REVUE DE STATISTIQUE APPLIQUÉE

FLORIAEN E. C. DE VYLDER
MARC J. GOOVAERTS

Estimation de la variance, dans un modèle classique, si les coefficients d'aplatissement des variables sont connus

Revue de statistique appliquée, tome 41, n° 3 (1993), p. 5-20
<http://www.numdam.org/item?id=RSA_1993__41_3_5_0>

Rev. Statistique Appliquée, 1993, **XLI** (3), 5-20

ESTIMATION DE LA VARIANCE, DANS UN MODÈLE CLASSIQUE, SI LES COEFFICIENTS D'APLATISSEMENT DES VARIABLES SONT CONNUS

Floriaen E.C. De Vylder (1), Marc J. Goovaerts (2)

(1) Université Cath. de Louvain, Place des Doyens 1
1348 Louvain-la-Neuve, Belgique
(2) Katholieke Universiteit Leuven de Bériotstraat 34
3000 Leuven, België

RÉSUMÉ

Le coefficient d'aplatissement de la variable non dégénérée X est défini par

$$e(X) = (EY^4/E^2Y^2) - 3 \quad \text{où} \quad Y = X - EX.$$

Le modèle statistique considéré est défini par les variables indépendantes $X_1, \ldots, X_n$ telles que

$$EX_i = m, \quad \text{Var } X_i = s^2/w_i, \quad e(X_i) = e_i \quad (i = 1, \ldots, n).$$

Si m est connu, l'estimateur non biaisé de variance minimum de s^2 dans la famille $\left\{ \sum c_i(X_i - m)^2 / c_1, \ldots, c_n \in \mathbb{R} \right\}$ est

$$\widehat{s}^2 = c \sum \frac{w_i}{2 + e_i}(X_i - m)^2 \quad \text{où} \quad \frac{1}{c} = \sum \frac{1}{2 + e_i}.$$

Si m n'est pas connu, un estimateur optimal de s^2 se laisse dériver, mais il est trop compliqué pour être pratique et nous en proposons une variante.

L'estimateur classique de s^2 est optimal si les e_i sont nuls, donc en particulier si les variables X_i sont normales et de façon plus générale si les X_i ont même loi (auquel cas les e_i sont égaux, de même que les w_i).

Mots clés : *Estimateur, Variance minimum, Non biaisé, Coefficient d'aplatissement, Théorie de la crédibilité.*

SUMMARY

The kurtosis of the non degenerated random variable X is defined to be

$$e(X) = (EY^4/E^2Y^2) - 3 \quad \text{where} \quad Y = X - EX.$$

The considered statistical model is defined by the random variables $X_1, \ldots, X_n$ satisfying

$$EX_i = m, \quad \text{Var}\, X_i = s^2/w_i, \quad e(X_i) = e_i \ (i = 1, \ldots, n).$$

When m is known, the minimum-variance unbiased estimator of s^2 in the family $\left\{ \sum c_i (X_i - m)^2 / c_1, \ldots, c_n \in \mathbb{R} \right\}$ equals

$$\widehat{s}^2 = c \sum \frac{w_i}{2 + e_i} (X_i - m)^2 \quad \text{where} \quad \frac{1}{c} = \sum \frac{1}{2 + e_i}.$$

If m is non known, an optimal estimator for s^2 can be found, but it is too complicated to be practical and we suggest a variant of it. The classical estimator for s^2 is optimal if the kurtosisses vanish, thus certainly when the random variables are normally distributed.

Key Words : *Estimator, Minimum variance, Unbiased, Kurtosis, Credibility theory.*

1. Coefficient d'aplatissement d'une variable aléatoire

Soit X une variable non dégénérée en un point, avec moment d'ordre quatre fini. Le coefficient d'aplatissement $e(X)$ de X est défini par

$$e(X) = \frac{EY^4}{E^2Y^2} - 3, \text{ où } Y = X - EX.$$

Ce coefficient d'aplatissement n'est autre que le coefficient γ_2 de Fisher.

