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

ressources-actuarielles.net

Modèles fréquence – coût :

Quelles perspectives d'évolution ?

Version 0.7

Mars 2014

Frédéric PLANCHET
frederic@planchet.net
Guillaume SERDECZNY
guillaume.serdeczny@maif.fr

Réunion du 21 mars 2014

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.

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.

On se propose ici de revisiter les principales étapes de la construction d'un tarif en examinant l'intérêt de l'utilisation de ces approches alternatives.

Préambule

Les étapes d'une tarification

La réalisation d'un tarif 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

Périmètre de la présentation

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.

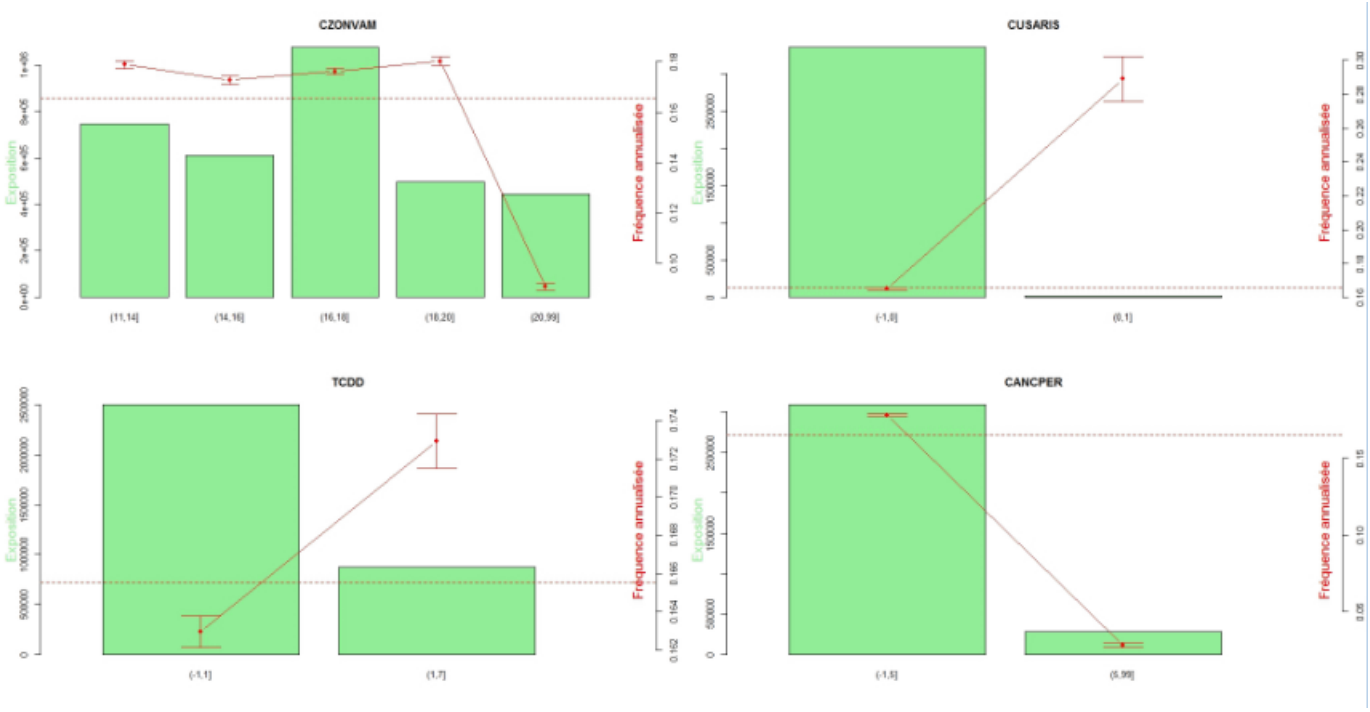
Du fait de ces restrictions, l'impact de la réassurance (non proportionnelle) n'est pas abordé.

Préambule

Impact des caractéristiques du client sur la fréquence

Il s'agit de modéliser 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*.

