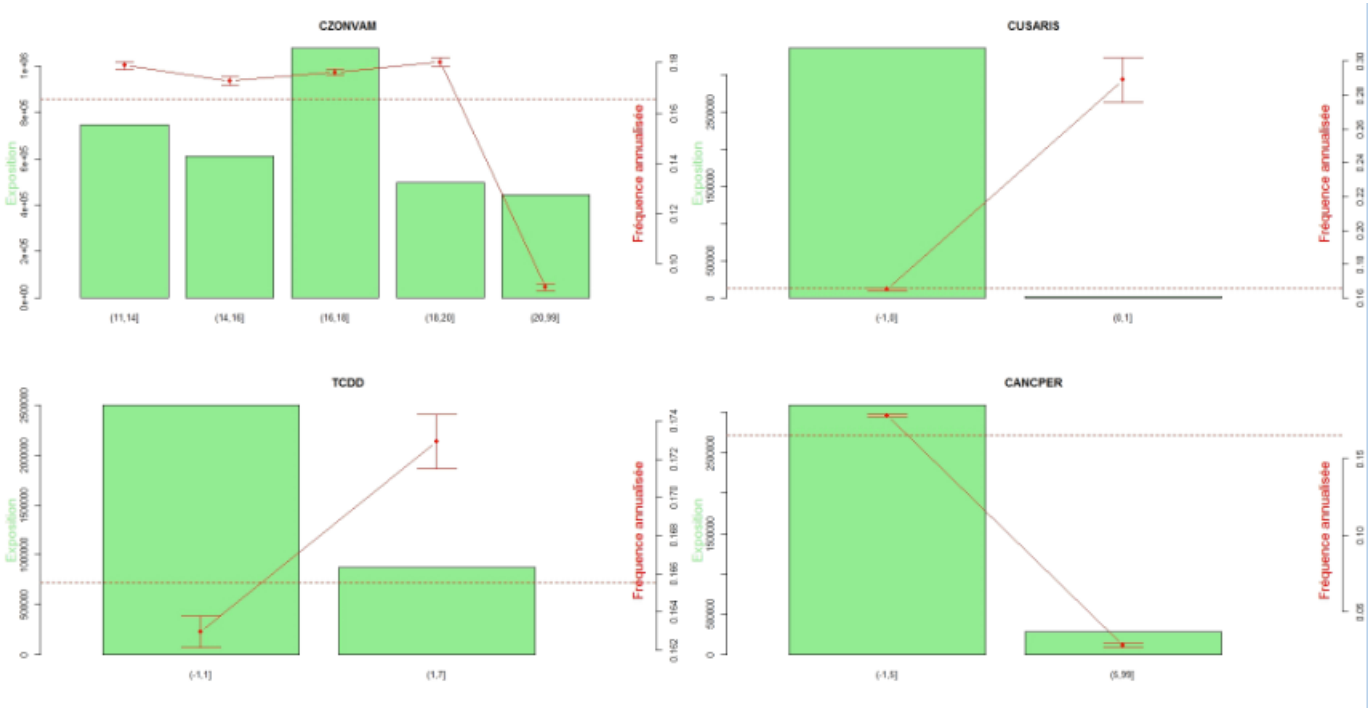
Du fait de ces restrictions, l'impact de la réassurance (non proportionnelle) n'est pas abordé.

Préambule

Impact des caractéristiques de l'assuré sur la fréquence ou le coût

Il s'agit de rendre compte des effets que l'on constate par de simples statistiques descriptives :

La modélisation est indispensable pour régulariser les estimateurs empiriques que l'on peut calculer dans chaque « case » d'une segmentation *a priori*.



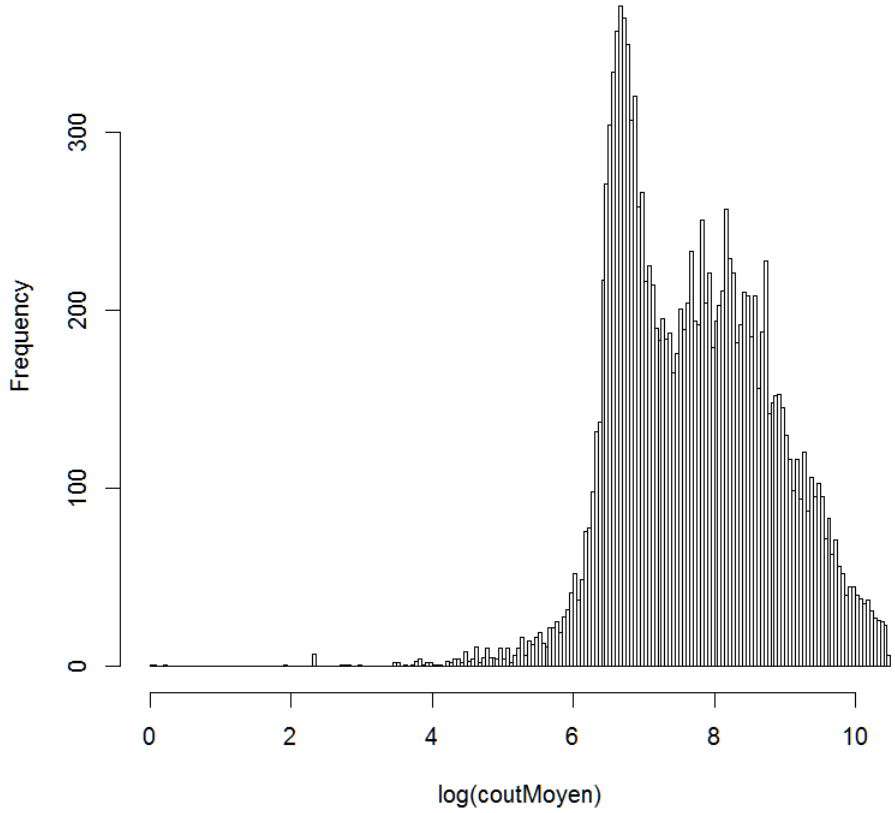
NB : attention aux effets d'hétérogénéité

Préambule

Impact des caractéristiques de l'assuré sur la fréquence ou le coût

On peut aussi par ailleurs observer que la distribution du logarithme du coût moyen n'a pas de forme simple.

L'allure de cette distribution met en évidence l'hétérogénéité sous-jacente et légitime le recours à une décomposition en fonction de variables explicatives.



Préambule

Impact des caractéristiques de l'assuré sur la fréquence ou le coût

Dans la mesure où l'on différencie les risques puis les tarifs, ceux-ci ont tendance à être de plus en plus individualisés.

Mais la mesure du risque n'est qu'une appréciation de l'espérance de coût attachée à une personne soumise à des facteurs de risques. Ainsi, une différenciation très poussée conduit à des tarifs quasiment individuels et non mutualisés. Dans ces conditions, l'analyse *ex ante* du risque devient de plus en plus personnalisée et une erreur d'estimation croissante risque d'apparaître en l'absence de base statistique suffisante.

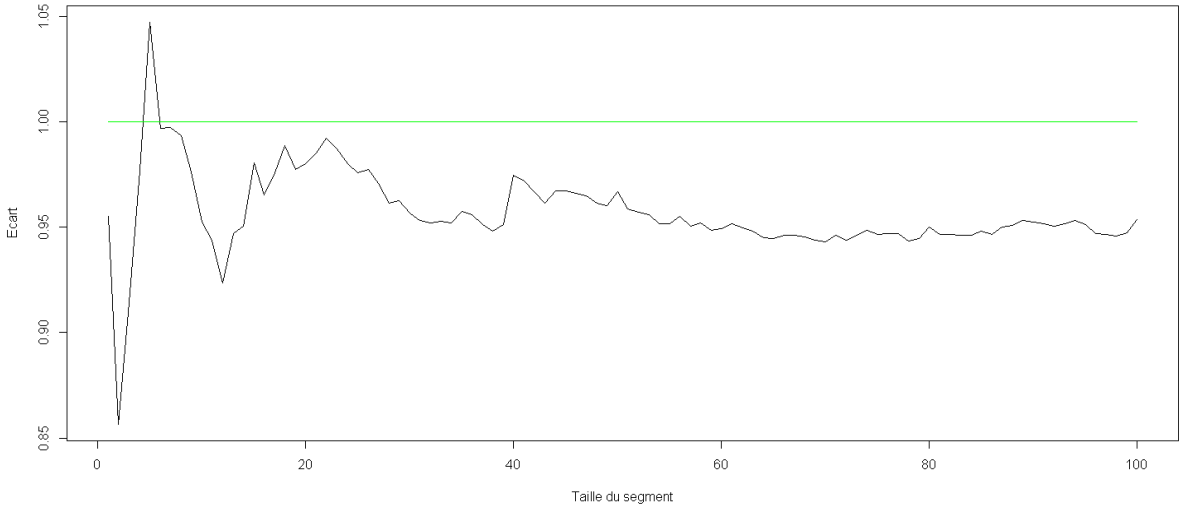
En effet, tant que le tarif (au travers de la fréquence et du coût) est construit sur la base de méthodes statistiques, il faut disposer d'un volume de données suffisant pour que la loi des grands nombres s'applique et que la volatilité de l'estimateur soit faible.

Préambule

Impact des caractéristiques de l'assuré sur la fréquence ou le coût

Si l'on suppose à titre d'exemple que l'on doive estimer un coût moyen de 1000 € entaché d'une incertitude log-normale, l'estimation de ce coût moyen est sensiblement biaisée pour les échantillons de petite taille, comme l'illustre l'exemple suivant :

Au-delà de cette mesure « individuelle », l'addition de ces erreurs d'estimation au niveau de chaque segment tarifaire augmente la volatilité globale des grandeurs estimées (coût ou fréquence).



Voir LEROY et PLANCHET [2016].

Préambule

Le cadre usuel de tarification

En pratique la tarification IARD est souvent effectuée dans le cadre très général des modèles fréquence-coût :

$$S = \sum_{i=1}^{N} C_i + I_G \times G$$

avec N le nombre de sinistres (souvent supposé suivre une loi de Poisson), C le coût unitaire d'un sinistre (en général gamma ou log-normal), I_G l'indicatrice de survenance d'un sinistre grave et G le coût d'un sinistre grave (par exemple de type Pareto).

Préambule

Le cadre usuel de tarification

Sous réserve de l'indépendance de la fréquence et des coûts, la prime pure à l'intérieur d'une classe de risque est de la forme :

$$E\left[S\middle|X\right]=E\left[N-I_G\middle|X\right]\times E\left[C\middle|X\right]+P\left(I_G=1\middle|X\right)\times E\left(G\middle|X\right)$$

On se ramène ainsi à modéliser l'espérance conditionnelle du nombre de sinistres et l'espérance conditionnelle du coût unitaire.

Il s'agit donc d'estimer des espérances conditionnelles, ce qui est le cadre général des modèles de régression, et plus particulièrement des modèles de régression non linéaires (GLM).

Préambule

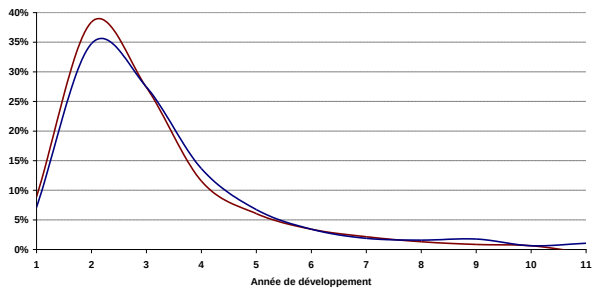
Identification des sinistres graves et sériels

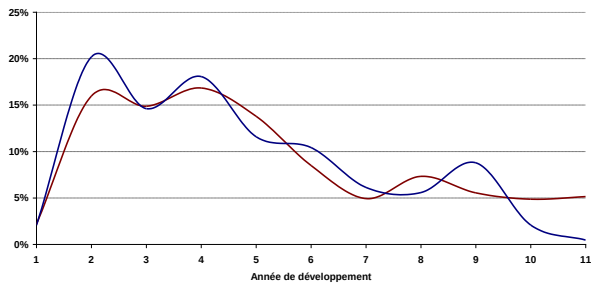
L'identification des sinistres sériels s'appuie sur la mise en relation du sinistre avec un événement, en général codé dans la base de données.

Pour les sinistres graves, il s'agit de déterminer le seuil de gravité pertinent. Pour cela on peut considérer différents critères :

- un sinistre grave étant rare doit être mutualisé sur un ensemble plus large et donc la segmentation tarifaire est *a priori* plus grossière ;

- le comportement du sinistre en termes de déroulement peut aussi être considéré pour les branches longues, par exemple :





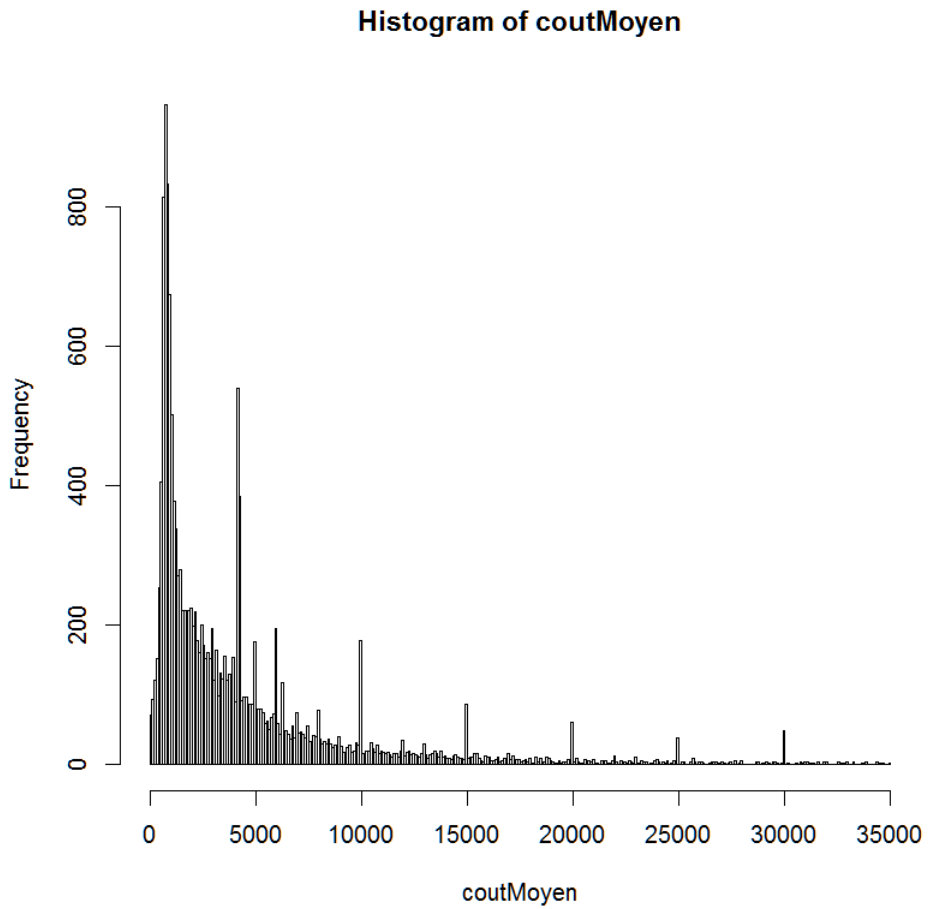
Préambule

Remarque sur le coût moyen

Dans les branches longues, on peut devoir traiter spécifiquement la prise en compte de forfaits à l'ouverture qui induisent des discontinuités dans la distribution des coûts :

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    1.0   956.1  2491.0  4381.0  5212.0 34960.0
```

La distribution empirique des coûts fait apparaître des masses sur les montants entiers en K€.



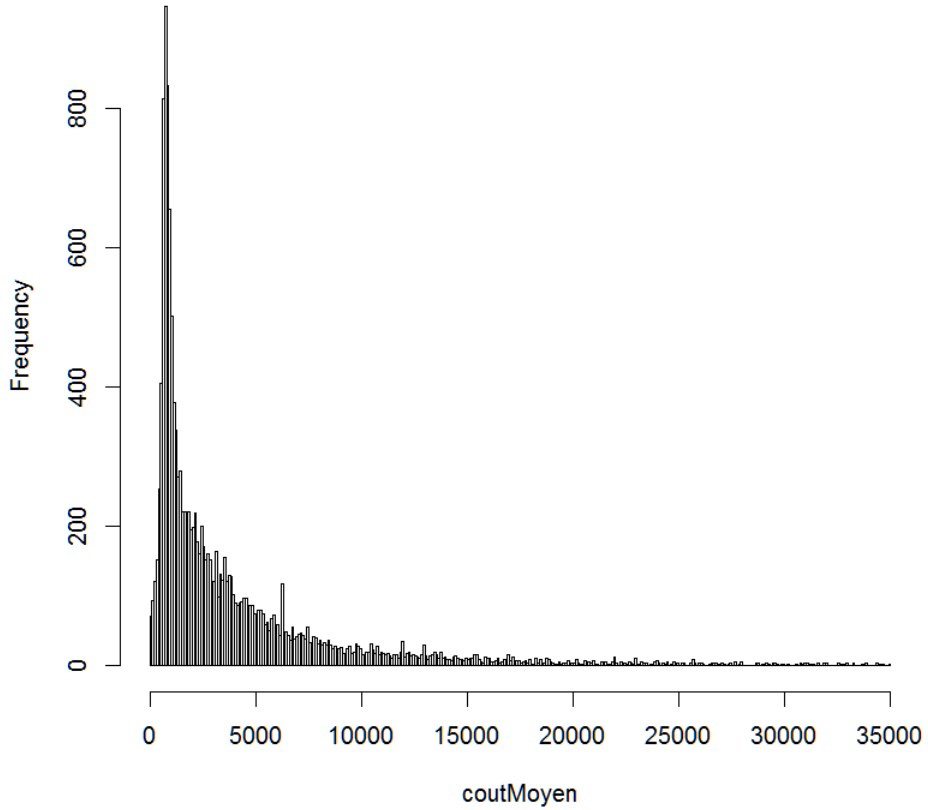
Préambule

Ces montants forfaitaires à l'ouverture doivent être exclus de l'étude.

Les caractéristiques des lignes restantes sont les suivantes :

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 1.0   895.9  2074.0  4059.0  4953.0 34960.0
```

On peut noter sur l'exemple présenté que la suppression des forfaits d'ouverture fait baisser le coût moyen d'environ 8,5 %, ce qui laisse penser que les forfaits d'ouverture sont (trop ?) prudents.