On a

$$e(aX + b) = e(X) \quad (a \neq 0),$$
$$e(X) = \frac{EY^4 - E^2Y^2}{E^2Y^2} - 2 = \frac{\text{Var}\, Y^2}{\text{Var}^2 Y} - 2 \geqslant -2.$$

La valeur extrême -2 est atteinte si et seulement si X prend 2 valeurs distinctes, chacune avec la probabilité $1/2$. En effet :

$$e(X) = -2 \Leftrightarrow$$
$$\text{Var}\, Y^2 = 0 \Leftrightarrow$$
$$\text{Il existe } c > 0 \text{ telle que } Y^2 = c \text{ p.s.} \Leftrightarrow$$
$$\text{Il existe } c > 0 \text{ telle que } P(Y = -c^{1/2}) = \tfrac{1}{2}$$
$$\text{et } P(Y = +c^{1/2}) = \tfrac{1}{2}.$$

Les variables normales ont un coefficient d'aplatissement nul. En quelque sorte la distribution normale sert comme référence lorsque des excès sont considérés.

Quelques autres coefficients d'aplatissement sont mentionnés ci-dessous.

Si N est une Poisson de paramètre λ :

$$e(N) = \frac{1}{\lambda}.$$

Si N est binomiale de paramètres n, p :

$$e(N) = \frac{1 - 6pq}{npq} \text{ (où } q = 1 - p)$$

Si N est uniforme sur $\{0, 1, \ldots, n\}$:

$$e(N) = -\frac{6}{5}\Big(1 + \frac{2}{n(n+2)}\Big)$$

Si X est uniforme sur $[0, a]$:

$$e(X) = -\frac{6}{5}$$

Si X est gamma avec densité $\frac{1}{\Gamma(a)}\, x^{a-1}e^{-x}$ $(a > 0)$ sur $\mathbb{R}_+$, alors :

$$e(X) = \frac{6}{a}.$$

Si X est Pareto avec densité $ax^{-(1+a)}$ $(a > 4)$ sur $[1, +\infty[$, alors

$$e(X) = \frac{3(a-2)(3a^2 + a + 2)}{(a-4)(a-3)a} - 6.$$

Si X est lognormale avec densité $\frac{1}{x\sigma(2\pi)^{1/2}}\exp\Big(-\frac{(\log x - m)^2}{2\sigma^2}\Big)$ sur $\mathbb{R}_+$, alors

$$e(X) = e^{4\sigma^2} + 2e^{3\sigma^2} + 3e^{2\sigma^2} - 6.$$

Le cas de la loi de Poisson montre déjà que le coefficient d'aplatissement peut prendre des valeurs arbitrairement grandes.

2. Espérance d'une forme homogène de degré quatre

Ci-dessous n est un entier fixé strictement positif. Les indices i, j, p, q prennent toujours toutes les valeurs $1, 2, \ldots, n$.

Théorème 1

Soient $Y_1, \ldots, Y_n$ des variables centrées, non dégénérées, indépendantes, avec un moment d'ordre quatre fini chacune. Notons

$$EY_i^2 = u_i, \quad e(Y_i) = e_i.$$

Alors

$$E(Y_iY_jY_pY_q) = \delta_{ij}\delta_{pq}u_iu_p + \delta_{ip}\delta_{jq}u_iu_j + \delta_{iq}\delta_{jp}u_iu_j + \delta_{ijpq}u_i^2e_i \qquad (*)$$

où δ_{ij} est le symbole de Kronecker et δ_{ijpq} le symbole prenant la valeur 1 si $i = j = p = q$ et la valeur 0 sinon.

$$E\left(\sum_{ijpq} a_{ijpq}Y_iY_jY_pY_q\right) = \sum_{ij}(a_{iijj} + a_{ijij} + a_{ijji})u_iu_j + \sum_i a_{iiii}u_i^2e_i. \quad (**)$$

Démonstration

On vérifie que la formule $(*)$ est vraie dans tous les cas possibles

$$(i = j = p = q), (i = j = p \neq q), \ldots$$

A titre d'exemple :

Pour $i = j = p = q$, $(*)$ est $EY_i^4 = 3u_i^2 + u_i^2\left(\frac{EY_i^4}{u_i^2} - 3\right)$.