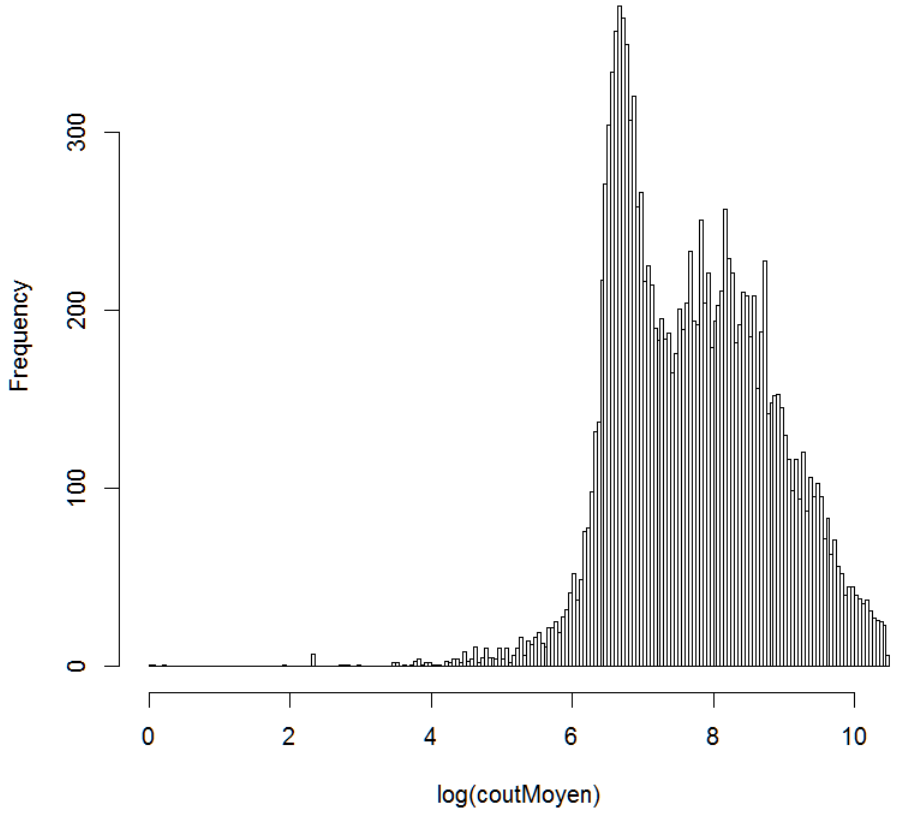


Préambule

Impact des caractéristiques du client sur la fréquence

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.



SOMMAIRE

1. Le cadre standard

Le cadre usuel de tarification

En pratique la tarification IARD est en général 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).

1. Le cadre standard

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 de prédire des espérances conditionnelles, ce qui est le cadre général des modèles de régression, et plus particulièrement des modèles de régression non linéaires (GLM).

1. Le cadre standard

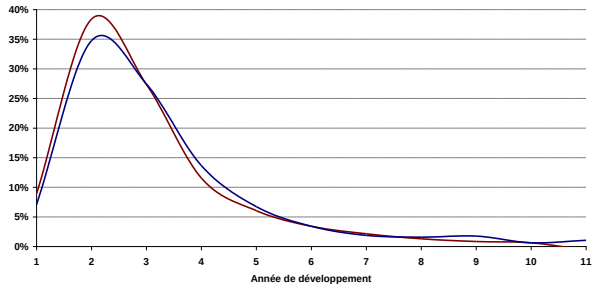
Identification des sinistres graves et sériels

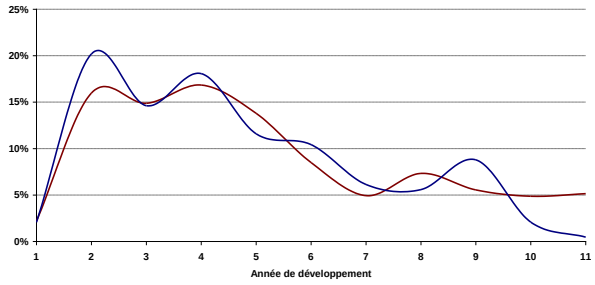
L'identification des sinistres sériels s'appuie sur la mise en relation du sinistre avec un événement, en général codé dans le base de données.

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 :



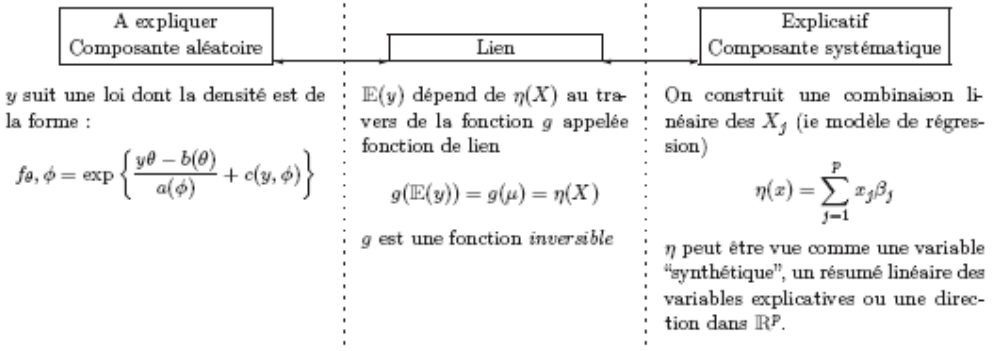


1. Le cadre standard

On a souvent recours aux modèles linéaires généralisés (*Generalized Linear Models*, GLM) pour les aspects de segmentation de l'offre.

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]).

La théorie des GLM bénéficie d'un avantage par rapport aux modèles linéaires classiques : le caractère normal de la variable à expliquer Y n'est plus imposé, seule l'appartenance à une famille exponentielle est indispensable.