Sommaire

1. **Les modèles GLM**
2. Les régressions pénalisées
3. Agrégation de modèles
4. Points divers

1. Les modèles GLM

Les modèles linéaires généralisés (*Generalized Linear Models*, GLM) constituent le cadre de référence pour modéliser l'effet des variables de segmentation sur le tarif.

Les GLM ont fait leur apparition dans NELDER et WEDDERBURN [1972]. Ils sont adaptés à de nombreuses problématiques et sont d'utilisation courante dans le domaine de la statistique et de l'actuariat (*cf*. DENUIT et CHARPENTIER [2005]). Les GLM permettent de ne pas imposer le caractère normal de la variable à expliquer Y et la relation entre les variables et l'espérance conditionnelle n'est pas nécessairement linéaire :

Modèles linéaires	Modèles linéaires généralisés
• variables aléatoires $Y_1, ..., Y_n$	• variables aléatoires $Y_1, ..., Y_n$
• mutuellement indépendant	• mutuellement indépendant
• $Y_i \rightarrow Normale$	• la distribution de Y_i n'est pas nécessairement normale mais doit être dans la famille exponentielle
• $\mathbb{E}[Y_i] = \mu_i = x_i'\beta$ et $\mathbb{V}[Y_i] = \sigma^2$	• $g(\mathbb{E}[Y_i]) = g(\mu_i) = x_i'\beta$ ou $\mu_i = g^{-1}(x_i'\beta)$

1. Les modèles GLM

Dans le contexte d'un modèle GLM, on considère que pour une variable aléatoire *Y*, qui correspond à la variable à expliquer, il existe une relation de la forme suivante :

$$g\left(\mathbf{E}\left[Y \middle| x_1, \ldots, x_p\right]\right) = \sum_{k=1}^{p} \beta_k x_k$$

entre avec *p* variables explicatives X_i (i = 1,…, p) et l'espérance conditionnelle de la variable à expliquer. La fonction *g* (strictement monotone et dérivable) est appelée fonction de lien du modèle. Elle détermine la relation entre le prédicteur linéaire et l'espérance de la variable expliquée.

Par exemple le choix (classique) $g(u)$ = **ln**(u) conduit au modèle multiplicatif suivant :

$$\mathbf{E}\left[Y \middle| x\right] = \mathbf{exp}\left(\sum_{k=1}^{p} \beta_k x_k\right) = \mathbf{exp}\left(\beta' x\right)$$

Il reste à spécifier la loi de la variable *Y*. On impose à *Y* d'appartenir à la famille dite exponentielle, pour laquelle la densité s'écrit :

$$f_{\theta,\varphi}(y) = \exp\left(\frac{y\times\theta - b(\theta)}{\varphi} + c(y,\varphi)\right)$$

avec *b* une fonction définie sur R deux fois dérivable et de dérivée première injective et *c* une fonction définie sur R^2. De nombreuses distributions classiques appartiennent à cette famille. On a en particulier :

$$\mathbf{E}(Y) = b'(\theta) \qquad \mathbf{V}(Y) = b''(\theta)\varphi = b''\left(b'^{-1}\left(\mathbf{E}[Y]\right)\right)\varphi = \upsilon\left(\mathbf{E}[Y]\right)\varphi$$

Le lien entre le paramètre et les variables explicatives est donc de la forme :

$$\theta(x) = b'^{-1}\left(E\left(Y|x\right)\right) = b'^{-1}\left(g^{-1}\left(\sum_{k=1}^{p}\beta_k x_k\right)\right)$$

On peut noter que le paramètre ϕ est un « paramètre de nuisance », l'espérance conditionnelle ne dépendant que de θ.

Exemples : lois Poisson et Gamma

Loi de probabilité	$\mathbf{Pr}(Y=y)=\exp\left(y\ln(\lambda)-\lambda+c(y)\right)$
θ	$\ln\lambda$
Φ	1
$b(\theta)$	$\exp(\theta)$
$\mathbf{E}[Y]$	λ
Fonction variance	$v(\lambda)=\lambda$

Loi de probabilité	$f(y)=\exp\left(\frac{-\frac{\mu}{\upsilon^2}y+\ln\frac{\mu}{\upsilon^2}}{\frac{1}{\upsilon}}+c(y,\upsilon)\right)$
θ	$-\frac{\mu}{\upsilon^2}$
Φ	$\frac{1}{\upsilon}$
$b(\theta)$	$\ln\frac{-1}{\theta}$
$\mathbf{E}[Y]$	μ
Fonction variance	$v(\mu)=\mu^2$

Autres exemples

Distribution de Y_i	θ_i	ϕ	$a_i(\phi)$	$b(\theta_i)$	$c(y_i, \phi)$
Normale$(\mu_i\,;\sigma^2)$	μ_i	σ^2	ϕ	$\frac{\theta^2}{2}$	$-\frac{1}{2}\left\{\frac{y^2}{\sigma^2} + log(2\pi\sigma^2)\right\}$
Poisson(μ_i)	$\log(\mu_i)$	1	ϕ	$\exp(\theta_i)$	$-\log y!$
Binomiale$\frac{1}{m_i}(m_i\,;\mu_i)$	$\log\left(\frac{\mu_i}{1-\mu_i}\right)$	$\frac{1}{\mu_i}$	ϕ	$\log(1+\exp\theta_i)$	$\log\binom{m_i}{m_i y_i}$
Gamma$(\mu_i\,;\alpha)$	$\frac{-1}{\mu_i}$	α^{-1}	ϕ	$-\log(-\theta)$	$\alpha\log(\alpha y) - \log y - \log\Gamma(\alpha)$
Inverse Gaussienne$(\mu_i\,;\sigma^2)$	$\frac{-1}{2\mu_i^2}$	σ^2	ϕ	$-(-2\theta)^{1/2}$	$-\frac{1}{2}\left\{\log(2\pi\phi y^3) + \frac{1}{\phi y}\right\}$

Remarque : des travaux spécifiques proposent des modèles prenant en compte les phénomènes de sous-déclaration des petits sinistres dans la fréquence. Les modèles « à inflation de zéros » font ainsi l'objet d'applications directe en tarification non-vie (cf. VASECHKO et *al*. [2009] et la section 5 de cette présentation).

En pratique on utilise souvent :

- la fonction de lien *log*, qui permet d'avoir un tarif multiplicatif ;
- la loi de Poisson ou la loi binomiale négative pour la fréquence ;
- la loi gamma ou normale pour le coût.

Remarques :

- la loi binomiale négative est le nombre d'échecs avant l'obtention de n succès dans une expérience où la probabilité de succès est p. Elle peut aussi s'interpréter comme un mélange de lois de Poisson lorsque le paramètre λ suit une loi gamma, ce qui s'interprète comme la prise en compte d'une hétérogénéité non observable.

- la loi globale de Y n'est pas en général de même nature que la loi sous-jacente (gamma, normale, *etc*.) mais est un mélange de ce type de lois.

Utilisation d'une variable *offset* dans un modèle de régression

Dans le cadre d'une régression pour expliquer un nombre de sinistres *N* avec un modèle poissonnien et une fonction de lien logarithme, on a :

$$\mathbf{E}\left[N \middle| x \right] = \mathbf{exp}\left(\sum_{k=1}^{p} \beta_k x_k \right) = \mathbf{exp}\left(\beta' x \right)$$

Si on veut tenir compte de l'exposition au risque *d*, on sait que l'espérance λ de la loi de Poisson devient $\lambda\, d$. La régression se réécrit alors :

$$\mathbf{E}\left[N \middle| x, d \right] = d \times \mathbf{exp}\left(\sum_{k=1}^{p} \beta_k x_k \right) = \mathbf{exp}\left(\beta' x + \mathbf{ln}\left(d \right) \right)$$

Tout se passe donc comme si l'on ajoutait une variable explicative pour laquelle le coefficient β est connu (ici égal à 1) et ne doit donc pas être estimé.

La variable $x_{p+1} = \mathbf{ln}\left(d \right)$ s'appelle une variable *offset*.

Utilisation d'une variable *offset* dans un modèle de régression

Cette idée peut être exploitée pour intégrer des variables de tarification avec des coefficients contraints (*i.e.* estimés par ailleurs). Si par exemple on veut intégrer dans le modèle les contraintes suivantes :

- zonier à : 1 = -5 %, 2 = 0 % et 3 = +5 % ;
- effectif à : 0 = -5 % et >0 = 0 %.

On définit alors la variable *t* par :

$$x_1 = \begin{cases} \ln(0,95) & ZONIER = 1 \\ 0 & ZONIER = 2 \\ \ln(1,05) & ZONIER = 3 \end{cases} \qquad t = \begin{cases} x_1 + \ln(0,95) & EFFECTIF = 0 \\ x_1 & EFFECTIF > 0 \end{cases}$$

L'introduction de *t* en variable *offset* permet d'estimer les coefficients des autres variables en tenant compte de ces contraintes tarifaires.

Utilisation d'une variable *offset* dans un modèle de régression

Cette approche est notamment utilisée lorsque l'on procède à un lissage des coefficients d'une variable issue de la régression : la prise en compte de l'impact du lissage sur les autres coefficients conduit à refaire une régression en utilisant la variable lissée comme variable *offset*.

Elle permet également de justifier la démarche de construction d'un zonier en effectuant une première régression à l'aide des variables tarifaire hors zone géographique puis d'ajouter cette information *ex-post* pour augmenter la part de variance expliquée.

La construction du zonier est une problématique à part entière qui peut mobiliser des outils mathématiques élaborés (cf. BOSKOV et VERRALL [1994] dont le modèle est utilisé dans MATHIS [2009]).

Estimation d'un modèle GLM

La log-vraisemblance (pondérée) d'une observation est de la forme :

$$l_{\theta,\varphi}(y) = \frac{y\times\theta - b(\theta)}{\varphi/w} + c(y,\varphi)$$

L'estimation par maximum de vraisemblance conduit à résoudre les équations normales

$$\frac{\partial}{\partial\beta_j}\sum_{i=1}^{n} l_{\theta_i,\varphi,w_i}(y_i) = 0$$

dont on peut montrer après quelques calculs qu'elles peuvent se mettre sous la forme :

$$\sum_{i=1}^{n} \frac{\tilde{w}_i \times x_{ij}}{\varphi} g'(\mu_i)(y_i - \mu_i) = 0$$

avec $\tilde{w}_i = \frac{w_i}{V(\mu_i)}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2$

Validation d'un modèle GLM – Déviance

Pour mesurer la qualité de l'ajustement d'un modèle GLM on utilise souvent la déviance, égale par définition à :

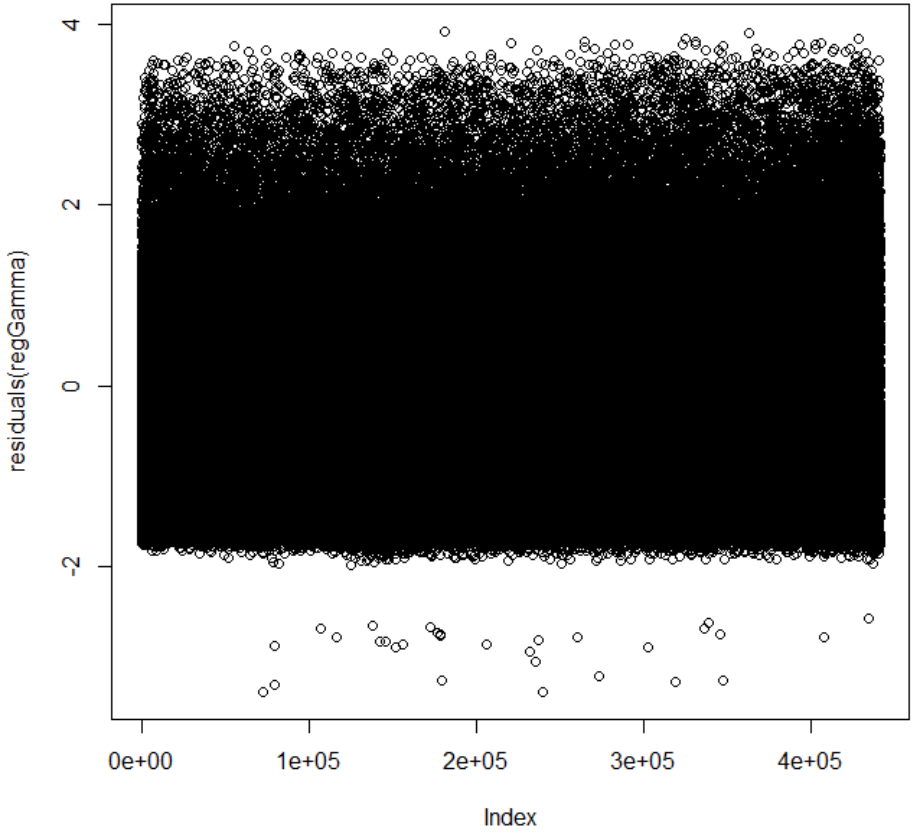
$$D = 2 \times \left(\ln L\left(Y \middle| Y\right) - \ln L\left(\hat{\mu} \middle| Y\right) \right)$$

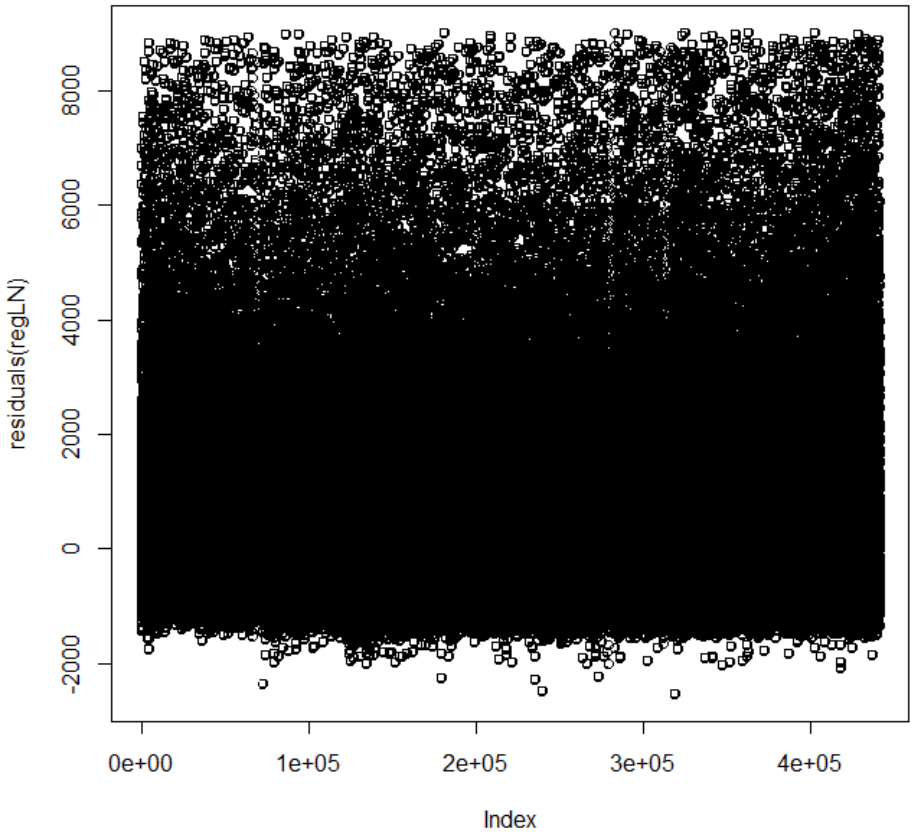
D est positif et « petit » pour un modèle de bonne qualité. Cette statistique suit asymptotiquement, du fait de résultats généraux sur les rapports de vraisemblance, une loi du Khi-2 à $n - p$ - 1 degrés de liberté (son espérance est donc $n - p - 1$).