Pour $i = j = p \neq q$, $(*)$ est $EY_i^3 \cdot EY_q = 0$.

Pour $i = j \neq p = q$, $(*)$ est $EY_i^2 \cdot EY_p^2 = u_iu_p$.

La formule $(**)$ résulte directement de la formule $(*)$.

3. Définition du modèle considéré

Nous considérons le modèle statistique défini par les variables indépendantes avec moment d'ordre quatre fini $X_1, \ldots, X_n$ telles que

$$EX_i = m, \ \text{Var}\, X_i = \frac{s^2}{w_i}, \ e(X_i) = e_i \quad (i = 1, \ldots, n).$$

Nous supposons $s^2 > 0$. Les variables X_i ne sont donc pas dégénérées. Nous supposons connus les poids $w_i > 0$ et les $e_i \geqslant -2$. Le problème est d'estimer m et s^2.

L'estimateur classique de m est

$$\widehat{m} = \sum \frac{w_i}{w_\bullet} X_i \quad \text{où} \quad w_\bullet = \sum w_i.$$

Il est bien connu, et on le vérifie facilement, qu'il est *optimal* dans la classe des estimateurs

$$\left\{\sum c_iX_i / c_1, \ldots, c_n \in \mathbb{R}\right\}.$$

Dans cette note, *optimal* est synomyme de *non biaisé de variance minimum.*

Le problème restant est donc l'estimation de s^2.

Nous supposerons d'abord que m est connu. Ce cas met bien en évidence le rôle joué par les coefficients d'aplatissement dans l'optimalité des estimateurs.

Nous notons

$$Y_i = X_i - EX_i = X_i - m,$$
$$u_i = \frac{1}{w_i}.$$

Dans les optimisations, nous appliquerons la méthode classique du multiplicateur de Lagrange.

4. Estimation de s^2 si m est connu

Cherchons l'estimateur optimal de s^2 dans la classe des estimateurs de la forme

$$S_m^2 = \sum c_i (X_i - m)^2 \quad \text{où} \quad c_1, \ldots, c_n \in \mathbb{R}.$$

On a

$$ES_m^2 = \sum c_i EY_i^2 = s^2 \sum c_i u_i$$

et l'absence de biais se traduit donc par la *contrainte* linéaire

$$\sum c_i u_i = 1.$$

Par la formule $(*)$ du théorème précédent, on a

$$\begin{aligned} ES_m^4 &= E\left(\sum_{ij} c_i c_j Y_i Y_i Y_j Y_j\right) \\ &= s^4 \sum_{ij} c_i c_j (\delta_{ii}\delta_{jj} u_\iota u_j + \delta_{\iota j}\delta_{\iota j} u_i u_i + \delta_{ij}\delta_{ij} u_\iota u_\iota + \delta_{iij_J} u_i^2 e_i) \\ &= s^4 \left(\sum_i c_\iota u_\iota \cdot \sum_j c_j u_j + 2 \sum_i c_\iota^2 u_\iota^2 + \sum_\iota c_i^2 u_\iota^2 e_\iota\right). \end{aligned}$$

D'où

$$\operatorname{Var} S_m^2 = s^4 \sum_\iota c_\iota^2 u_i^2 (2 + e_\iota).$$

qui est à minimiser sous la contrainte $\sum c_\iota u_\iota = 1$ sur les coefficients inconnus c_ι.

Supposons d'abord $2 + e_i \neq 0$ $(i = 1, \ldots, n)$. Soit -2λ le multiplicateur de Lagrange correspondant à la contrainte. Alors on a à rendre minimum

$$F = \sum c_i^2 u_i^2 (2 + e_i) - 2\lambda \sum c_i u_i.$$

L'annulation de la dérivée partielle en c_i fournit le système

$$c_i u_i^2 (2 + e_i) - \lambda u_i = 0.$$

D'où, car $u_i \neq 0$, $c_i = \frac{\lambda w_i}{2 + e_i}$. Puis $\lambda \sum \frac{1}{2 + e_i} = 1$ par la contrainte.