1. Le cadre standard

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) $g(u) = \mathbf{ln}(u)$ conduit au modèle multiplicatif suivant :

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

1. Le cadre standard

Il reste à spécifier la loi de la variable *Y*. On retient une 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 :

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

1. Le cadre standard

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	$v(\mu)=\mu^2$

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])

1. Le cadre standard

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 log-normale pour le coût.

Remarque : 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.

1. Le cadre standard

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*.

1. Le cadre standard

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 :

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

$$t = \begin{cases} x_1 + \ln(0,95) & EFFECTIF = 0 \\ x_1 & EFFECTIF > 0 \end{cases}$$

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

1. Le cadre standard

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]).

1. Le cadre standard

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 :

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

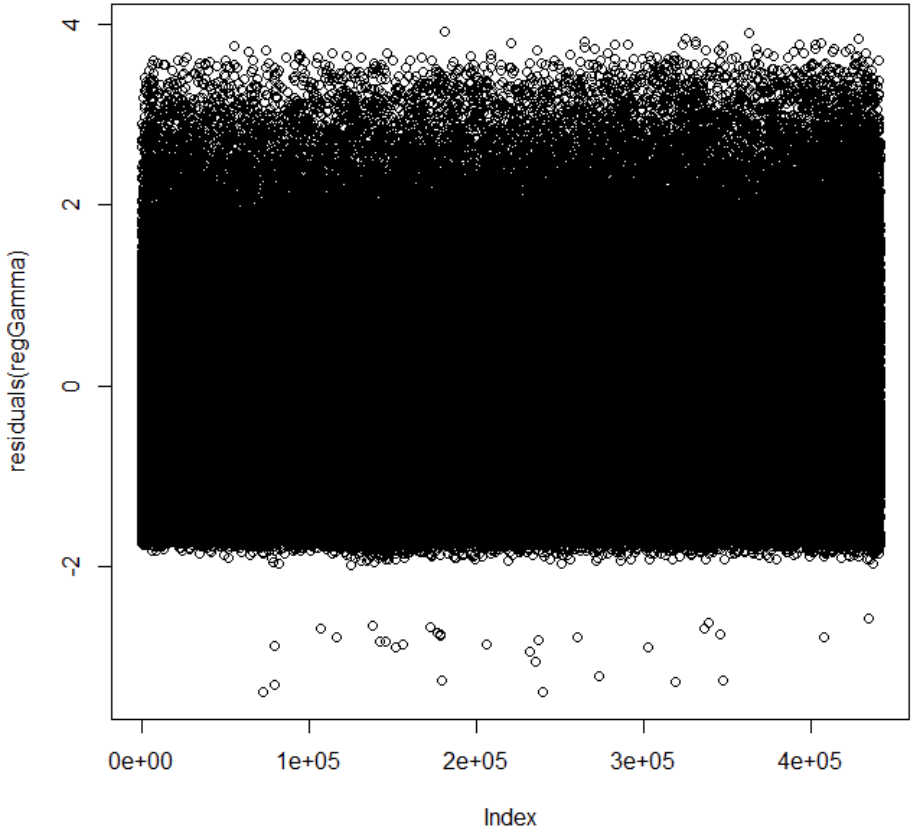
D est positif et « petit » pour un modèle de bonne qualité. Cette statistique suit asymptotiquement, du fait de résultats généraux sur les rapports de vraisemblance, une loi du Khi-2 à $n - p$ - 1 degrés de liberté (son espérance est donc $n - p - 1$).

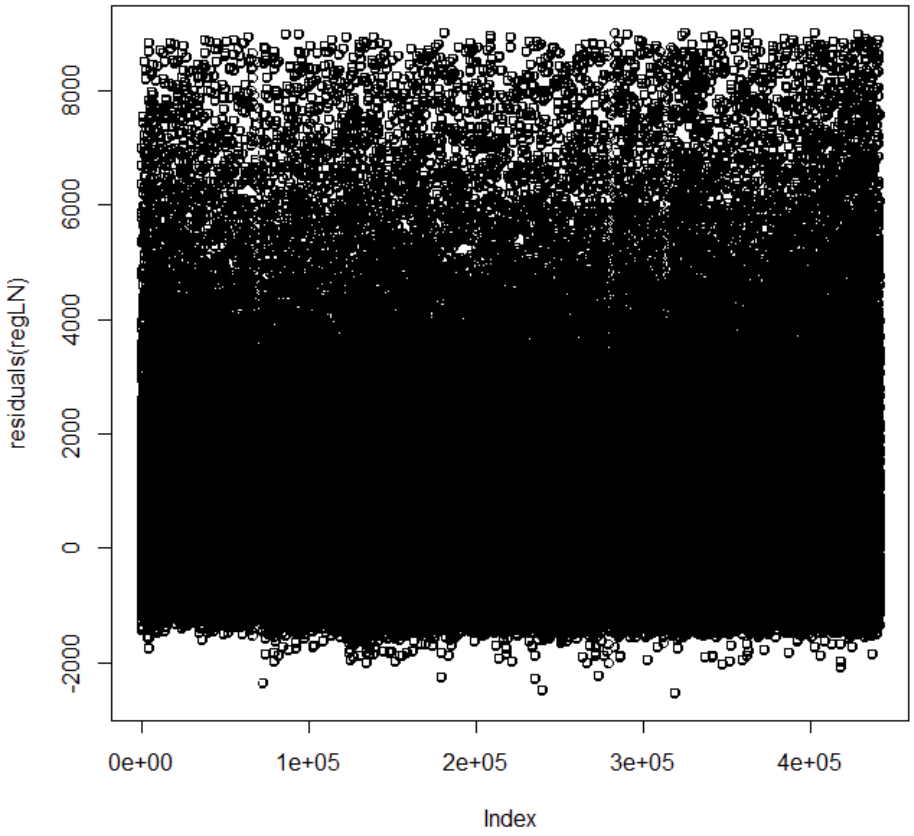
Cet indicateur global est en pratique complété par une analyse observation par observation ; cette analyse se base souvent sur l'analyse des résidus.