Cet indicateur global est en pratique complété par une analyse observation par observation ; cette analyse se base souvent sur l'analyse des résidus.

Validation d'un modèle GLM – Résidus

Les graphiques ci-dessous mettent par exemple en évidence que le modèle gamma (à gauche) est mieux adapté que le modèle LN (à droite) :





Validation d'un modèle GLM – Résidus

Les résidus peuvent être calculés de différentes manières. Les deux principales sont les résidus de Pearson et les résidus de déviance :

- Résidus de Pearson $r_i^P = \sqrt{\omega_i}\,\frac{y_i - \mu_i}{\sqrt{V(\mu_i)}}$

- Résidus de déviance $r_i^P = \varepsilon(y_i - \mu_i)\sqrt{d_i}$

On peut noter que la somme des carrés des résidus est dans les deux cas, asymptotiquement, un Khi-2 à $n - p - 1$ degrés de liberté.

Choix d'un modèle GLM – Critères AIC et BIC

La comparaison de 2 modèles nécessite de tenir compte de la complexité de chaque modèle. Les critères AIC et BIC pénalisent la log-vraisemblance du modèle avec le nombre de paramètres :

$$AIC = 2 \times \left(p - \ln L\left(\hat{\mu} \middle| Y \right) \right) \qquad BIC = -2 \times \ln L\left(\hat{\mu} \middle| Y \right) + p \times \ln\left(n \right)$$

L'AIC est asymptotiquement optimal lorsque l'on souhaite sélectionner le modèle avec l'erreur quadratique moyenne, si l'on fait l'hypothèse que le modèle générant les données n'est pas parmi les candidats, ce qui est en fait presque toujours le cas en pratique (cf. YANG [2005]). Ce n'est pas le cas du BIC. Yang montre également que la vitesse de convergence de l'AIC vers l'optimum est, dans un certain sens, la meilleure possible.

https://fr.wikipedia.org/wiki/Crit%C3%A8re_d'information_d'Akaike

Choix d'un modèle GLM – Critères AIC et BIC

À partir de ces critères, on peut imaginer des processus de sélection des variables à prendre en compte dans le modèle :

- on part du modèle avec seulement la constante et on ajoute la variable qui conduit à la plus forte baisse de l'AIC ou du BIC à chaque étape ; on s'arrête lorsque l'indicateur ne baisse plus (sélection ascendante) ;

- on part du modèle avec toutes les variables et on effectue des suppressions pas à pas (sélection descendante) ;

- un mélange des deux techniques ci-dessus (sélection ascendante avec possibilité de suppression d'une variable déjà sélectionnée).

Choix d'un modèle GLM – Critères AIC et BIC

```
# sélection de modèle ascendante
logit <- glm(Cible~1,data=train,family=binomial(link = "logit"))
summary(logit)
selection <- step(logit,direction="forward",trace=TRUE,k = log(nrow(train)),
scope=list(upper=formule))
summary(selection)
```

```
glm(formula = Cible ~ Comptes + Duree_credit + Garanties + Autres_credits +
    Age, family = binomial(link = "logit"), data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8336  -0.7597  -0.4377   0.8726   2.4797

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                        -2.3030     0.5980  -3.851 0.000118 ***
ComptesCC < 0 euros                 0.5610     0.2345   2.393 0.016714 *
ComptesCC > 200 euros              -0.4430     0.4010  -1.105 0.269354
ComptesPas de compte               -1.6734     0.2638  -6.344 2.24e-10 ***
Duree_credit(15,36]                 1.0028     0.2161   4.641 3.46e-06 ***
Duree_credit(36,Inf]                1.4934     0.3359   4.446 8.73e-06 ***
GarantiesSans garant                1.5787     0.5416   2.915 0.003559 **
Autres_creditsCrédits extérieurs    0.6355     0.2344   2.711 0.006713 **
Age(25,Inf]                        -0.6296     0.2406  -2.617 0.008871 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Choix d'un modèle GLM – Critères AIC et BIC

L'approche pas–à–pas ne permet pas de trouver le meilleur modèle parmi les 2^p modèles possibles.

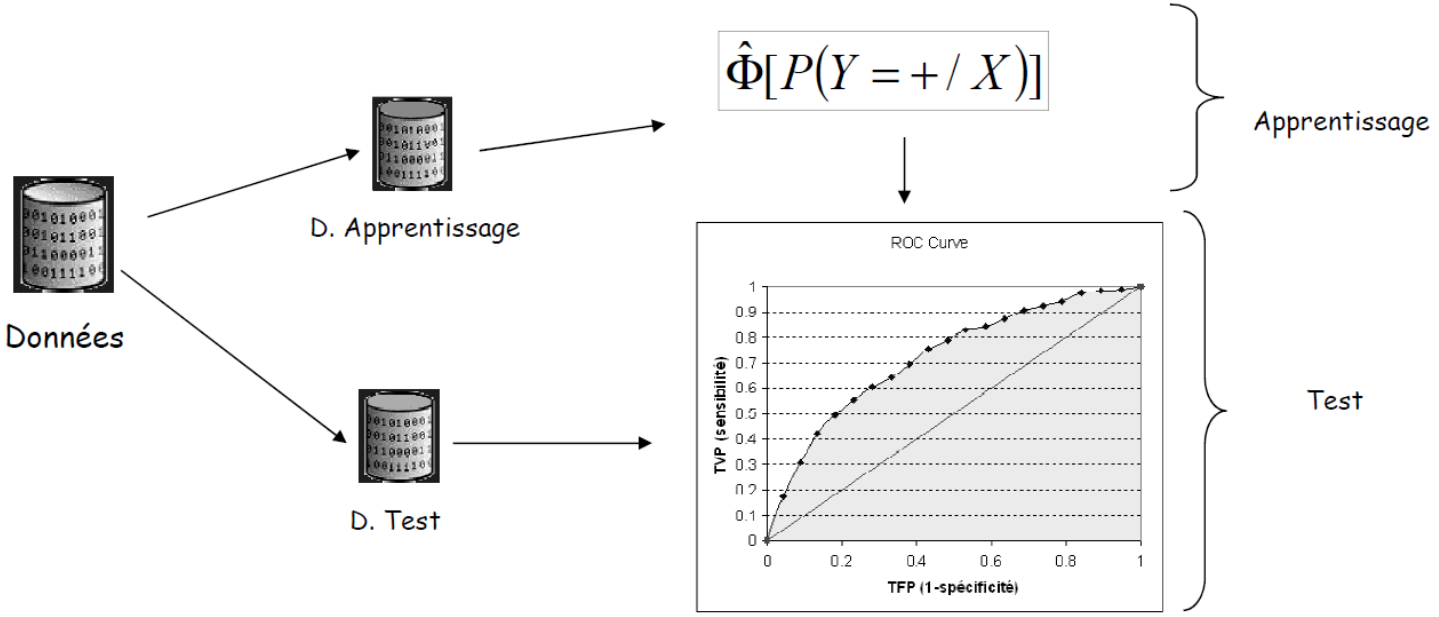
Dans certains cas (régression linéaire ou régression logistique par exemple), il est possible d'utiliser une approche « force brute » en l'optimisant (cf. FURNIVAL et WILSON [1974]), mais pour un modèle GLM général, il n'est pas possible d'examiner toutes les possibilités.

On est donc conduit à examiner d'autres techniques, par exemple en utilisant des arbres CART sur les variables pour expliquer l'indicateur de performance (erreur d'estimation ou AUC).

On s'intéresse dans la suite à des approches différentes utilisables en grande dimension.

Cas particulier d'un modèle logistique – Courbe ROC

Dans le cas particulier où la valeur estimée est une probabilité (typiquement dans un modèle logistique), la courbe ROC est un outil supplémentaire de mesure de la qualité du modèle :



http://eric.univ-lyon2.fr/~ricco/cours/slides/roc_curve.pdf

Cas particulier d'un modèle logistique – Courbe ROC

Dans le contexte de l'affectation à une classe dans un modèle binaire, il est naturel de s'intéresser au taux de « vrais positifs » (TVP, prédiction « + » pour une observation « + ») et au taux de « faux positifs » (TFP, prédiction « + » pour une observation « - »).

En se basant sur la règle de classement qui affecte « + » lorsque $P\left(Y=+\middle|x\right)\geq\alpha$ et en faisant varier le seuil α, , on recalcule les indicateurs ci-dessus et on peut représenter les points

$$\left(TFP\left(\alpha\right),TVP\left(\alpha\right)\right)$$

On obtient une courbe telle que l'aire sous la courbe (AUC) indique la probabilité pour que la fonction score place un positif devant un négatif. Dans le meilleur des cas, AUC=1. L'AUC est directement lié à la statistique de Mann-Whitney

$$S=\sum_{i/y_i=+}r_i-\frac{n_+\left(n_++1\right)}{2}\quad\Longrightarrow\quad AUC=\frac{S}{n_+\times n_-}$$

Limites de l'approche classique

Les modèles GLM classiques nécessitent de disposer d'un échantillon de taille n relativement importante et, surtout, que le nombre p de variables de variables explicatives ne soit pas « trop » grand. Il est par exemple évident que si $p>n$, il n'est plus possible de calculer l'estimateur MCO d'une régression linéaire.

De nombreuses techniques ont été développées pour faire face à ce problème. Dans la suite de cette présentation, on s'intéresse plus particulièrement à deux approches :

- l'agrégation de modèles ;
- la pénalisation de la régression.

Le principe de la régression PLS est également décrit.

Les exemples présentés ci-après sont adaptés de TUFFERY [2015].

De la statistique vers l'apprentissage

Comme nous le verrons :

- si, dans une démarche statistique « classique », on construit un estimateur sur un jeu de données unique et une théorie asymptotique permet de juger de sa qualité et de construire des intervalles de confiance,

- les évolutions des modèles proposées s'inscrivent dans la logique de la théorie de l'apprentissage : les données sont en général coupées en deux, avec des échantillons d'apprentissage et de validation et la qualité n'est plus jugée via des critères asymptotiques, mais à l'aune de l'adéquation à l'échantillon de validation.

On va chercher non plus un modèle, mais des modèles pour construire une synthèse plutôt qu'un modèle unique et l'adéquation n'est plus envisagée de la même manière.

De la statistique vers l'apprentissage : l'erreur d'estimation

Dans un modèle paramétrique estimé par maximum de vraisemblance, on dispose de résultats de convergence qui fournissent une information sur la loi de l'écart entre l'estimateur et la vraie valeur

$$\sqrt{n}\left(\hat{\beta}_n - \beta\right) \xrightarrow[n\to\infty]{Loi} N\left(0, I\left(\beta\right)^{-1}\right)$$

et on dispose d'un estimateur de la variance asymptotique

$$\hat{V}\left(\hat{\beta}\right) = -\left[\frac{\partial^2 \ln L\left(y \middle| x; \hat{\beta}\right)}{\partial\beta\partial\beta'}\right]^{-1}$$

On dispose donc d'une mesure de l'erreur d'estimation. L'analyse des résidus complète cette analyse avec une indication sur la qualité de l'ajustement.

De la statistique vers l'apprentissage : l'erreur de prédiction

Dès que l'on sort d'un cadre de type « maximum de vraisemblance », mesurer la qualité de l'adéquation du modèle nécessite, pour éviter le sur-apprentissage, de distinguer un échantillon d'ajustement et un échantillon de validation. La qualité de l'ajustement est alors mesurée sur l'échantillon de validation avec un critère quadratique :

$$e(\hat{\mu}) = \sum_{i \in valid} \left(Y_i - \hat{\mu}_i\right)^2$$

Ce critère est « naturel » dans le cadre de l'estimation d'une espérance conditionnelle.

D'autres critères peuvent parfois être utilisés en fonction du contexte (l'AUC par exemple pour une régression logistique).

De la statistique vers l'apprentissage : l'erreur de prédiction

Par ailleurs, dans le cadre d'un modèle non paramétrique, la vitesse de convergence de l'estimateur n'est plus en « racine de n » ; on peut montrer que la vitesse de convergence optimale d'un estimateur non paramétrique d'une fonction *k* fois dérivables en dimension *p* est de l'ordre de

$$n^{-\frac{k}{(2k+p)}}$$

En particulier, lorsque *d* est grand, la vitesse de convergence est fortement pénalisée.

De la statistique vers l'apprentissage : la validation croisée (*k-fold*)

Le découpage de l'échantillon en apprentissage et validation induit une perte d'un certain volume de données pour calculer les estimateurs.

On peut alors couper l'échantillon aléatoirement en *k* sous-échantillons de taille égale, puis :

- prendre l'un de ces échantillons pour échantillon de validation, le reste pour l'apprentissage ;

- refaire tourner la méthode en changeant les rôles, *k* fois pour que chaque sous-échantillon ait été utilisé une fois en validation ;

- Moyenner les *k* estimateurs pour obtenir l'estimateur final ;

- Moyenner les erreurs sur les échantillons de validation.

Sommaire

1. Les modèles GLM
2. **Les régressions pénalisées**
3. Agrégation de modèles
4. Points divers

2. Les modèles de régression pénalisés

Le principe de la pénalisation d'une régression est de contraindre les coefficients à ce pas être « trop grands », qui se traduit par l'introduction d'une contrainte de la forme

$$\sum_{j=1}^{p}\left|\beta_j\right|^{\delta} \leq C$$

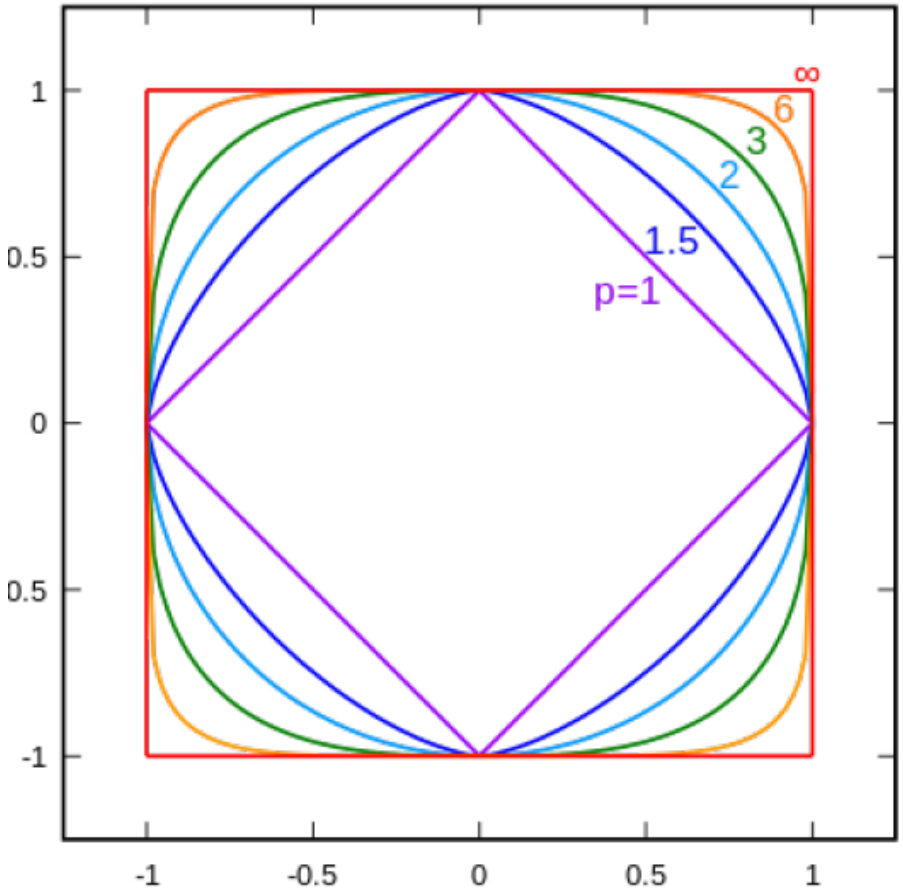
ou, de manière équivalente, à pénaliser la log-vraisemblance du modèle :

$$l_{\theta,\varphi}(y) = \frac{y\times\theta - b(\theta)}{\varphi/w} + c(y,\varphi) + \lambda\sum_{j=1}^{p}\left|\beta_j\right|^{\delta}$$

Lorsque le modèle intègre une constante, elle est exclue de la contrainte. Les choix les plus classiques pour δ sont 0 (pénalisation en fonction du nombre de coefficients), 1 (LASSO) et 2 (Ridge).

2. Les modèles de régression pénalisés

La forme de la région autorisée dépend de la valeur du paramètre δ :



Source : https://fr.wikipedia.org/wiki/Norme_(math%C3%A9matiques)#/media/File:Vector-p-Norms_qtl1.svg

La régression RIDGE

Dans le cas d'une régression linéaire, la colinéarité induit un problème de conditionnement de la matrice X^TX qui devient difficile à inverser.

L'ajout de la pénalisation résout cette difficulté puisqu'alors on a la solution explicite :

$$\hat{\beta}_{ridge} = \left(X^T X + \lambda I_p \right)^{-1} X^T Y$$

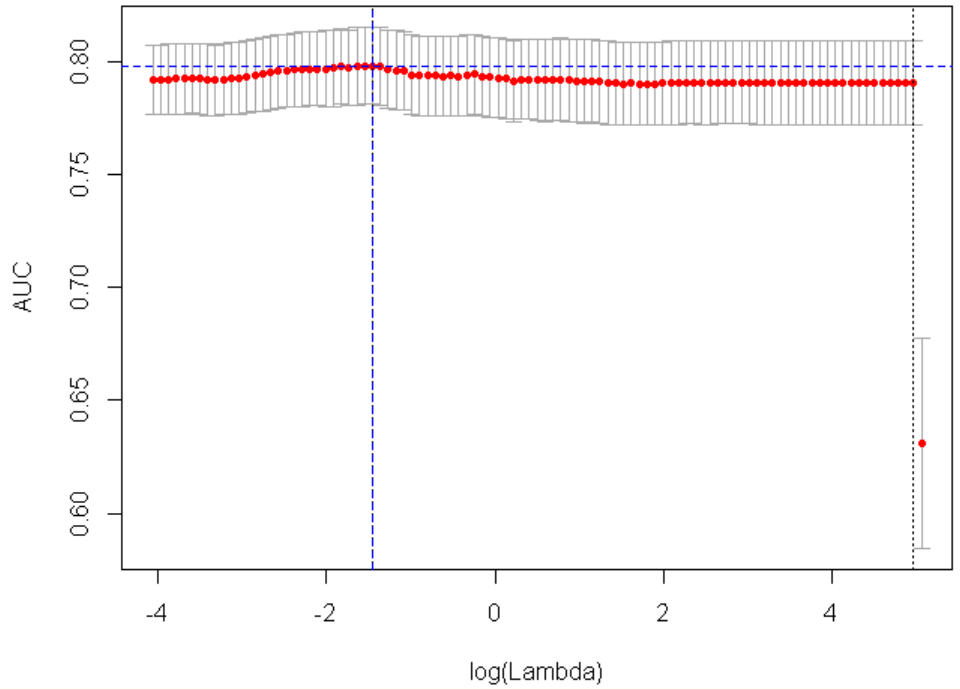
ce qui revient à ajouter λ à toutes les valeurs propres de X^TX. Cette logique se transpose dans le cas d'une régression non linéaire (GLM), au prix d'une résolution numérique du programme d'optimisation.

La pénalisation est un paramètre supplémentaire du modèle qui peut être choisi en effectuant les calculs avec différentes valeurs.

La régression RIDGE

