Modèles de régression à directions révélatrices Applications en assurance non-vie

Eun Yung Kim, Michel Delecroix

MODELES DE REGRESSION A DIRECTIONS REVELATRICES APPLICATION EN ASSURANCE NON-VIE

Eun Jung KIM[1,2] & Michel DELECROIX[1]

[1] *Laboratoire de Statistique Théorique et Appliquée, Université Paris VI, Paris, France*
[2] *Laboratoire de Statistique, CREST-INSEE, Malakoff, France*
E-mail : Eun-Jung.KIM@ensae.fr

Résumé

L'usage des modèles linéaires généralisés est universel en tarification automobile. Les modèles utilisés supposent connue la forme de la loi de la variable expliquée Y(occurence des sinistres, fréquence des sinistres et coûts des sinistres) et supposent qu'elle est reliée au vecteur des régresseurs X par une relation du type: $E[Y|X = x] = g(x \cdot \beta)$ ($\cdot$ représente le produit scalaire) pour une fontion g réelle spécifiée. Nous proposons d'élargir cette approche en utilisant des modélisations semi-paramétriques où la seule hypothèse est :

$$\exists \beta_1, ..., \beta_k (k \geq 1), \exists g : E[Y|X = x] = g(x \cdot \beta_1, ..., x \cdot \beta_k)$$

où g n'est pas spécifiée. En employant la méthode « rMAVE »(en anglais, refined Minimum Average Variance Estimation) pour la détermination de k, des β_i et g, on concrétise cette approche sur des données simulées et des jeux de données classiques en assurance non-vie et on compare les résultats obtenus avec ceux que donne le modèle linéaire généralisé.

Mots clés : modèle semi-paramétrique, réduction de dimension par directions révélatrices, données de comptage, assurance non-vie

Abstract

Generalized linear models are universally used in pricing vehicle insurance. The used models assume a known distribution for the response variable Y (occurence of claims, claim frequency and claim costs) and assume that it is related to a vector of regressors X by a relationship : $E[Y|X = x] = g(x \cdot \beta)$ ($\cdot$ signifies the scalar product) for a specified real function g. We extend this approach by using a semiparametric modelization where the only assumption is :

$$\exists, \beta_1, ..., \beta_k (k \geq 1), \exists g : E[Y|X = x] = g(x \cdot \beta_1, ..., x \cdot \beta_k)$$

where g is not specified. By using the method « rMAVE »(refined Minimum Average Variance Estimation) to determine k, β_i and g, we motivate this approach by simulation argument and an application to classical real data in non-life insurance and compare the obtained results with those given by a generalized linear model.

Keywords : semiparametric model, dimension reduction by multiple index, count data, non-life insurance

1 INTRODUCTION

La théorie des modèles linéaires généralisés (MLG) a été largement développée par McCullagh et Nelder (1989). En tarification automobile, elle est utilisée largement pour analyser l'occurence des accidents des conducteurs, les fréquences de ces accidents ou les côuts des sinistres.

Dans tous les cas, il s'agit de modéliser l'influence d'un vecteur des variables explicatives $X \in R^p$ sur une variable expliquée $Y \in R$ en estimant la fonction de régression m. Les MLG supposent connue la forme de la loi de la variable Y qui appartient à la famille exponentielle et supposent qu'elle est dépendante du vecteur des régresseur $X \in R^p$ par une relation du type:

$$\exists \beta \in R^p, \quad E[Y|X = x] = m(x) = g(x \cdot \beta) \tag{1}$$

où $(\cdot)$ représente le produit scalaire de R^p et g une fonction réelle spécifiée. Nous proposons d'élargir cette approche en estimant m sous la *seule* hypothèse:

$$\exists \beta_1, ..., \beta_k, (k \geq 1), \quad E[Y|X = x] = g(x \cdot \beta_1, ..., x \cdot \beta_k) \tag{2}$$

où g n'est pas spécifiée. On impose une hypothèse d'orthonormalité sur les β_i pour éviter les problèmes d'identifiabilté (Ichimura and Lee, 1991). C'est à dire qu'on suppose finalement l'existence d'une matrice $B = (\beta_1, ..., \beta_k)$ de dimension pxk telle que :

$$E[Y|X = x] = g(B'x), B'B = Id_k \tag{3}$$

(B' représente la transposée de B.)

Pour estimer la valeur de k, les paramètres β_i, et la fonction de régression m, nous employons la méthode « rMAVE »(en anglais, refined Minimum Average Variance Estimation) introduite par Xia, Li, Tong et Zhang (2002). Cette méthode sera développée au paragraphe 2. La trosième partie sera consacrée à la présentation des données étudiées. Ces données résultent de simulations, pour une part, et proviennent d'un fichier d'assurance automobile d'autre part. Dans la quatrième partie, nous comparerons les résultats obtenus sur ces données par la méthode rMAVE et la méthode MLG, pour conclure enfin.