1. Le cadre standard

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) :





1. Le cadre standard

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.

1. Le cadre standard

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 :

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

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

1. Le cadre standard

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 : $E(l_i) = 0$

1. Le cadre standard

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
```

1. Le cadre standard

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
```

1. Le cadre standard

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
```

1. Le cadre standard

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 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
```

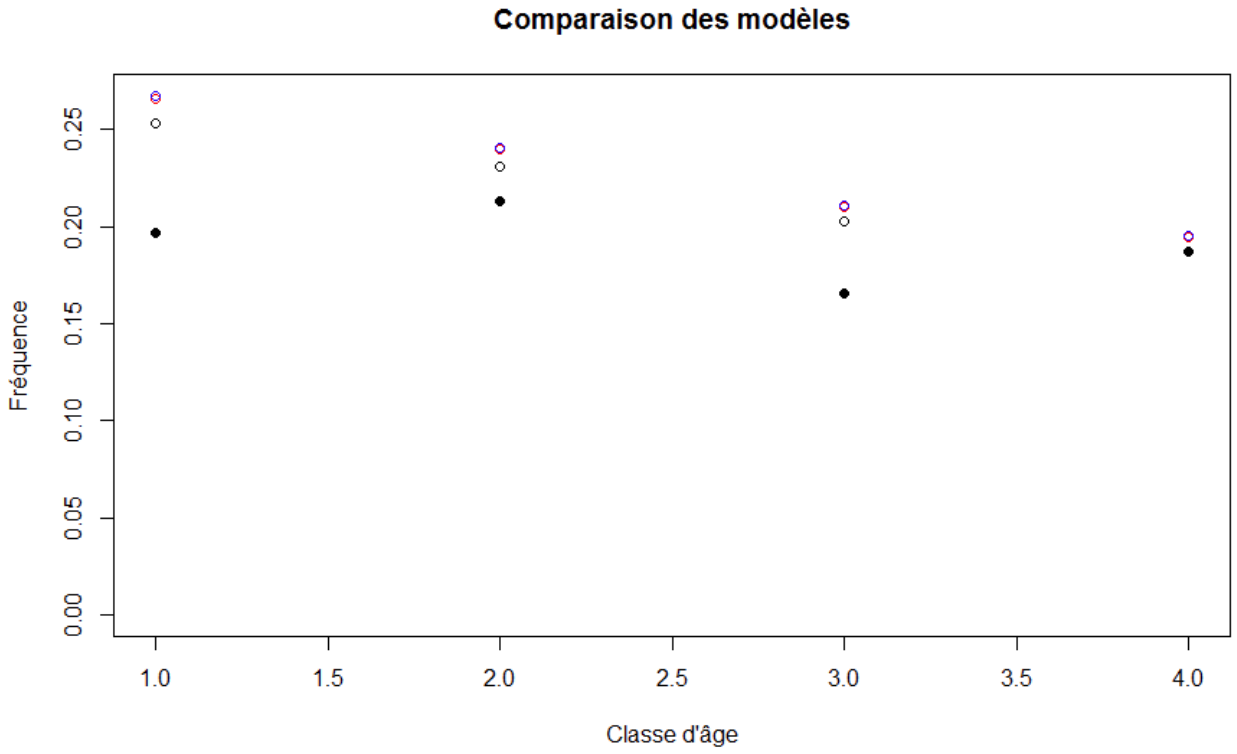
1. Le cadre standard

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.