```
cvfit = cv.glmnet(x,y,alpha=0,family = "binomial",type="auc",nlambda=100)
x11()
plot(cvfit)
abline(h=cvfit$cvm[which(cvfit$lambda==cvfit$lambda.min)],col='blue',lty=2)
abline(v=log(cvfit$lambda.min),col='blue',lty=2)
# calcul de la régression pour une plage de valeurs de lambda
listeLambda=seq(cvfit$lambda[1],cvfit$lambda[length(cvfit$lambda)],length=10000)
fit = glmnet(x,y,alpha=0,family = "binomial",lambda=listeLambda,standardize = T)
# affichage des coefficients
x11()
plot(fit,xvar="lambda",label="T")
```

On utilise la validation croisée pour déterminer le niveau de l'AUC en fonction de λ.

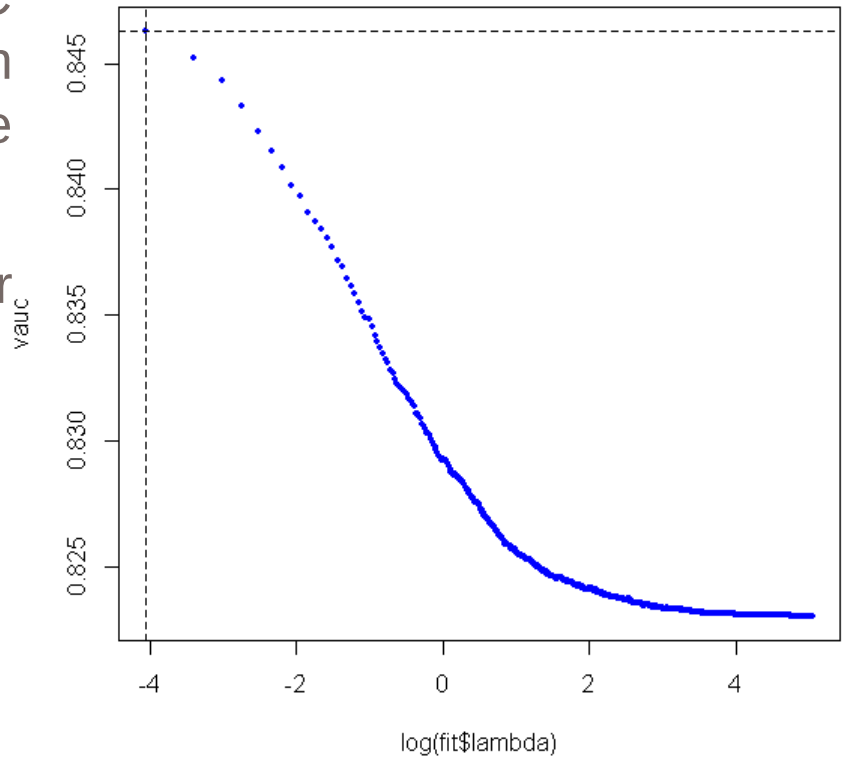


La régression RIDGE

On ne peut pas utiliser l'échantillon d'apprentissage pour estimer directement λ, car cela conduit à un sur-apprentissage

Ici par exemple, cela conduirait à un AUC de presque 0,85 obtenu avec la pénalisation la plus faible, donc avec un modèle logistique simple.

Il faut donc utiliser l'échantillon de test pour trouver le paramètre optimal.

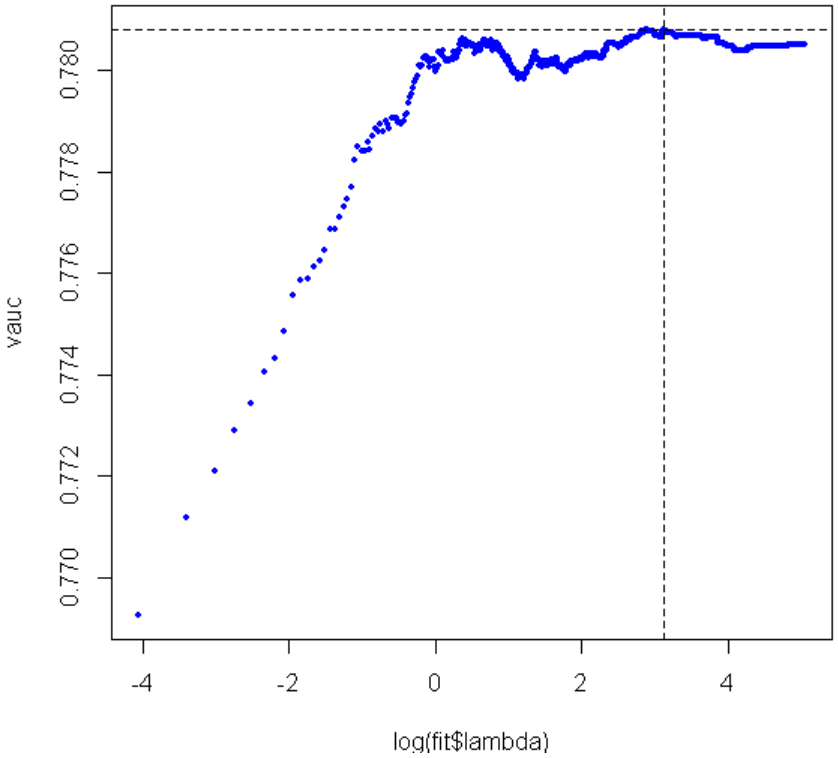


La régression RIDGE

```
#Travail sur le base de tests
xt <- model.matrix( ~ . -1,data=valid[,-which(names(valid)=="Cible")])
yt <- as.numeric(valid[,"Cible"])
ytpred <- predict(fit,newx=xt,type="response")
roc <- function(x) { performance(prediction(ytpred[,x],yt),"auc")@y.values[[1]]
vauc <- Vectorize(roc)(1:length(fit$lambda))
```

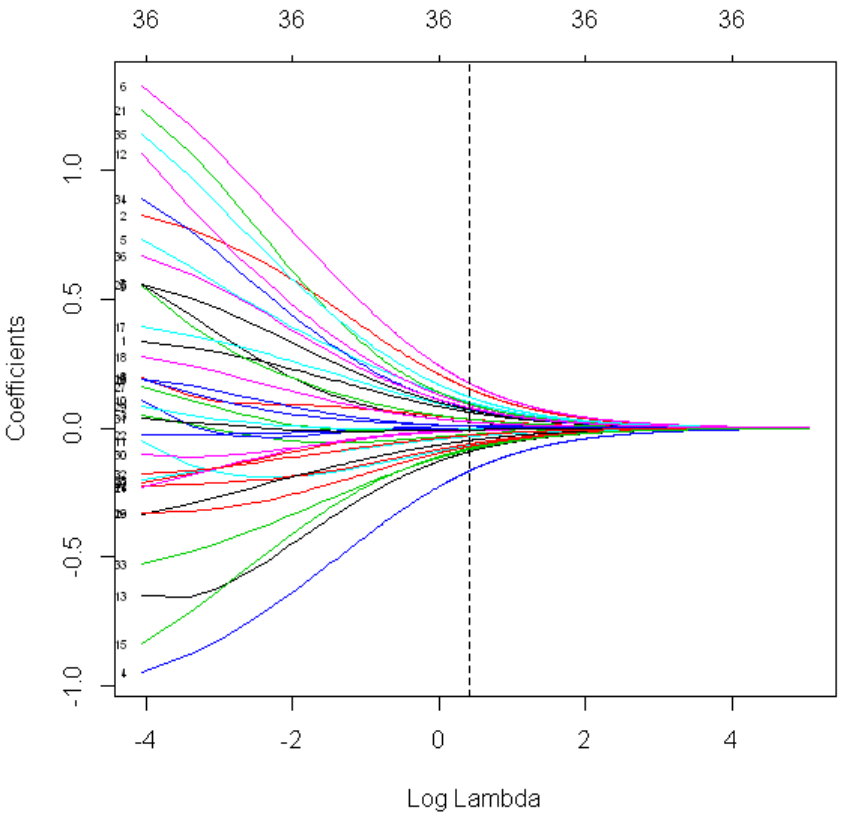
On obtient un AUC maximal de l'ordre de 0,7806 pour une pénalisation de 22,5.

L'AUC est élevé sur une plage assez large de valeurs et on note que l'on obtient un AUC de 0,7805 avec une pénalisation plus faible de l'ordre de 1,5.



La régression RIDGE

À l'issue de ces étapes, on obtient les coefficients ajustés



(Intercept)	-0.918931690
ComptesCC [0-200 euros[	0.062072682
ComptesCC < 0 euros	0.153377179
ComptesCC > 200 euros	-0.024914197
ComptesPas de compte	-0.167161880
Historique_creditImpayé en cours dans autre banque	0.101236172
Historique_creditImpayés passés	0.175797806
Historique_creditPas de crédits ou en cours sans retard	0.023572991
Objet_creditBusiness	0.034750657
Objet_creditEtudes	0.034801912
Objet_creditIntérieur	0.007902700
Objet_creditVidéo - HIFI	-0.067581376
Objet_creditVoiture neuve	0.092225695
Objet_creditVoiture occasion	-0.091908202
Epargne> 500 euros	-0.062047160
EpargneSans épargne	-0.079078381
Anciennete_emploientre 1 et 4 ans	0.007210973
Anciennete_emploiSans emploi ou < 1 an	0.070680961
Taux_effort	0.024511659
Situation_familialeHomme célibataire/marié/veuf	-0.048291916
Situation_familialeHomme divorcé/séparé	-0.008143156
GarantiesSans garant	0.095429234
Anciennete_domicile	-0.001523318
BiensImmobilier	-0.032448223
BiensNon immobilier	-0.007750683
Autres_creditsCrédits extérieurs	0.071933456
Statut_domicilePropriétaire	-0.071237865
Nb_credits	-0.008257262
Type_emploiA172	0.008384147
Type_emploiA173	-0.005113748
Type_emploiA174	-0.010107593
Nb_pers_charge2	-0.006256117
TelephoneA192	-0.028196146
Age(25,Inf]	-0.084595833
Duree_credit(15,36]	0.077876728

Le « fléau de la dimension »

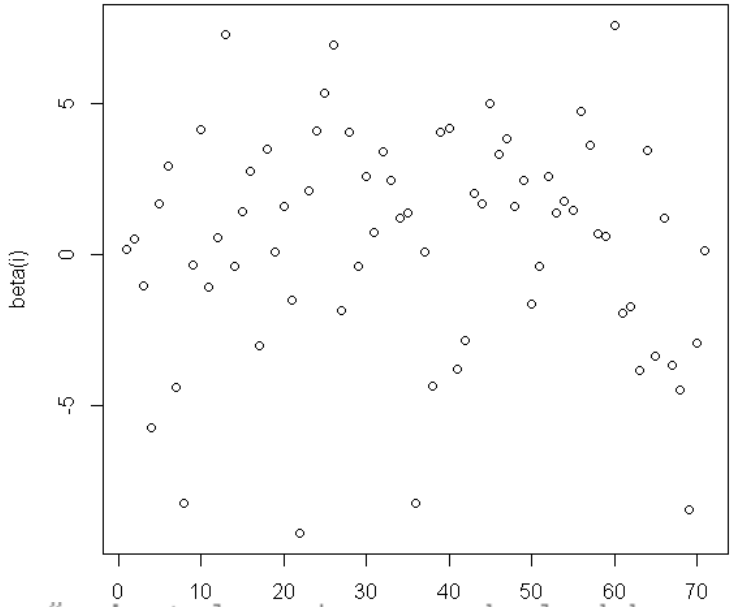
On observe que, si beaucoup de variables explicatives sont significatives, alors on ne parviendra pas à bien estimer les coefficients associés, car leur nombre est proche de n.

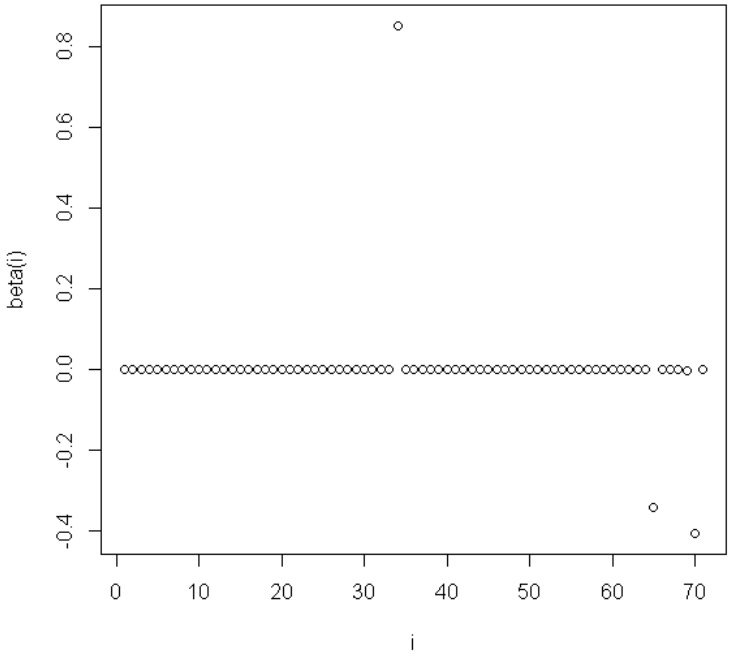
On est alors conduit à formuler une hypothèse de « parcimonie », en supposant que le nombre de coefficients non nuls est égal à k, avec $k << n$. Si on connaissait quels sont les coefficients non nuls, on reviendrait à un modèle en petite dimension.

L'approche par pénalisation LASSO fournit un moyen d'identifier les coefficients non nuls.

La régression LASSO (*Least Absolute Shrinkage and Selection Operator*)

L'idée de la pénalisation LASSO est de contraindre les coefficients « petits » à être nuls pour rendre les autres coefficients plus significatifs

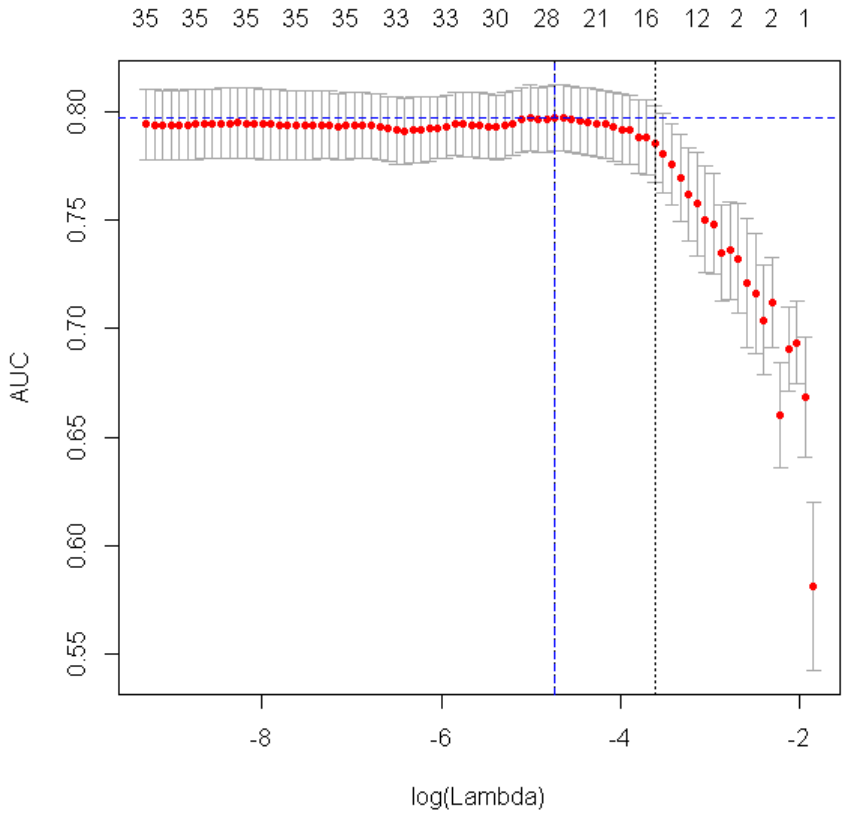