2 ALGORITHMES MAVE et rMAVE

On dispose d'un échantillon $(X_i, Y_i)_{i=1,...,n}$ observé, de vecteurs aléatoires indépendants et de même loi, où $X_i \in R^p$ et $Y_i \in R$. On suppose vérifiée l'hypothèse (3) pour les (X_i, Y_i). Pour estimer B, en supposant k connue, on remarque que, puisque :

$$E[(Y_i - m(X_i))^2] = \inf_h E[(Y_i - h(X_i))^2]$$

on a :

$$E[(Y_i - g(B'X_i))^2] = \inf_h E[(Y_i - h(X_i))^2] \tag{4}$$

Appelons alors $g_{\overline{B}}$ l'espérance conditionnelle de Y_i par rapport à $\overline{B}'X_i$ pour toute matrice $\overline{B}$, de dimension pxk :

$$g_{\overline{B}}(u) = E[Y_i|\overline{B}'X_i = u]$$

On en déduit $g = g_{\overline{B}}$, et d'après (4) :

$$E[(Y_i - g_B(B'X_i))^2] = \underset{\overline{B}}{\text{Inf}}\, E[(Y_i - g_{\overline{B}}(\overline{B}'X_i))^2] \tag{5}$$

On aura a fortiori :

$$E[(Y_i - g_B(B'X_i))^2] = \underset{\overline{B}}{\text{Inf}}\, E[(Y_i - g_{\overline{B}}(\overline{B}'X_i))^2|X_i] \tag{6}$$

Si les fonctions $g_{\overline{B}}$ étant connues, un estimateur naturel de B serait donc :

$$\widetilde{B} = \underset{B}{\text{Argmin}} \sum_{i=1}^{n} \{\sum_{j=1}^{n} (y_j - g_{\overline{B}}(\overline{B}'x_j))^2 w(x_i, x_j)\} \tag{7}$$

en remplaçant le critère théorique (espérance) par un critére empirique deduit de l'échantillon observé $(x_i, y_i), 1 \leq i \leq n$ et l'espérance conditionnelle par un estimateur de type Nadaraya Watson avec :

$$w(x_i, x_j) = \frac{K_h(x_j - x_i)}{\sum_{j=1}^{n} K_h(x_j - x_i)} \qquad K_h(u) = K(\frac{u}{h})$$

où K est un "noyau de Parzen-Rosenblatt" sur R^p.

Comme $g_{\overline{B}}$ est inconnue, on remarque que dans les sommes en j intervenant en (7) seules les observations x_j proches de x_i, interviennent réellement, compte tenu des poids $w(x_i, x_j)$. Pour ces observations, en effectuant un développement limité au premier ordre, on peut écrire :

$$y_j - g_{\overline{B}}(\overline{B}'x_j) \sim y_j - a_i - b_i(\overline{B}'(x_j - x_i)) \tag{8}$$

où $a_i = g_{\overline{B}}(\overline{B}'x_i)$ et b_i est le vecteur des dérivées partielles de $g_{\overline{B}}$, calculé en $\overline{B}x_i$.
On peut donc admettre que, en posant :

$$(a_i, b_i) = \underset{a \in R, b \in R^p}{\text{Argmin}} \sum_{j=1}^{n} [y_j - a - b(\overline{B}'(x_j - x_i))]^2 w(x_i, y_j) \tag{9}$$

l'équation (7) fournit finalement un estimateur plausible de B, $\widehat{B}$ défini par :

$$\widehat{B} = \underset{\overline{B}}{\text{Argmin}} \sum_{i=1}^{n} \{\sum_{j=1}^{n} [y_j - a_i - b_i(\overline{B}'(x_i - x_j))]^2 w(x_i, x_j)\} \tag{10}$$

La méthode MAVE consiste à utiliser (9) et (10) simultanément. Pour $\overline{B}$ donné, on détermine les a_i et b_i pour $1 \leq i \leq n$, par (9), puis, on détermine $\widehat{B}$ par (10).
$\widehat{B}$ obtenu on définit une nouvelle estimation des a_i,b_i par (9), et (10) permet alors une ré-estimation de B, etc. Xia et al.(2002) ont démontré la convergence de l'algorithme et de l'estimateur ainsi défini.

La méthode rMAVE permet un raffinement menant à une plus grande efficacité algorithmique : en modifiant dans (6) le conditionnement utilisé, on voit que (7) peut se réécrire en :

$$\widetilde{B} = \underset{\overline{B}}{\text{Argmin}} \sum_{i=1}^{n} \{ \sum_{j=1}^{n} E[(Y_j - g_{\overline{B}}(\overline{B}' X_i)^2 | \overline{B}' X_i] \}$$

d'où un estimateur qui s'écrit :

$$\widehat{\widehat{B}} = \underset{\overline{B}}{\text{Argmin}} \sum_{i=1}^{n} \{ \sum_{j=1}^{n} [y_j - a_i - b_i(\overline{B}'(x_i - x_j))]^2 w(\overline{B}' x_i, \overline{B}' x_j) \} \tag{11}$$

Au départ des itérations, on utilise les poids $w(x_i, x_j)$ pour définir une première matrice estimée qu'on introduit ensuite dans le critère. Cette idée a été proposée sous une autre forme par Hristache, Juditsky, Polzehl et Spokoiny (2001) pour estimer B à partir de l'estimateur du gradient de g.