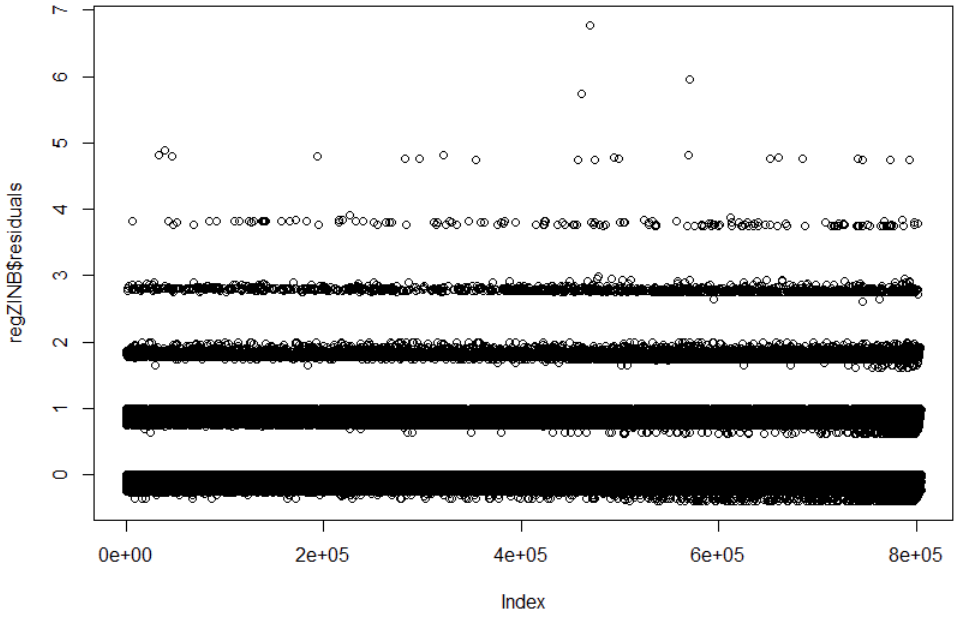
Les valeurs prédites sont assez éloignées des valeurs brutes…

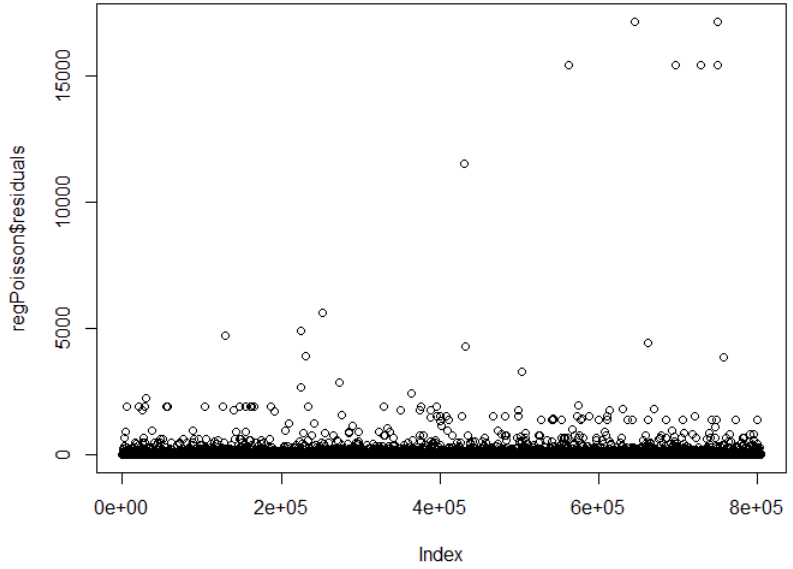


1. Le cadre standard

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 :