```
pp=9/20; # c'est la puissance de lambda
gamma=2/5;
N=144 # nombre d'observations
lambda=N^{pp} #on définit lambda
#Modele linéaire:
lasso.lm=lqa(Y~X-1,penalty=lasso(lambda=lambda),standardize=TRUE)
coeff.lasso=lasso.lm$coef;
```

La régression LASSO

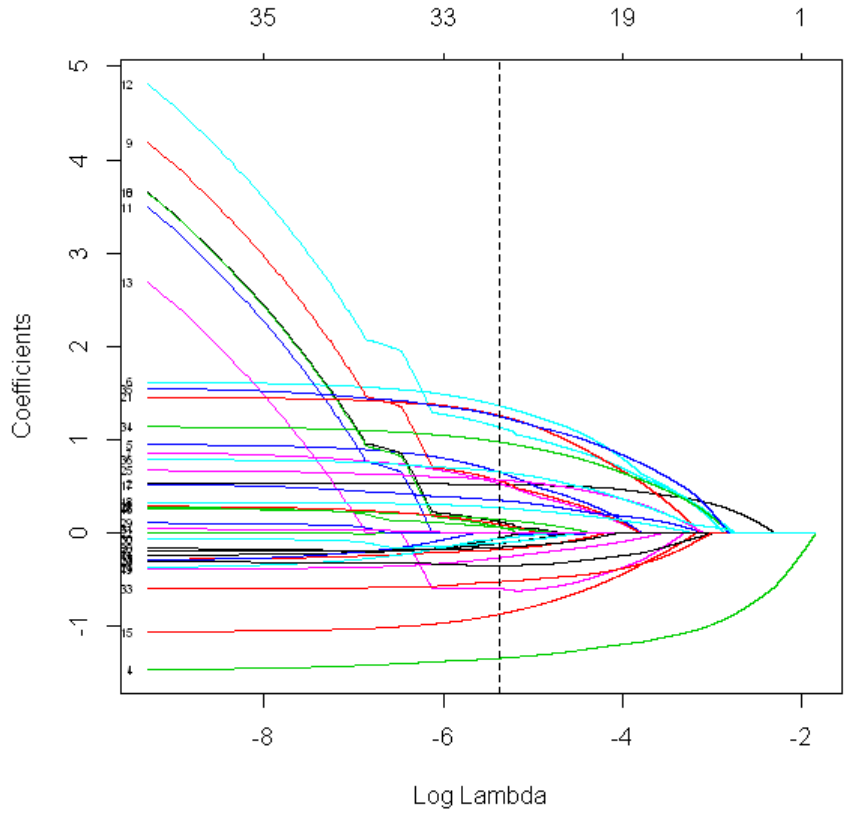
La démarche est globalement la même que pour la pénalisation Ridge, en utilisant le *package* glmnet.



(Intercept)	-3.48279072
ComptesCC [0-200 euros[	.
ComptesCC < 0 euros	0.52163811
ComptesCC > 200 euros	-0.17792518
ComptesPas de compte	-1.34334708
Historique_creditImpayé en cours dans autre banque	0.65367387
Historique_creditImpayés passés	1.36602572
Historique_creditPas de crédits ou en cours sans retard	0.53193916
Objet_creditBusiness	0.12065936
Objet_creditEtudes	0.56085507
Objet_creditIntérieur	0.06426693
Objet_creditVidéo - HIFI	.
Objet_creditVoiture neuve	1.13497339
Objet_creditVoiture occasion	-0.60051137
Epargne> 500 euros	-0.11216806
EpargneSans épargne	-0.87143536
Anciennete_emploientre 1 et 4 ans	0.14511484
Anciennete_emploiSans emploi ou < 1 an	0.35086248
Taux_effort	0.27152966
Situation_familialeHomme célibataire/marié/veuf	-0.26658154
Situation_familialeHomme divorcé/séparé	-0.05364326
GarantiesSans garant	1.25937805
Anciennete_domicile	.
BiensImmobilier	.
BiensNon immobilier	-0.06175868
Autres_creditsCrédits extérieurs	0.56945584
Statut_domicilePropriétaire	-0.34755366
Nb_credits	0.09355490
Type_emploiA172	0.06172710
Type_emploiA173	.
Type_emploiA174	-0.10508881
Nb_pers_charge2	.
TelephoneA192	-0.15320311
Age(25,Inf]	-0.52313899
Duree_credit(15,36]	0.97600559

La régression LASSO

La manière de contraindre les coefficients est toutefois très différente :



La régression LASSO

Lorsque l'hypothèse de parcimonie est satisfaite, il se peut toutefois que la méthode LASSO ne permette pas d'identifier les coefficients nuls correctement.

Une variante introduisant des poids sur les coefficients dans la pénalisation, la méthode LASSO adaptative, permet de résoudre cette difficulté.

$$l_{\theta,\varphi}(y) = \frac{y \times \theta - b(\theta)}{\varphi / w} + c(y, \varphi) + \lambda \sum_{j=1}^{p} \omega_j \times |\beta_j|$$

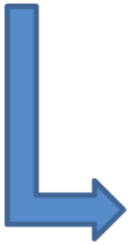
On peut montrer (cf. DENOYER et GUILLOT [2013] pour le cas linéaire) qu'avec un choix judicieux des poids et du coefficient de pénalisation, l'algorithme converge.

On peut par exemple prendre $\omega_j = \frac{1}{|\hat{\beta}_j^{mco}|}$

Synthèse : l'*elastic net*

L'idée est d'utiliser une pondération avec une partie Ridge et une partie LASSO, soit :

$$l_{\theta,\varphi}(y) = \frac{y\times\theta - b(\theta)}{\varphi/w} + c(y,\varphi) + \lambda\left(\alpha\sum_{j=1}^{p}|\beta_j| + (1-\alpha)\sum_{j=1}^{p}\beta_j^2\right)$$



```
# calcul de l'AUC maximale en test, pour chaque valeur de alpha
elastic <- function(a)
{
set.seed(235)
cvfit = cv.glmnet(x,y,alpha=a, family = "binomial",type="auc",nlambda=100)
# calcul de la régression pour une plage de valeurs de lambda
fit = glmnet(x,y,alpha=a,family = "binomial",lambda=seq(cvfit$lambda[1],
cvfit$lambda[length(cvfit$lambda)],length=10000),standardize = T)
# prédiction sur une plage de lambda sur la base de test
ytpred <- predict(fit,newx=xt,type="response")
roc <- function(x) { performance(prediction(ytpred[,x],yt),"auc")@y.values[[1]] }
vauc <- Vectorize(roc)(1:length(fit$lambda))
return(vauc[which.max(vauc)])
}
```

Sommaire

1. Les modèles GLM
2. Les régressions pénalisées
3. **Agrégation de modèles**
4. Points divers

3. Agrégation de modèles

,

L'agrégation de modèles est une technique d'estimation non paramétrique consistant à construire un estimateur comme moyenne pondérée d'estimateurs issus de différents modèles :

$$\hat{m}(x) = \frac{1}{B}\sum_{b=1}^{B} w_b \times \hat{m}_b(x)$$

L'idée est de réduire l'erreur de prévision sur un échantillon de validation :

$$e(\hat{m}) = \sum_{i=n+1}^{n+m} \left(y_i - \hat{m}(x_i)\right)^2$$

Ce critère est cohérent avec le fait que l'on estime une espérance conditionnelle, qui minimise l'écart quadratique avec le sous-espace engendré par les variables explicatives. La mesure d'erreur doit être adaptée avec la quantité estimée : par exemple pour la médiane, on retiendrait la norme L^1 et pour un modèle logistique, comme on l'a fait auparavant, l'AUC.

3. Agrégation de modèles

,

Le critère d'agrégation s'adapte lorsqu'il s'agit de prédire une variable qualitative ; par exemple s'il s'agit d'une variable binaire et que

$$m(x) = P(Y = 1 | x)$$

on peut soit calculer l'estimateur agrégé de *m* comme précédemment puis prédire *Y*, soit prédire *Y* pour chaque modèle *b* et en déduire la prédiction agrégée par un vote à la majorité. Cela se généralise à une variable catégorielle ; si $\hat{m}_{b,i}(x)$ est l'estimateur de $P(Y = i | x)$ dans le modèle *b*, il est naturel de poser

$$\hat{p}_b(x) = \arg\max_i \hat{m}_{b,i}(x)$$

pour prédire la classe et d'en déduire la classe estimée par le modèle agrégé par un vote majoritaire.

Le calcul de la variance de l'estimateur agrégé n'est simple que lorsque les échantillons utilisés sont indépendants :

$$V\left(\hat{m}\right)=\frac{1}{B^2}\sum_{b=1}^{B}w_b^2\times V\left(\hat{m}_b\right)$$

Mais il n'est pas en pratique possible d'avoir *B* échantillons indépendants (sauf à avoir un très grand volume de données).

La variance de la moyenne de *B* variables i.i.d., chacune de variance σ^2, est σ^2 / B. Si les variables sont identiquement distribuées mais avec une corrélation ρ des variables prises deux à deux, la variance de la moyenne devient :

$$\rho\times\sigma^2+\frac{1-\rho}{B}\sigma^2$$

La présence de corrélation entre les échantillons empêche donc la variance de tendre vers 0 et il est nécessaire de contourner cette difficulté.

3. Agrégation de modèles

,

Pour réduire la variance, le principe est de construire *B* échantillons *bootstrap* à partir de l'échantillon d'origine, d'ajuster un modèle sur chacun et d'agréger ensuite les résultats.

Le principe de la méthode *bootstrap* consiste à remarquer que pour un échantillon de taille suffisante, la fonction de répartition de la loi sous-jacente peut être approchée par la fonction de répartition empirique :

$$I(\varphi) = E(\varphi(X)) = \int \varphi dF \approx I_n(\varphi) = \int \varphi dF_n = \frac{1}{n}\sum_{i=1}^{n} \varphi(X_i) \quad F_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i \le x\}}$$

Évaluer des statistiques par simulation se ramène alors à générer des échantillons à l'aide de la distribution empirique. Hors, un tirage dans la distribution empirique s'obtient simplement par un tirage avec remise des *n* valeurs dans l'échantillon initial.

On obtient ainsi au plus n^n échantillons *boostrapés* à partir desquels on peut calculer les estimateurs empiriques des statistiques d'intérêt.

La théorie asymptotique fournie des informations sur la loi de la statistique $I_n(\varphi)$: le théorème central limite permet en effet de prouver que la loi asymptotique est normale et d'en déduire, par exemple, des intervalles de confiance de la forme

$$J_\alpha = \left[I_n(\varphi) - \frac{\sigma_\varphi}{\sqrt{n}} N^{-1}\left(1-\frac{\alpha}{2}\right), I_n(\varphi) + \frac{\sigma_\varphi}{\sqrt{n}} N^{-1}\left(1-\frac{\alpha}{2}\right) \right]$$

avec $\sigma_g^2 = \mathrm{Var}\left(\varphi(X)\right)$. La méthode Bootstrap permet notamment d'estimer cette variance et conduit typiquement à

$$\hat{\mu}\left[I_n^1(\varphi),\ldots,I_n^B(\varphi)\right] = \frac{1}{B}\sum_{b=1}^{B} I_n^b(\varphi) \qquad \hat{\sigma}^2\left[I_n^1(\varphi),\ldots,I_n^B(\varphi)\right] = \frac{1}{B-1}\sum_{b=1}^{B}\left[I_n^b(\varphi)-\hat{\mu}\right]^2$$

$$J_\alpha = \left[\hat{\mu} - \frac{\hat{\sigma}}{\sqrt{B}} N^{-1}\left(1-\frac{\alpha}{2}\right), \hat{\mu} + \frac{\hat{\sigma}}{\sqrt{B}} N^{-1}\left(1-\frac{\alpha}{2}\right) \right]$$

Le *bootstrap* est souvent biaisé et il existe des adaptations pour diminuer le biais (par exemple la méthode BCa).

3. Agrégation de modèles

,

Pour chaque échantillon *bootstrap*, il faut construire le modèle GLM « optimal » ce qui, on l'a vu, peut être envisagé de plusieurs manières.

En particulier, il faut décider si à chaque simulation on conserve l'ensemble des variables possibles (*bagging*) ou si on ne sélectionne qu'un sous-ensemble de ces variables (forêts aléatoires).

La seconde option permet en pratique de baisser la variance de l'estimateur agrégé et doit être préférée. La randomisation est donc effectuée à deux niveaux :

- sélection *bootstrap* de l'échantillon ;

- sélection aléatoire de *d* variables parmi *p* pour chaque échantillon.

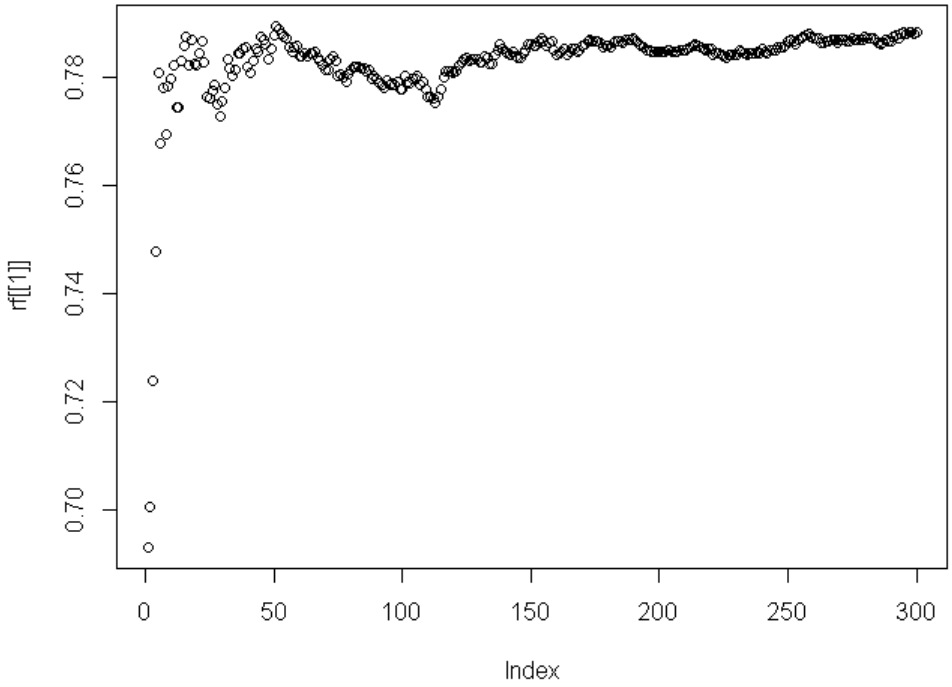
Le paramètre *d* est un paramètre de la méthode.

Une fois fixé l'échantillon et *d*, il reste à déterminer le modèle optimal : une sélection pas-à-pas constitue un choix raisonnable simple à mettre en œuvre.

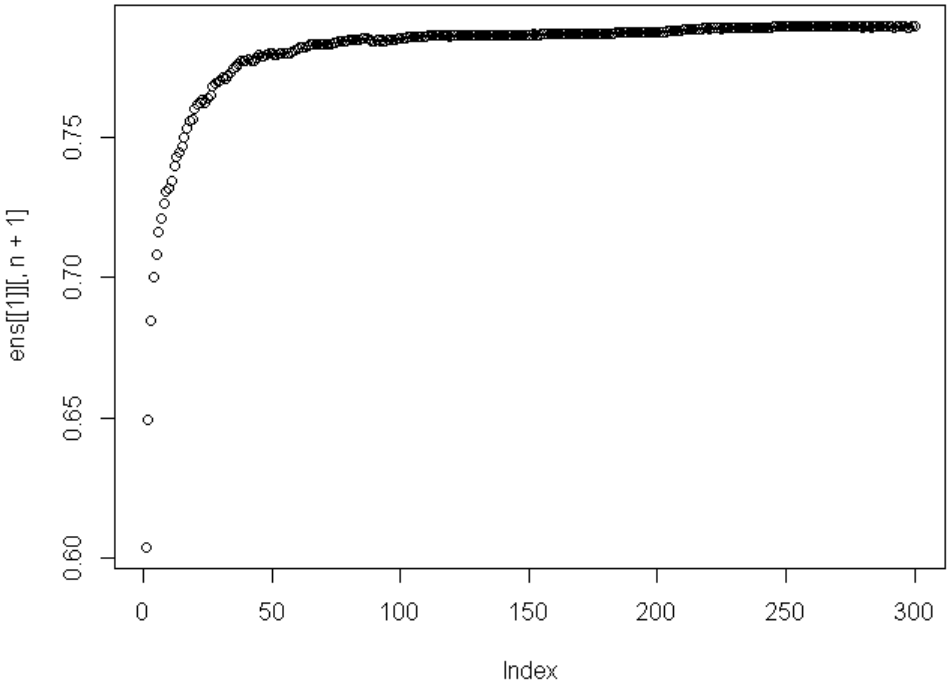
NB : le *boosting* (agrégation adaptative), n'est pas abordé ici.

3. Agrégation de modèles ,

On illustre cette technique avec 300 modèles agrégés en prenant 4 variables à chaque étape :



On peut lisser ces résultats en les exécutant 30 fois :



3. Agrégation de modèles

L'agrégation de modèle est très exigeante d'un point de vue numérique et les traitements doivent être en pratique parallélisés.

Le *package* « foreach » permet d'exécuter des traitements parallélisés