Le nombre de directons k se détermine par la technique de validation croisée. Si $\widehat{m}_k$ est l'estimateur de m obtenu par la méthode précédente pour une matrice B de dimension pxk, on compare les sommes des carrés des résidus :

$$\widehat{k} = \underset{k}{\text{Argmin}} \sum_{i=1}^{n} (Y_i - \widehat{m}_k(X_i))^2)$$

3 LES JEUX DE DONNEES

1) Données simulées

Le modèle de régression considéré est défini par :

$$E(Y_i | X_i) = \frac{|\beta_1' X_i|}{(\beta_2' X_i)^4 + 1}$$

où $X_i = (X_{i1}, X_{i2}, X_{i3}, X_{i4}, X_{i5}, X_{i6}, X_{i7})'$ avec : $X_{i1}, X_{i2}, X_{i5} \sim N(0,1)$, $X_{i3} \sim B(0.2)$, $X_{i4}, X_{i6} \sim B(0.5)$, $X_{i7} \sim B(0.7)$ et la loi de Y_i conditionnellement à X_i est une loi de Poisson. Les paramètres sont $\beta_1 = (1, 4, 2, 3, 0, 0, 0)/\sqrt{30}$ et $\beta_2 = (0, 0, 0, 0, 1, -10, -20)/\sqrt{501}$.

On a simulé des échantillons de taille n, pour n=100, n=200 et n=400. Pour chaque taille, on génére 100 échantillons. Les observations ($n = 100, 200, 400$) sont divisées en deux groupes. La moitié des observations ($n = 50, 100, 200$) sert à estimer les paramètres et le reste est utilisé pour ajuster le modèle à partir des estimateurs des paramètres obtenus.

2) Jeu de données en assurance automobile

Ces donnèes sont accessible en ligne (www.act.mq.eud.au/GLMsforInsuranceData). Les 65856 polices sont observées en 2004 ou en 2005. Afin de diminuer le temps de calcul, on prélève un échantillon de taille n=10000, en utilisant la procédure " Surveyselect " de SAS.

La distribution des accidents sur la population totale, ou l'échantillons prélèvé, est la suivante :

Nombre des accidents	0	1	2	3	Total
Nombre de polices obs.	9343	609	44	4	65856
Pourcentage	93.43	6.09	0.44	0.04	100

TAB. 1 – Nombre d'accidents observés dans la population

Le but de l'étude est d'analyser le nombre des accidents Y en utilisant deux types de variable explicative, les trois variables (X_1, X_2, X_3) relatives au véhicule assuré d'une part et les deux variables (X_4, X_5) relatives à l'assuré d'autre part. Les variables considérées sont :

Variable	Définition	Moyenne	Écart-type
X_1	valeur du véhicule : \$0-\$23 (en dix millions)	1.791	1.241
X_2	caractère du véhicule : bus(1), convertible(2), coupé(3), à rayon arrière(4), coupé convertible (5), caravane ou combi motorisé(6), minibus(7), van(8), roadster(9), berline(10), pick-up(11), camion(12), utilitaire(13)	8.568	3.212
X_3	âge du véhicule : 1(le plus récent), 2, 3, 4	2.663	1.077
X_4	âge de l'assuré : 1(le plus jeune), 2, 3, 4, 5, 6	3.468	1.417
X_5	résidence de l'assuré : A(1), B(2), C(3), D(4), E(5), F(6)	2.791	1.432

TAB. 2 – Description des variables

4 RESULTATS

On compare la méthode rMAVE à l'utilisation des "GLM" sur les deux jeux de données. La méthode rMAVE se révèle aussi performante sans être décisivement meilleure que la téchnologie GLM.

Bibliographie

[1] Denuit M. et Charpentier A. (2005) *Mathématiques de l'assurance non-vie Tome 2 : Tarification et provisonnement*, Economica, Cambridge University Press.
[2] De Jong P. et Heller G. J. (2008) *Generalized linear models for insurance data*, International Series on Actuarial Science, Cambridge University Press.
[3] Foncel,J., Hristache,M. et Patilea,V. (2004) *Semiparametrice single-index poisson regression model with unobserved heterogeneity*, Série des documents de travil du CREST, INSEE.
[4] Hristache,M., Juditsky,A., Polzehl,J. et Spokoiny,V. (2001) *Structure adaptive approach for dimension reduction*, Ann.Statist.29, no.6, 1537-1566.
[5] Ichimura, H. et Lee, L.F. (1991) *Semiparametric least squares estimation of multiple index models : single equation estimation*, Cambridge University Press, 3-50.
[6] McCullagh, P. et Nelder, J.A.,(1989) *Generalized linear models*, second ed., Chapman and Hall, London.
[7] Xia, Y., Li, W.K., Tong, H. et Zhang, D. (2002) *An adaptive estimation of dimension reduction space*, J. Roy, Statist. Soc. Ser. B 64, 363-410.