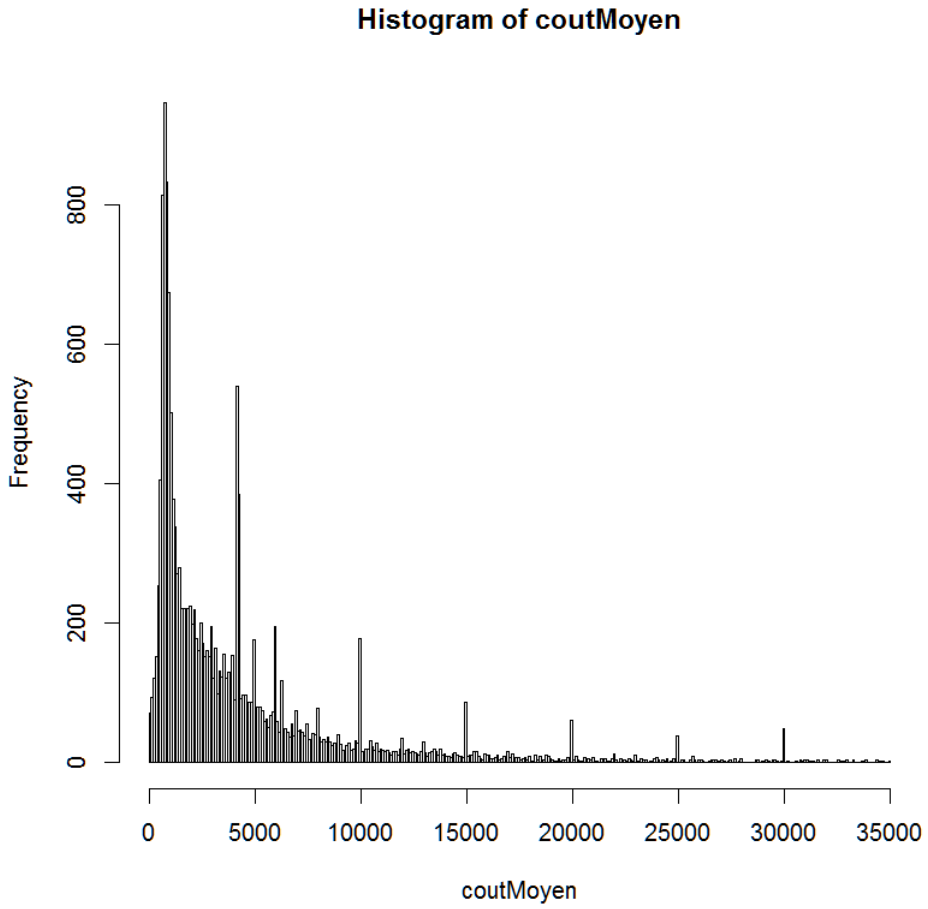
1. Le cadre standard

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€.



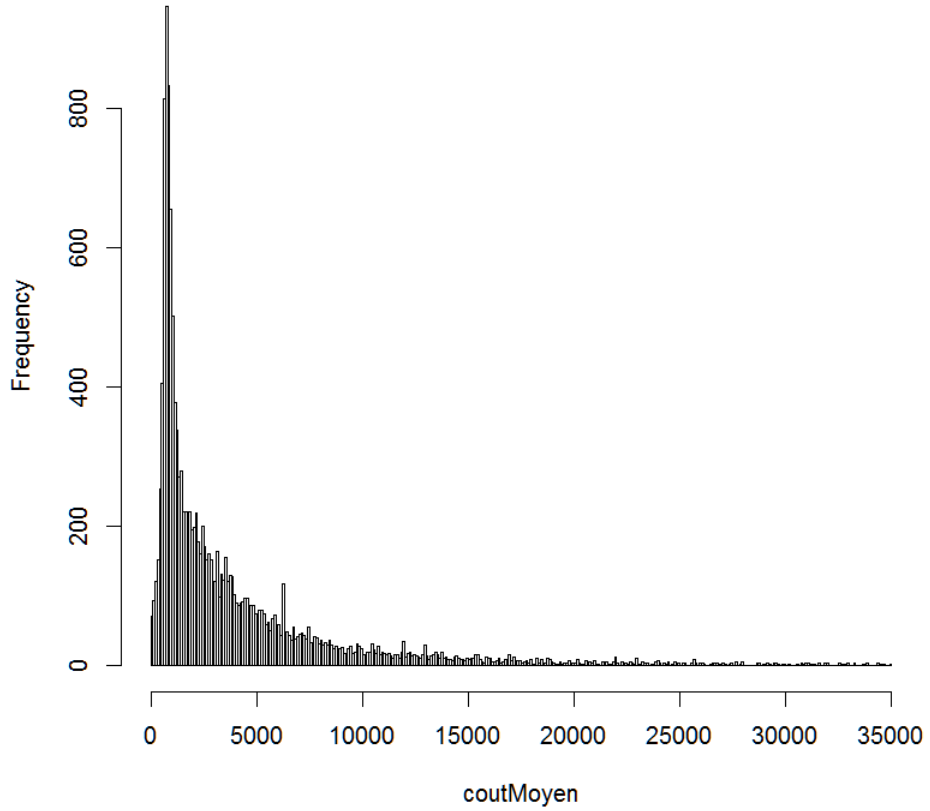
1. Le cadre standard

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

2. Les approches alternatives,

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].

2. Les approches alternatives

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.

2. Les approches alternatives

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.