```
rf <- foreach(icount(nsimul), .combine='cbind') %dopar% {
    #instructions dans le boucle
   }
```

à condition d'avoir préalablement déclaré le nombre de cœurs, sur lesquels répartir les traitements, ce qui est possible avec les *packages* « parallel » (inclus dans le noyau de R) et « doSNOW »:

```
nCoeurs=detectCores()      #parallel
cl=makeClusters(nCoeurs)   #doSNOW
registerDoSNOW(cl)

rf <- foreach(icount(nsimul), .combine='cbind') %dopar% {
    #instructions dans le boucle
   }

stopCluster(cl)
```

Sommaire

1. Les modèles GLM
2. Les régressions pénalisées
3. Agrégation de modèles
4. **Points divers**

4. Points divers

La régression PLS (*Partial Least Squares Regression*, cf. TENENHAUS [1998])

La régression des moindres carrés partiels a été inventée en 1983 par Svante et Herman Wold. Dans le cadre linéaire, la régression PLS maximise la variance des prédicteurs *X* et maximise la corrélation entre *X* et la variable à expliquer *Y*.

Cet algorithme emprunte sa démarche à la fois à l'analyse en composantes principales (ACP) et à la régression. Plus précisément, la régression PLS cherche des composantes, appelées variables latentes, liées à *X* et à *Y*, servant à exprimer la régression de *Y* sur ces variables et finalement de *Y* sur *X*.

Le *package* R « plsRglm » permet une mise en œuvre aisée de cette technique dans le cadre de modèles GLM.

La régression PLS (*Partial Least Squares Regression*, cf. TENENHAUS [1998])

De manière plus précise, Pour régresser une variable *Y* (centrée) sur *p* variables explicatives $X = \left(X^1, \ldots, X^p \right)$ centrées, la méthode PLS propose de trouver de nouveaux facteurs qui joueront le même rôle que les variables explicatives initiales. Ces nouveaux facteurs sont appelés variables latentes ou composantes. Chaque composante est une combinaison linéaire des variables initiales, donc s'écrit

$$t^j = X \omega^j$$

et au global cela conduit à réécrire la matrice *X* $X = TP'$ avec des poids P à déterminer.

Les composantes *t* s'obtiennent par récurrence en utilisant un algorithme ; on peut montrer que le vecteur ω, de taille *p*, s'écrit :

$$\omega_j^1 = \frac{\mathbf{cov}\left(X^j, Y \right)}{\sqrt{\sum_{i=1}^{p} \mathbf{cov}\left(X^i, Y \right)}}$$

La régression PLS (*Partial Least Squares Regression*, cf. TENENHAUS [1998])

Ce coefficient est la pente de la droite des moindres carrés, passant par l'origine, du nuage de points $\left(X_i^j, Y_i\right)_{i=1,\ldots,n}$.

La composante de régression associée, qui est un vecteur de taille *n* s'écrit elle :

$$t_i^1 = \frac{1}{\left\|\omega^1\right\|^2} \sum_{j=1}^{p} \omega_j^1 X^j$$

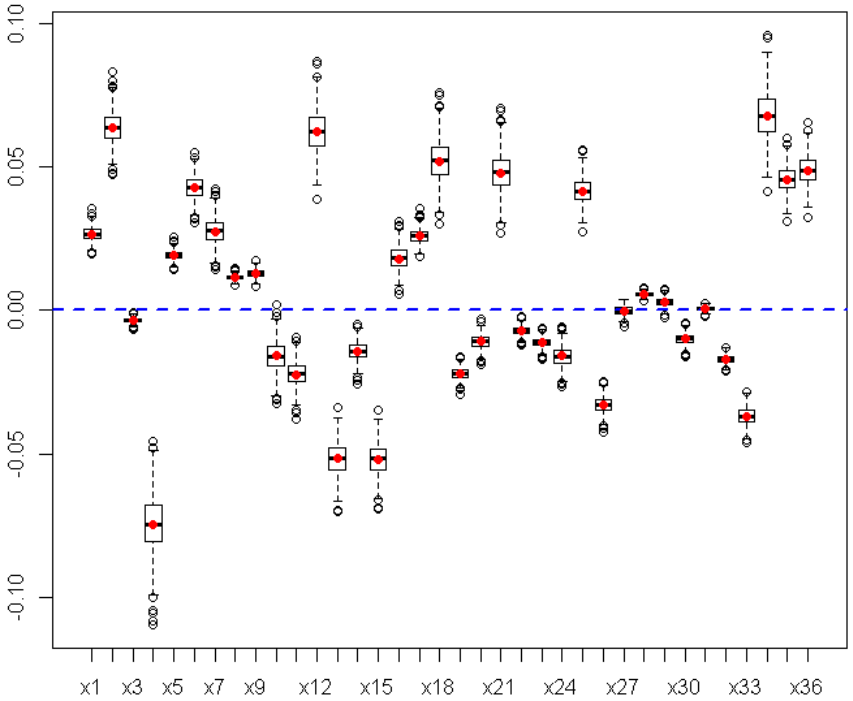
qui est la la pente de la droite des moindres carrés, passant par l'origine, du nuage de points $\left(\omega_j^1, X_i^j\right)_{j=1,\ldots,p}$.

https://fr.wikipedia.org/wiki/R%C3%A9gression_des_moindres_carr%C3%A9s_partiels

La régression PLS (*Partial Least Squares Regression*, cf. TENENHAUS [1998])

Avec les données utilisées jusqu'à présent, on obtient ainsi des résultats dont la forme est identique à un GLM classique :

L'AUC de 0,779 est proche du niveau obtenu avec une régression Ridge.



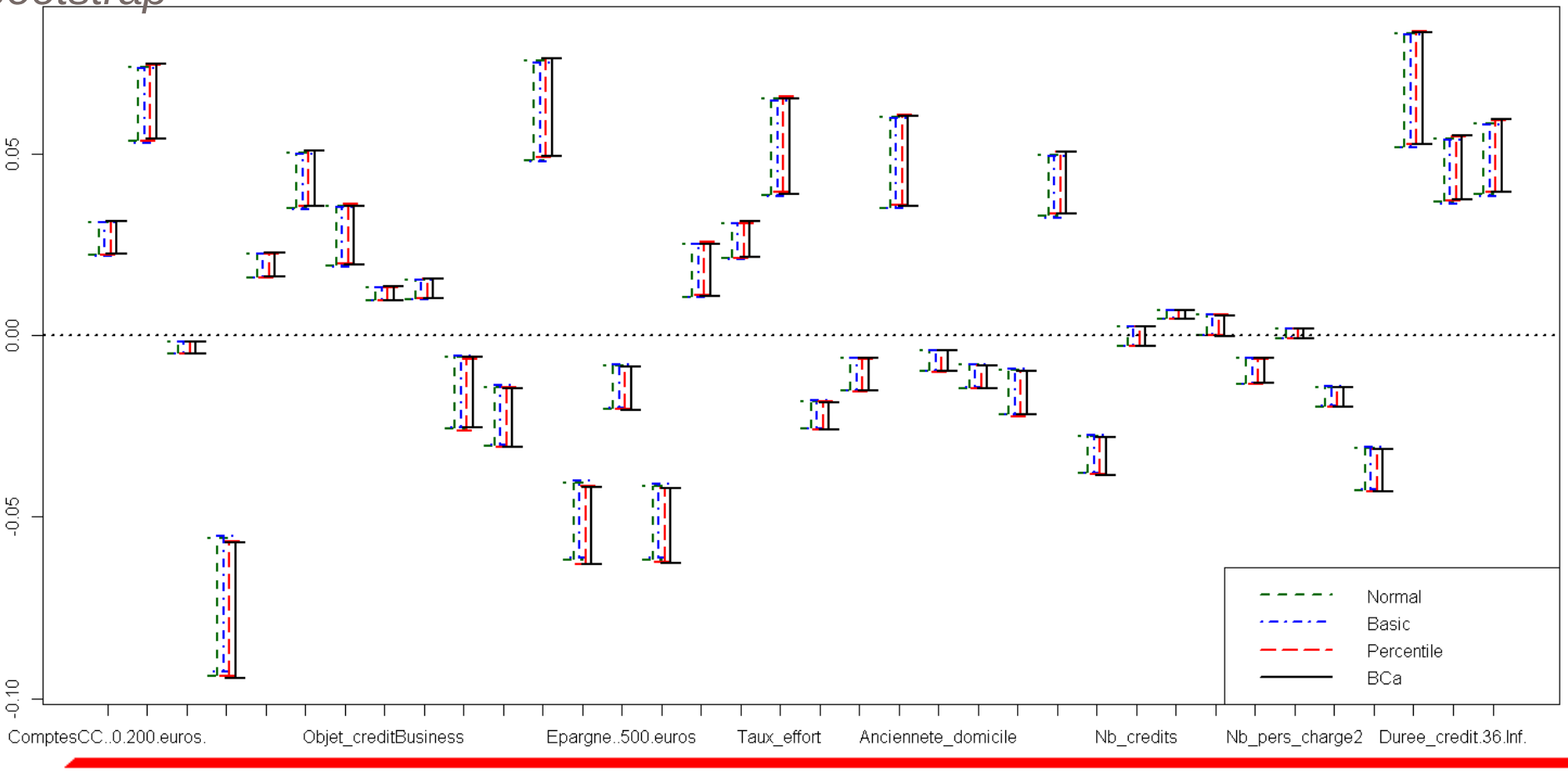
Intercept	-1.336607261
ComptesCC..0.200.euros.	0.396421654
ComptesCC...0.euros	0.922305481
ComptesCC...200.euros	-0.165925194
ComptesPas.de.compte	-1.292454430
Historique_creditImpayé.en.cours.dans.autre.banque	0.587905290
Historique_creditImpayés.passés	0.958498190
Historique_creditPas.de.crédits.ou.en.cours.sans.retard	0.125529463
Objet_creditEtudes	0.069781893
Objet_creditIntérieur	0.099021316
Objet_creditVidéo...HIFI	-0.510189580
Objet_creditVoiture.neuve	0.514395062
Objet_creditVoiture.occasion	-0.661512645
Epargne..500.euros	-0.478105470
EpargneSans.épargne	-0.537177851
Anciennete_emploientre.1.et.4.ans	0.022269205
Anciennete_emploiSans.emploi.ou...1.an	0.443396631
Taux_effort	0.140362221
Situation_familialeHomme.célibataire.marié.veuf	-0.312870577
Situation_familialeHomme.divorcé.séparé	-0.019977478
GarantiesSans.garant	0.619167262
Anciennete_domicile	-0.003693532
BiensImmobilier	-0.226111860
BiensNon.immobilier	-0.027516297
Autres_creditsCrédits.extérieurs	0.417052927
Statut_domicilePropriétaire	-0.442751817
Nb_credits	-0.066625159
Type_emploiA172	0.051171883
Type_emploiA173	-0.048050132
Type_emploiA174	-0.041216351
Nb_pers_charge2	-0.044270230
TelephoneA192	-0.180708552
Age.25.Inf.	-0.505461230
Duree_credit.15.36.	0.455527510
Duree_credit.36.Inf.	0.670248603
Montant_credit.4e.03.Inf.	0.464456508

4. Points divers

,

La régression PLS (*Partial Least Squares Regression*, cf. TENENHAUS [1998])

On obtient des intervalles de confiance des coefficients avec la méthode *bootstrap*



4. Points divers

,

Les modèles « à inflation de zéros » (cf. VASECHKO et *al*. [2009])

Le nombre de sinistres observé est décomposé en produit de deux variables :

$$Y = B \times Y^*$$

B est une indicatrice égale à 1 si le sinistre est déclaré et 0 sinon (elle n'est donc pas observable). Y^* est supposé suivre une loi de Poisson (modèle ZIP) ou binomiale négative (ZINB). On a donc typiquement des équations du type :

$$P\left(Y=0\middle|X\right)=q+\left(1-q\right)e^{-\lambda} \qquad P\left(Y=y\middle|X\right)=\left(1-q\right)e^{-\lambda}\frac{\lambda^y}{y!} \qquad q=\frac{\exp\left(X'\beta\right)}{1+\exp\left(X_i'\beta\right)}$$

pour la partie « inflation de zéro » et un modèle GLM usuel pour la variable Y^*. (qui n'est pas observable complètement).

4. Points divers

Les modèles « à inflation de zéros » (cf. VASECHKO et *al*. [2009])

Pour tester si la version avec inflation de zéro du modèle est préférable, on peut utiliser le test de Vuong, qui repose sur la statistique suivante :

$$Z = \frac{1}{\sigma_n \sqrt{n}} \sum_{i=1}^{n} l_i - \frac{p_1 - p_2}{2} \ln(n)$$

avec $l_i = \ln \frac{f_1(y_i|\beta_1)}{f_2(y_i|\beta_2)}$ et $\sigma_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (l_i - \bar{l})^2$

Cette statistique tend sous l'hypothèse nulle vers une loi normale centrée réduite.

NB : l'hypothèse nulle est simplement : $E(l_i) = 0$

4. Points divers

La lecture et l'interprétation des résultats présentent l'avantage d'être aisés et directs.

Ici un exemple avec la fonction de lien *log* et une réponse gamma :