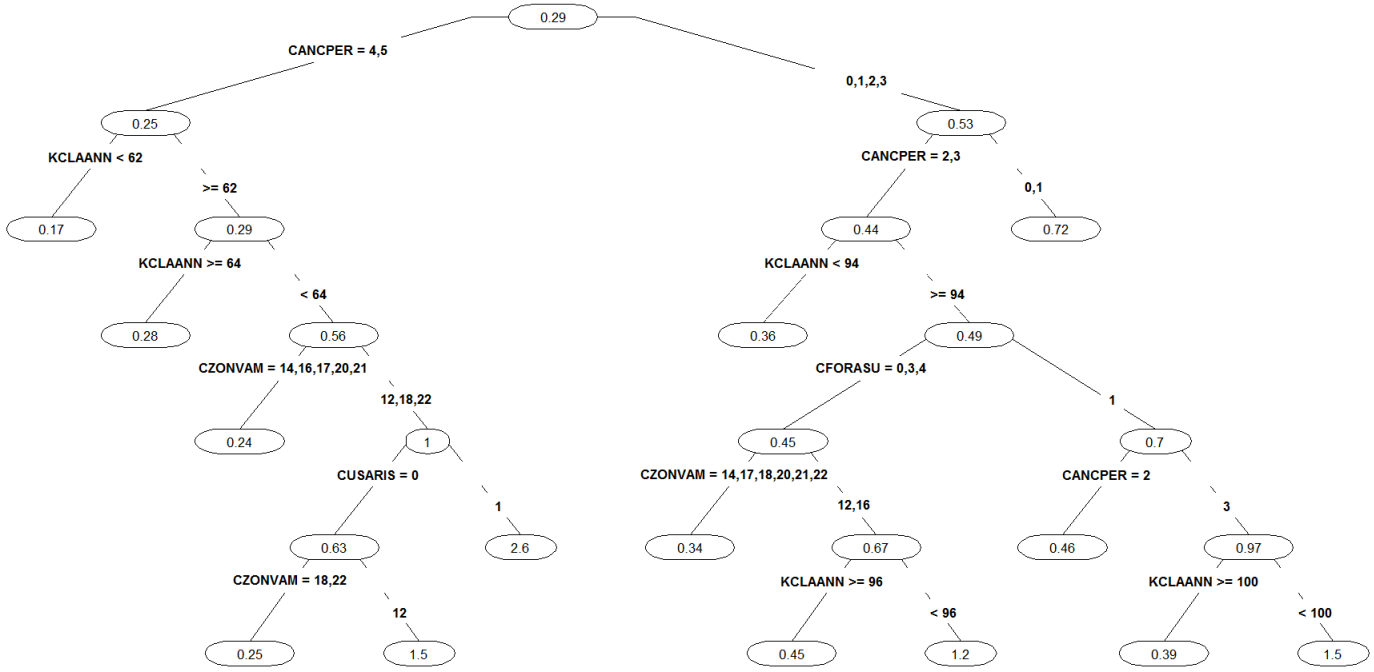
On doit disposer d'une base d'apprentissage et d'une base de validation.

2. Les approches alternatives

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 » :



2. Les approches alternatives

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.

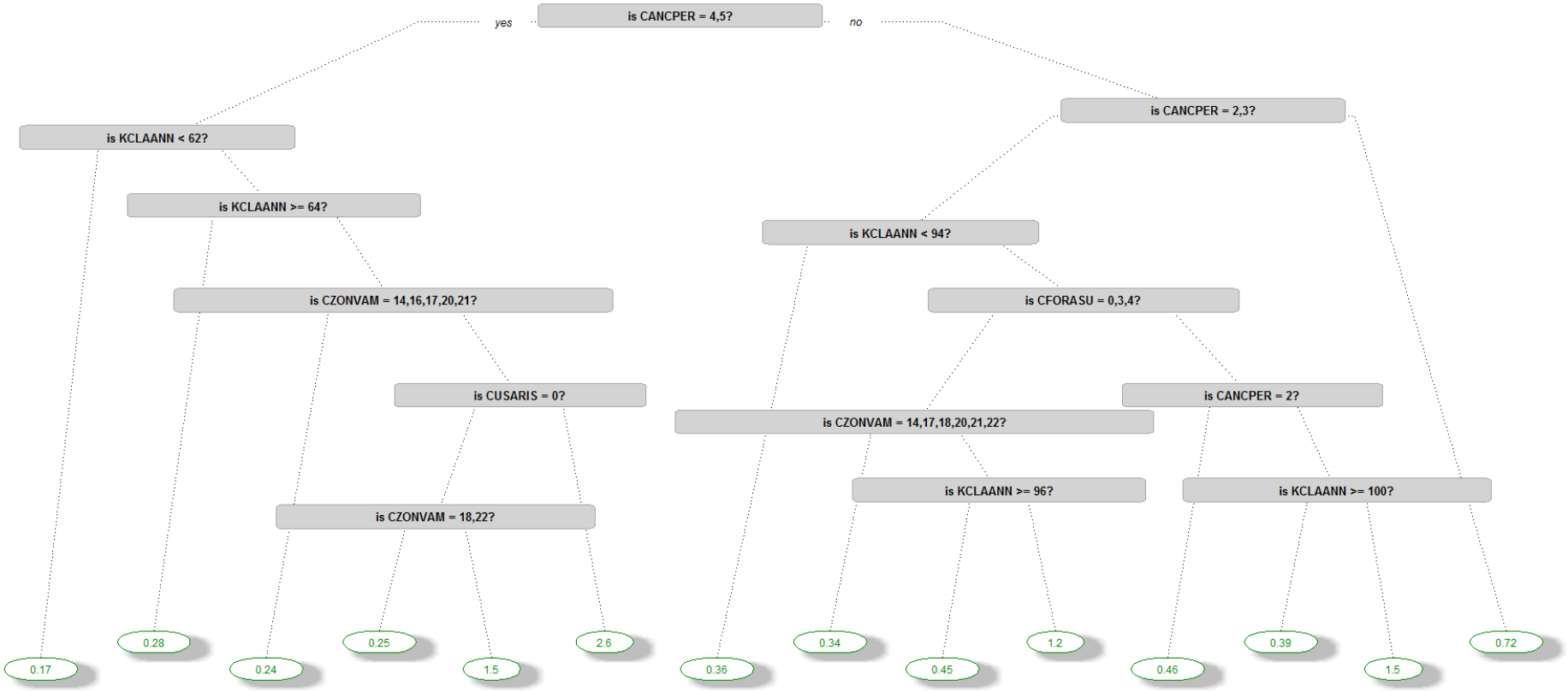
A l'intérieur de chaque nœud, la grandeur modélisée (fréquence ou coût moyen) est estimée par son espérance empirique.