```
Call:
glm(formula = formule, family = poisson(link = "log"), data = tFrequences,
    na.action = na.omit)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9400  -0.6482  -0.5171  -0.1362   7.1467

Coefficients:
                           Estimate Std. Error  z value Pr(>|z|)
(Intercept)               -1.353978   0.007001 -193.410   <2e-16 ***
CANCPER                    0.005391   0.000464   11.620   <2e-16 ***
CZONVAM18                  0.003298   0.005955    0.554     0.58
CZONVAM98                 -3.622889   0.053726  -67.433   <2e-16 ***
ClasseAge_en_2011(47,60]  -0.105110   0.007613  -13.807   <2e-16 ***
ClasseAge_en_2011(60,67]  -0.236671   0.008378  -28.249   <2e-16 ***
ClasseAge_en_2011(67,122] -0.312960   0.008226  -38.047   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 489961  on 804999  degrees of freedom
Residual deviance: 452733  on 804993  degrees of freedom
  (374 observations deleted due to missingness)
AIC: 669969

Number of Fisher Scoring iterations: 7
```

Ajustement d'un modèle de régression ZIP

Les résultats pour la composante de comptage sont les suivants :

Les résultats pour la composante d'inflation de zéros sont les suivants :

```
Call:
zeroinfl(formula = formule, data = tFrequences, na.action = na.omit, dist = "poisson")

Pearson residuals:
     Min       1Q   Median       3Q      Max
-0.52147 -0.42410 -0.37991 -0.09357 46.94874

Count model coefficients (poisson with log link):
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -0.9739031  0.0144721 -67.295  < 2e-16 ***
CANCPER                  0.0078058  0.0007328  10.651  < 2e-16 ***
CZONVAM18                0.0143409  0.0131852   1.088    0.277
CZONVAM98               -1.8450051  0.1429489 -12.907  < 2e-16 ***
ClasseAge_en_2011(47,60] -0.1032005  0.0163481  -6.313 2.74e-10 ***
ClasseAge_en_2011(60,67] -0.1866246  0.0184762 -10.101  < 2e-16 ***
ClasseAge_en_2011(67,122] -0.1722882  0.0182987  -9.415  < 2e-16 ***

Zero-inflation model coefficients (binomial with logit link):
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -0.629157   0.038647 -16.280  < 2e-16 ***
CANCPER                  0.007357   0.001427   5.155 2.53e-07 ***
CZONVAM18                0.032714   0.034318   0.953   0.3405
CZONVAM98                2.300712   0.191852  11.992  < 2e-16 ***
ClasseAge_en_2011(47,60] -0.029619   0.044749  -0.662   0.5080
ClasseAge_en_2011(60,67]  0.101351   0.048108   2.107   0.0351 *
ClasseAge_en_2011(67,122]  0.324660   0.044476   7.300 2.89e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ajustement d'un modèle de régression ZINB

Les résultats pour la composante de comptage sont les suivants :

Les résultats pour la composante d'inflation de zéros sont les suivants :

```
Call:
zeroinfl(formula = formule, data = tFrequences, na.action = na.omit, dist = "negbin")

Pearson residuals:
     Min       1Q   Median       3Q      Max
-0.52160 -0.42410 -0.37991 -0.09357 46.94873

Count model coefficients (negbin with log link):
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -0.9739004  0.0144729 -67.291  < 2e-16 ***
CANCPER                  0.0078009  0.0007331  10.641  < 2e-16 ***
CZONVAM18                0.0143421  0.0131855   1.088    0.277
CZONVAM98               -1.8441773  0.1429545 -12.900  < 2e-16 ***
ClasseAge_en_2011(47,60] -0.1031954  0.0163483  -6.312 2.75e-10 ***
ClasseAge_en_2011(60,67] -0.1866033  0.0184766 -10.099  < 2e-16 ***
ClasseAge_en_2011(67,122] -0.1722403  0.0182991  -9.413  < 2e-16 ***
Log(theta)              14.3015579         NA      NA       NA

Zero-inflation model coefficients (binomial with logit link):
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -0.629185   0.038651 -16.279  < 2e-16 ***
CANCPER                  0.007345   0.001428   5.142 2.71e-07 ***
CZONVAM18                0.032745   0.034320   0.954    0.340
CZONVAM98                2.302237   0.191923  11.996  < 2e-16 ***
ClasseAge_en_2011(47,60] -0.029577   0.044752  -0.661    0.509
ClasseAge_en_2011(60,67]  0.101412   0.048110   2.108    0.035 *
ClasseAge_en_2011(67,122]  0.324772   0.044478   7.302 2.84e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4. Points divers

Quel modèle retenir ?

- Le modèle à inflation de zéros domine largement les modèles de Poisson et Binomial Négatif
- Le modèle Binomial Négatif domine le modèle de Poisson.
- Le test ne permet pas de décider si le modèle ZIP domine le modèle ZINB.

```
> vuong(regPoisson,regZIP)
Vuong Non-Nested Hypothesis Test-Statistic: -34.45677
(test-statistic is asymptotically distributed N(0,1) under the
 null that the models are indistinguishible)
in this case:
model2 > model1, with p-value 1.782859e-260
> vuong(regNegBin,regZIP)
Vuong Non-Nested Hypothesis Test-Statistic: -35.34582
(test-statistic is asymptotically distributed N(0,1) under the
 null that the models are indistinguishible)
in this case:
model2 > model1, with p-value 5.811387e-274
> vuong(regPoisson,regNegBin)
Vuong Non-Nested Hypothesis Test-Statistic: -19.02048
(test-statistic is asymptotically distributed N(0,1) under the
 null that the models are indistinguishible)
in this case:
model2 > model1, with p-value 5.771123e-81
```