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

2. Les approches alternatives

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

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



2. Les approches alternatives

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.

2. Les approches alternatives

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).

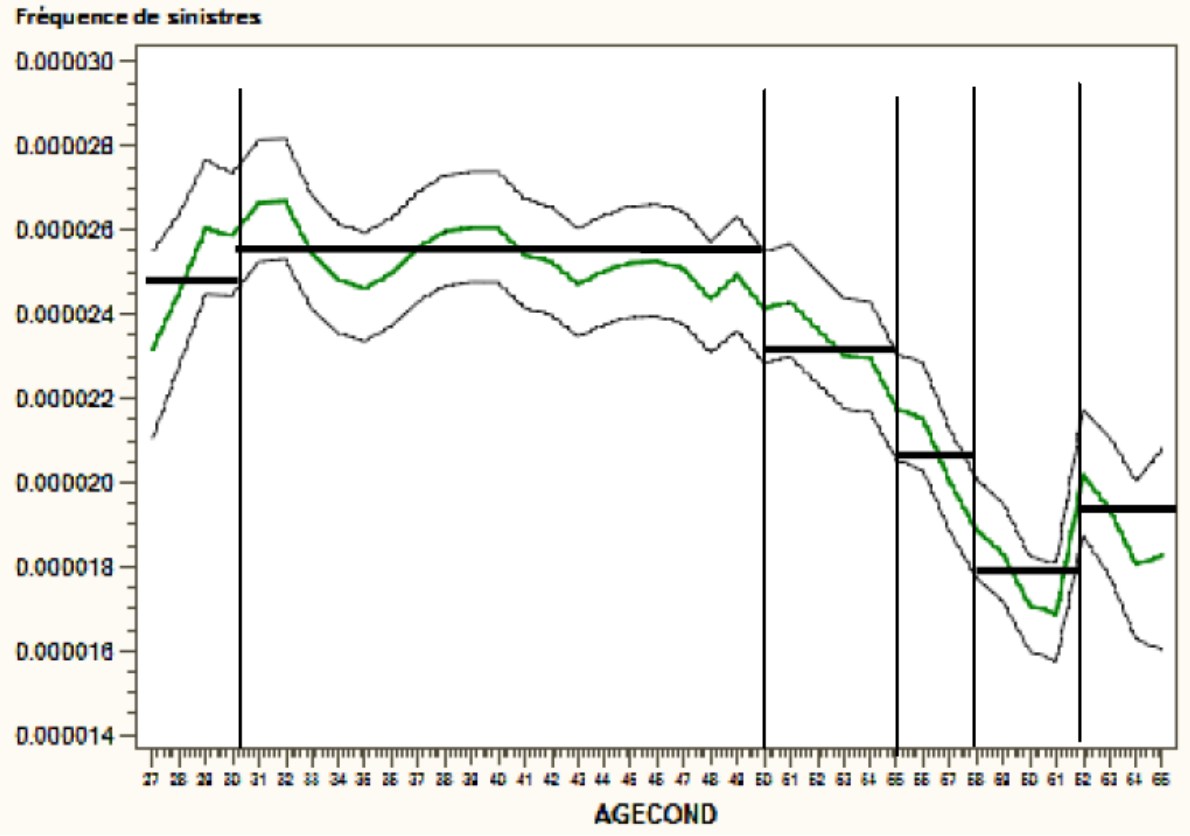
2. Les approches alternatives

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.



En guise de 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.

En guise de 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.

C'est le cadre des *big data*, qui donne lieu non seulement à des évolutions techniques mais aussi (surtout) à des évolutions des produits.

Références bibliographiques

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

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*.

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

R Development Core Team [2013] R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.R-project.org

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.

Contacts

Frédéric PLANCHET

frederic@planchet.net

Guillaume SERDECZNY

guillaume.serdeczny@maif.fr

Prim'Act

42 avenue de la Grande Armée

F - 75017 Paris

+33-1-42-22-11-00

MAIF

200, avenue Salvador Allende

F - 79038 Niort Cedex 09

+33-5-49-73-74-89

http://www.primact.fr – http://www.maif.fr

http://www.ressources-actuarielles.net

http://blog. ressources-actuarielles.net