```
> vuong(regZINB,regZIP)
Vuong Non-Nested Hypothesis Test-Statistic: -0.09978337
(test-statistic is asymptotically distributed N(0,1) under the
 null that the models are indistinguishible)
in this case:
model2 > model1, with p-value 0.4602582
```

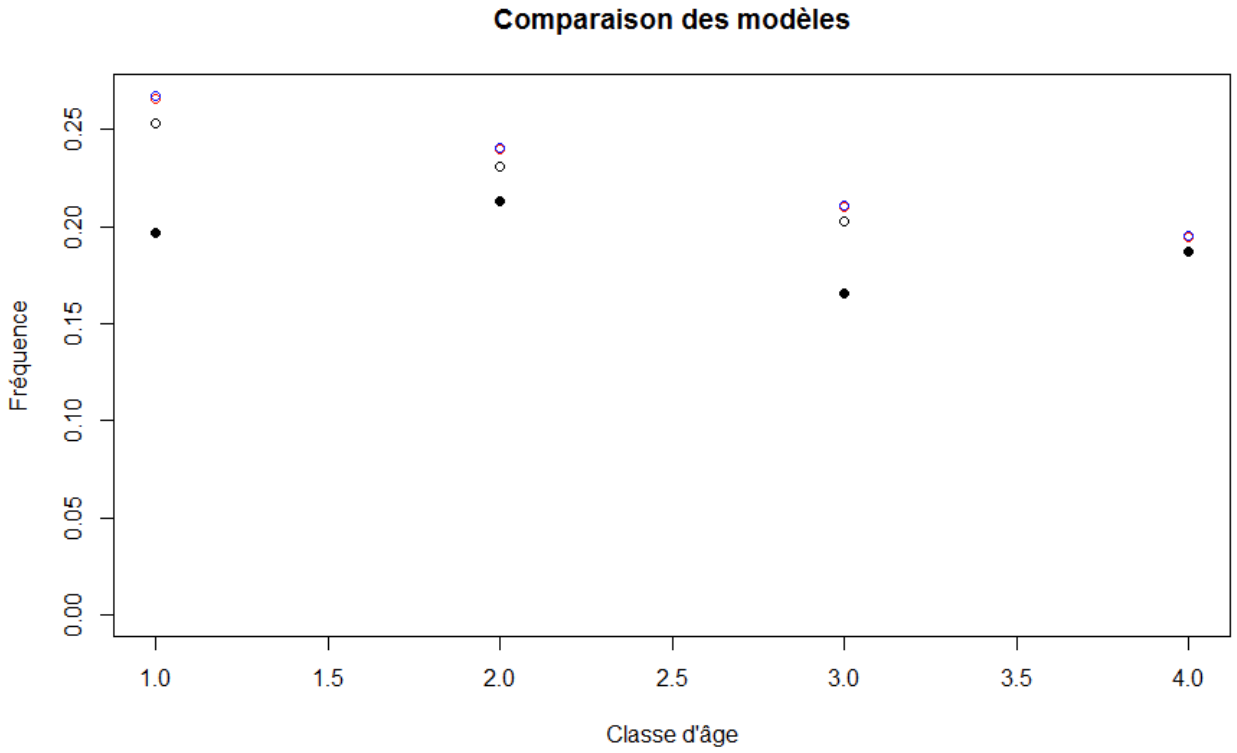
4. Points divers

Comparaison des modèles

Les prédictions de fréquences effectuées avec ces modèles sont en pratique parfois très proches.

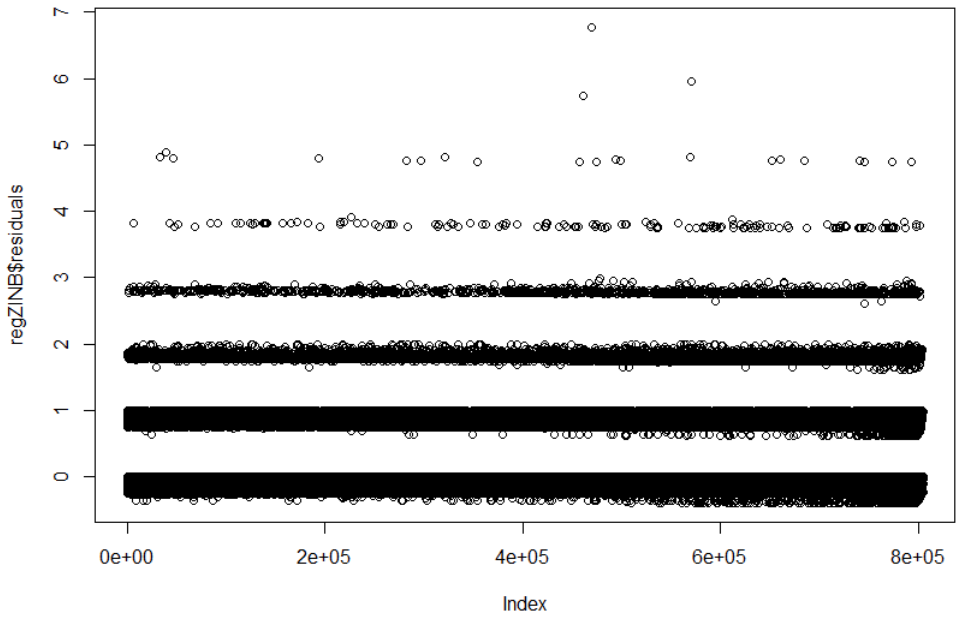
A titre d'illustration on présente les valeurs modélisées en fonction de la variable « classe d'âge » avec les modalités des autres variables fixées.

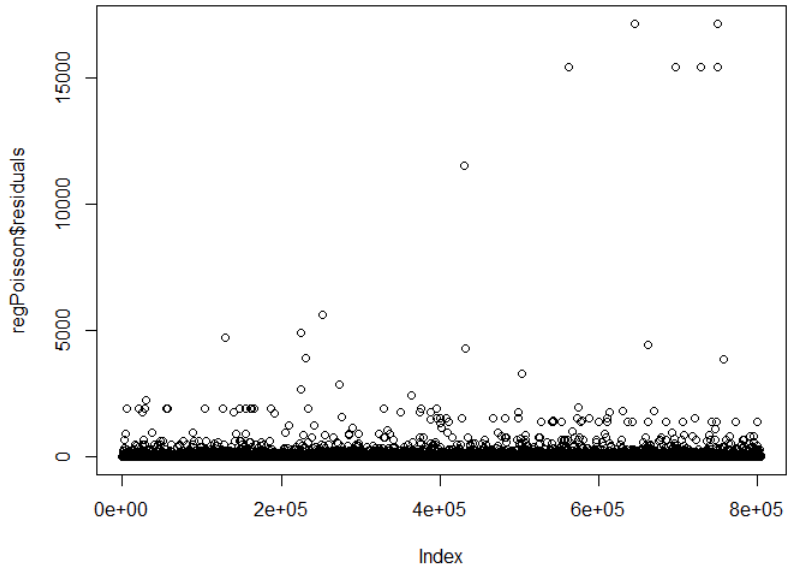
Les valeurs prédites sont assez éloignées des valeurs brutes…



Comparaison des modèles

L'analyse plus systématique de la pertinence d'un modèle passe également par l'analyse des résidus, ici de Pearson, qui met en évidence la supériorité du ZINB :





Utilisation et limites

L'approche GLM impose de faire une hypothèse sur la forme de la loi conditionnelle de la variable expliquée *Y* en fonction des explicatives. Cette hypothèse peut s'avérer fausse et on prend donc un risque de modèle.

On peut alors chercher à modéliser directement la forme de l'espérance conditionnelle, mais sans faire d'hypothèse sur la loi complète de la variable expliquée (régression non paramétrique, modèles GAM, réseaux de neurones, *etc*.).

Une telle démarche est de nature à faire diminuer le risque de modèle, les hypothèses sur lesquelles reposent l'évaluation de la prime pure étant moins restrictives.

Elle a été mise en œuvre par exemple dans DUPIN et *al*. [2003].

4. Points divers

Perspectives d'évolution méthodologiques

L'intérêt pour les données massives conduit les actuaires à s'intéresser à d'autres approches issues de la théorie statistique de l'apprentissage.

PAGLIA et PHELIPPE-GUINVARC'H [2011] proposent ainsi, dans la situation classique de la tarification d'un contrat d'assurance automobile, une comparaison entre les approches classiques par GLM et une méthode fondée sur la théorie de l'apprentissage.

La classification automatique (*clustering*) est un outil très utilisé en fouille de données (*data mining*) permet d'extraire d'un grand jeu de données des classes où les individus ont des caractéristiques similaires.

Perspectives d'évolution méthodologiques

Théorie de l'apprentissage

La statistique classique nécessite de formuler des hypothèses sur la distribution des données. La théorie de l'apprentissage statistique ne formule qu'une seule hypothèse : les données à prédire Y sont générées de façons identiques et indépendantes par un processus P à partir du vecteur des variables explicatives X.

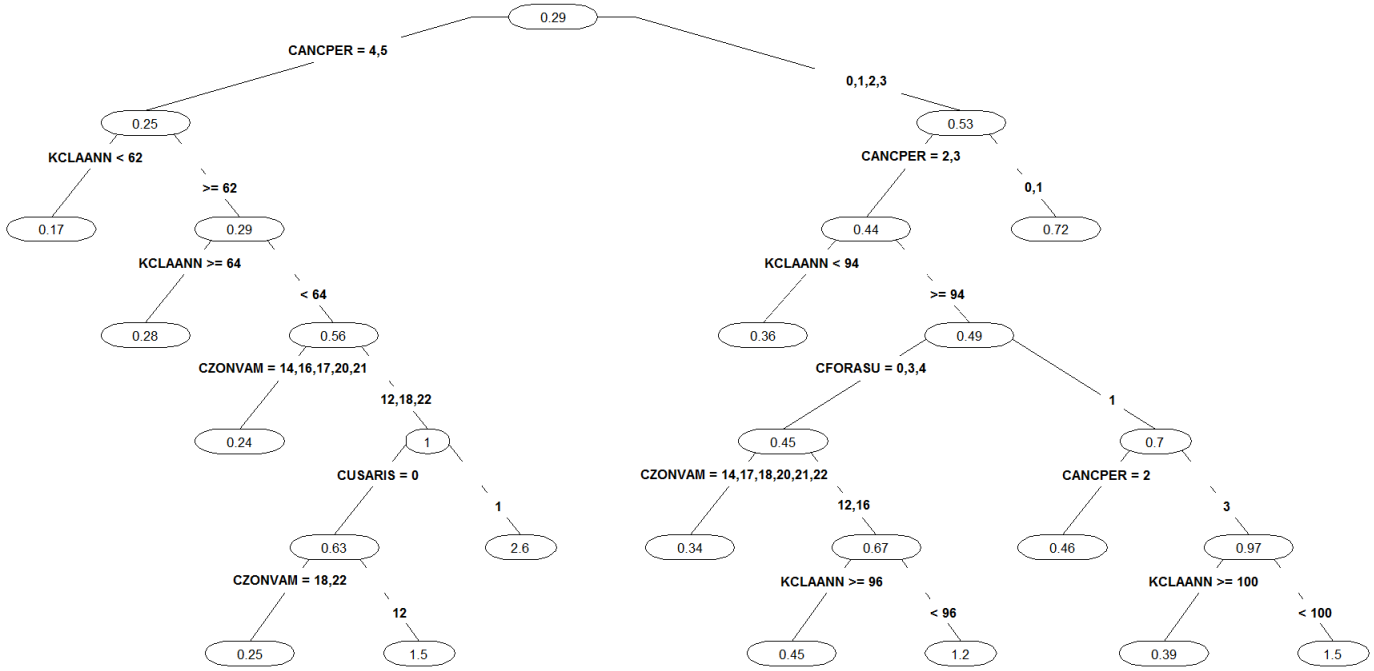
On cherche alors à construire un algorithme qui va apprendre à prédire la valeur de Y en fonction des valeurs explicatives X (i.e. $E[Y|X]$). Le résultat de cet apprentissage est une fonction $f(X,c)$. Elle fait intervenir les variables X et un paramètre de complexité c. Ce paramètre désigne par exemple le nombre de neurones dans un réseau de neurones (cf. AOUIZERATE [2012]) ou le nombre de nœuds dans un arbre de décision.

On doit disposer d'une base d'apprentissage et d'une base de validation.

Perspectives d'évolution méthodologiques

Théorie de l'apprentissage

Le tarif obtenu par une méthode de type CART (*Classification And Regression Tree*) présente une structure arborescente. Voici un exemple de modélisation de la fréquence des sinistres à partir de données « automobile » :



La méthode CART (BREIMAN et *al*. [1984])

La méthode consiste à construire un arbre binaire. A chaque nœud, l'algorithme recherche la séparation qui maximise le gain de variance, de sorte que la somme des variances intra groupe des nœuds fils soit plus faible que la variance du nœud père.

A l'intérieur de chaque nœud, la grandeur modélisée (fréquence ou coût moyen) est estimée par son espérance empirique.

L'intérêt de cette méthode est d'ordonner les variables des plus influentes en haut de l'arbre aux moins influentes en bas. L'utilisateur contrôle la complexité de l'arbre *via* le nombre de nœuds maximum et l'effectif minimum dans chaque nœud.

On peut résumer la démarche avec l'algorithme ci-après :

La méthode CART (BREIMAN et *al*. [1984])

1- Pour une une covariable donnée : soit elle est numérique, ordonnée et les partitionnements possibles de l'espace associé à cette covariable se situent entre 2 de ses valeurs successives observées, et ce pour toutes les valeurs ; soit elle est catégorielle et les partitionnements possibles de j associé à cette covariable sont toutes les combinaisons de modalités ;

2- On teste tous ces partitionnements en calculant un critère d'homogénéité par rapport à la quantité d'intérêt (réponse) ;

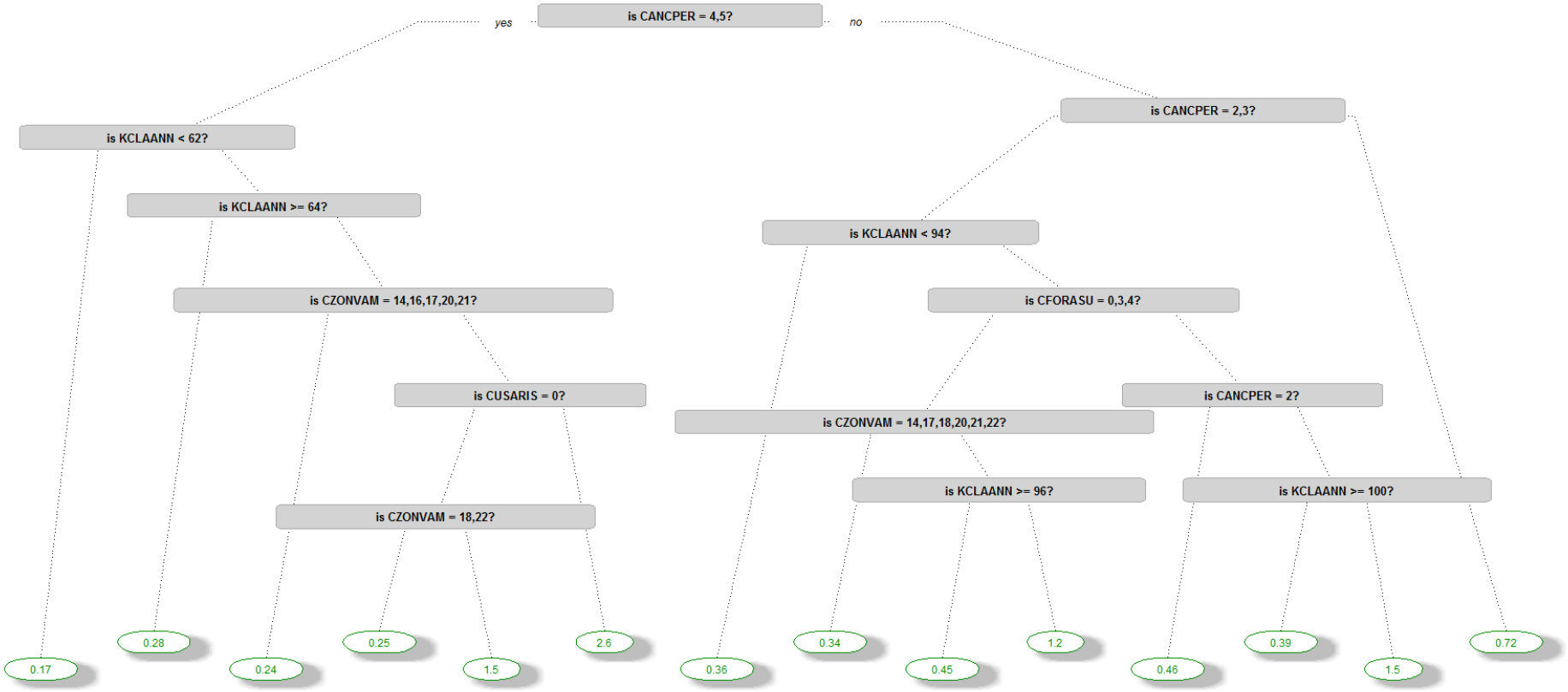
3- On choisit le partitionnement qui conduit à la plus grande homogénéité dans les sous-espaces créés ;

4 On répète les étapes (1)-(3) pour chacune des covariables : on obtiens une liste de *p* homogénéités max. ;

5 On choisit la covariable et son partitionnement qui maximisent l'homogénéité globalement.

La méthode CART (Breiman et *al*. [1984])

On obtient des résultats dont l'allure est la suivante (pour la fréquence) :



La méthode CART (BREIMAN et *al*. [1984])

Pour intégrer des ajustements *ex-post* dans le tarif (équivalents aux lissages des coefficients dans un modèle GLM), on peut directement modifier le tarif associé à un nœud et répartir la perte ou le gain sur les autres nœuds par exemple au prorata de l'exposition. Mais la règle de redressement est moins claire que dans le cadre d'un modèle GLM.

L'arbre optimal sur l'échantillon d'apprentissage n'est pas forcément le meilleur pour la prédiction. Il est donc en pratique nécessaire d'effectuer des ajustements pour éviter le sur-apprentissage.

La méthode du *bagging* (*bootstrap aggregation*) qui consiste à construire des arbres par *bootstrap* puis à utiliser la moyenne des prédicteurs de chaque arbre comme prédiction en est une illustration. Cela permet de diminuer la variance de la prédiction mais on perd la principale qualité d'un arbre de décision : la lisibilité du tarif. La méthodes des forêts aléatoires en constitue une variante plus souple.

Les modèles GAM

L'idée des modèles additifs est de relâcher l'hypothèse de linéarité du prédicteur que l'on impose dans un GLM :

$$g\left(\mathbf{E}\left[Y\middle|x_1,\ldots,x_p\right]\right)=\sum_{k=1}^{p}\beta_k\,x_k$$

en supposant la forme plus générale

$$g\left(\mathbf{E}\left[Y\middle|x_1,\ldots,x_p\right]\right)=\sum_{k=1}^{p}f_k\left(x_k\right)$$

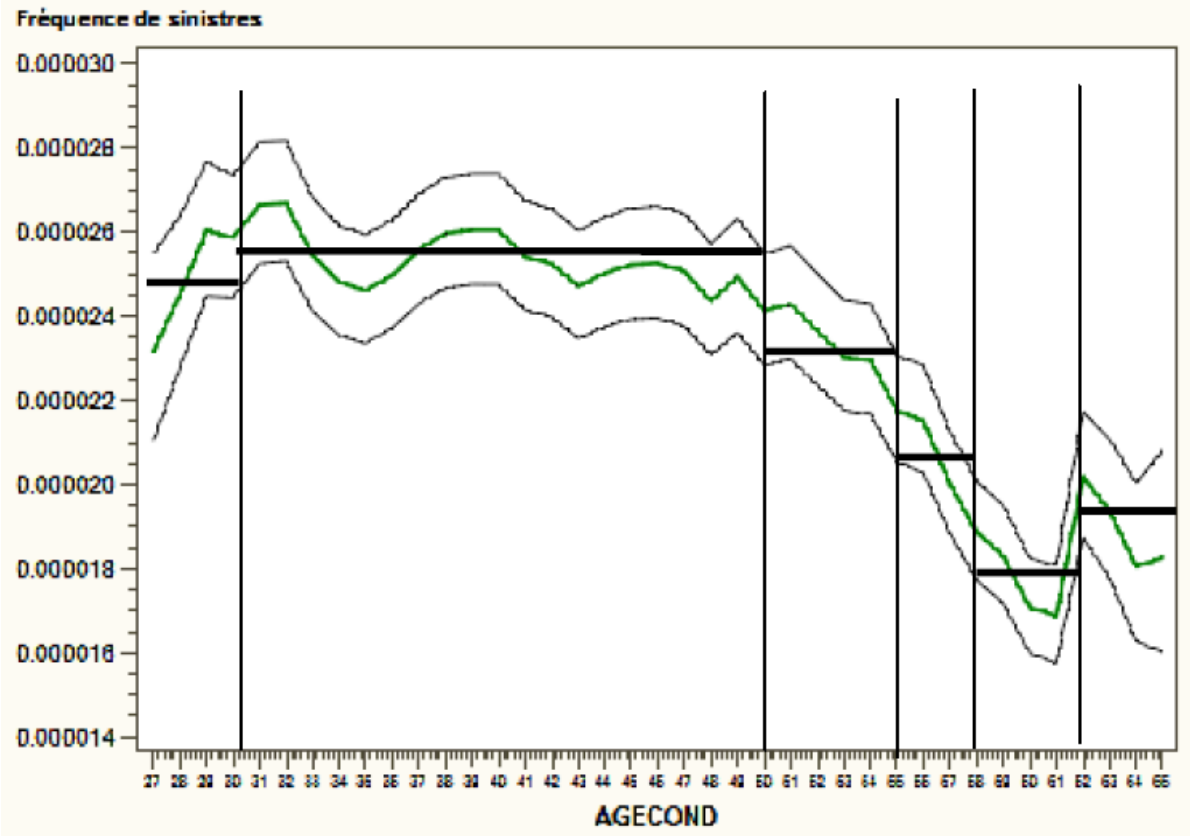
L'estimation des fonctions associées aux variables explicatives est effectuée par des méthodes semi-paramétriques de lissage (splines pénalisés par exemple).

Les modèles GAM

Les modèles GAM peuvent être utilisés en amont d'un modèle GLM pour définir le découpage en classes d'une variable continue dont l'effet est non linéaire.

Le graphique suivant, repris de POUNA-SIEWE [2010], illustre ce type d'utilisation en indiquant les classes construites à l'aide des intervalles de confiance de la courbe marginale estimée.

NB : avec un modèle CART cette étape est inutile.



Conclusion

Un tarif est un objet complexe dont la construction mobilise différents modèles en fonction des composants à décrire :

- discrétisation de variables continues (GAM) ;

- zonier (modèles bayésiens) ;

- structure tarifaire de base (GLM) avec, pour la fréquence, une attention particulière portée à la sur-dispersion et à la sous-déclaration des petits sinistres.

Il n'existe *a priori* pas de modèle unique qui permette de rendre compte de tous ces effets de manière globale, y compris dans le cadre « standard » discuté ici d'un tarif construit avec un nombre relativement restreint de variables explicatives.

Conclusion

La possibilité de prendre en compte dans certains contextes (santé, automobile, MRH) des données beaucoup plus fines conduit à reconsidérer le cadre même de tarification.

On se trouve en effet confronté à des situations dans lesquelles le nombre de variables tarifaires devient très grand, ce qui dégrade la qualité des estimateurs de la fréquence et du coût moyen.

Les techniques de pénalisation et d'agrégation permettent de répondre à cette problématique de manière efficace, au prix d'une utilisation intensive des ressources informatiques de l'ordinateur.

Références

AOUIZERATE J.M. [2012] «Alternative neuronale en tarification santé », *Bulletin Français d'Actuariat*, vol. 12, n°23.

BOSKOV M.; VERRALL R.J. [1994] « Premium Rating by Geographic Area Using Spatial Models », *ASTIN Bull.*, 24 (1994), No 1, 131-143.

BREIMAN L., OLSHEN L., FRIEDMAN R., STONE J. [1984] *Classification and regression trees*, Chapman & Hall

CHARPENTIER A. (editor) [2014] *Computational Actuarial Science*, with R, The R Series. Chapman and Hall.

DENOYER A., GUILLOT T. [2013] Méthodes de type LASSO pour modélisation et sélection de variables en très grande dimension, mémoire de master.

DENUIT M., CHARPENTIER A. [2005] *Mathématiques de l'assurance non-vie. Tome II : tarification et provisionnement,* Paris : Economica.

DUPIN G.; MONFORT A.; VERLÉ J.P. [2003] « Robust inference in rating models » *Proceedings of the 34th ASTIN Colloquium*.

FURNIVAL, G.M.; WILSON, R.W. [1974]. Regression by leaps and bounds. *Technometrics*, 16, 499-511.

LEROY G., PLANCHET F. [2016] « Un regard actuariel sur les évolutions de l'assurance automobile », *Risques*, n°105.

MATHIS J. [2009] « Elaboration d'un zonier en assurance de véhicules par des méthodes de lissage spatial basées sur des simulations MCMC », ISFA, mémoire d'actuariat.

NELDER J., WEDDERBURN R. [1972] « Generalized linear models », *Journal of Roy. Stat. Soc.* B, vol. 135, 370-384.

PAGLIA A., PHELIPPE-GUINVARC'H M.V. [2011] « Tarification des risques en assurance non-vie, une approche par modèle d'apprentissage statistique », *Bulletin Français d'Actuariat*, vol. 11, n°22.

PARTRAT C., BESSON J.L., [2004] *Assurance non-vie – modélisation, simulation*, Paris : Economica.

PLANCHET F., THÉROND P.E., JUILLARD M. [2011] *Modèles financiers en assurance*, seconde édition, Paris : Economica.

POUNA SIEWE V. [2010] Modèles additifs généralisés : Intérêts de ces modèles en assurance automobile, ISFA, Mémoire d'actuariat

Références

R Development Core Team [2016] R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.R-project.org

TENENHAUS M. [1998] *La régression PLS – Théorie et pratique*, Éditions TECHNIP.

TUFFERY S. [2015] *Modélisation prédictive et apprentissage statistique avec R*, Éditions TECHNIP.

VASECHKO O.A.; GRUN-REHOMME M.; BENLAGHA N. [2009] « Modélisation de la fréquence des sinistres en assurance automobile », *Bulletin Français d'Actuariat*, vol. 9, n°18.

YANG Y. [2005] « Can the strengths of AIC and BIC be shared? », *Biometrika*, vol. 92, p. 937–950.