Considérons ensuite le cas où un e_i, par exemple e_1, égale -2. Alors on vérifie que l'estimateur optimal de s^2 est

$$\hat{s}^2 = w_1 (X_1 - m)^2, \quad \text{avec} \quad \text{Var}(\hat{s}_m^2) = 0.$$

On a alors :

Théorème 2

L'estimateur optimal de s^2 dans la classe

$$\left\{ \sum c_i (X_i - m)^2 / c_1, \ldots, c_n \in \mathbb{R} \right\}$$

est

$$\hat{s}_m^2 = c \sum \frac{w_i}{2 + e_i} (X_i - m)^2 \quad \text{où} \quad \frac{1}{c} = \sum \frac{1}{2 + e_i},$$

si tous les e_i dépassent strictement -2.

Si un coefficient d'aplatissement est égal à -2, soit e_1 par exemple, alors l'estimateur optimal est

$$\hat{s}_m^2 = w_1 (X_1 - m)^2$$

et il est de variance nulle.

Comparaison avec l'estimateur classique

L'estimateur classique de s^2, dans le cas m connu, est

$$\hat{s}_{m.cl}^2 = \frac{1}{n} \sum w_i (X_i - m)^2.$$

De l'expression générale de Var S_m^2 dérivée dans la démonstration précédente, résulte que

$$\operatorname{Var}\widehat{s}_m^2 = \frac{s^4}{\sum \frac{1}{2+e_i}}, \quad \operatorname{Var}\widehat{s}_{m,cl}^2 = \frac{s^4}{n^2}\sum(2+e_i).$$

On observe alors que :

a) L'estimateur classique $\widehat{s}_{m,cl}^2$ égale l'estimateur optimal $\widehat{s}_m^2$ si et seulement si $e_1 = \ldots = e_n$. En particulier, l'estimateur classique est optimal si les coefficients d'aplatissement sont nuls, ce qui est en particulier réalisé si les variables X_i sont normales.

b) Le rapport

$$\frac{\operatorname{Var}\widehat{s}_{m,cl}^2}{\operatorname{Var}\widehat{s}_m^2} \to \infty$$

si un coefficient d'aplatissement tend vers -2 ou vers l'infini.

c) Par définition de l'optimalité de $\widehat{s}_m^2$, on a

$$\operatorname{Var}\widehat{s}_m^2 \leqslant \operatorname{Var}\widehat{s}_{m,cl}^2.$$

Cette inégalité équivaut à

$$n^2 \leqslant \sum(2+e_i)\cdot\sum\frac{1}{2+e_i},$$

qui résulte aussi de l'inégalité de Schwarz

$$\left(\sum x_i y_i\right)^2 \leqslant \sum x_i^2 \cdot \sum y_i^2,$$

avec

$$x_i = (2+e_i)^{1/2}, \quad y_i = (2+e_i)^{-1/2}.$$

5. Estimation de s^2 si m est non connu

Problème

Nous cherchons l'estimateur optimal de s^2 de la forme

$$S^2 = \sum c_i(X_i - \widehat{m})^2 \quad \text{où} \quad c_1 \ldots . c_n \in \mathbb{R}$$

et où $\widehat{m}$ est l'estimateur classique de m.

Autre expression de S^2

On a

$$X_i - \widehat{m} = X_\imath - \sum \frac{w_p}{w_\bullet} X_p = \sum_p \frac{w_p}{w_\bullet}(X_i - X_p) = \sum_p \frac{w_p}{w_\bullet}(Y_i - Y_p).$$

Alors

$$\begin{aligned} S^2 &= \sum_{ipq} c_i \frac{w_p}{w_\bullet} \frac{w_q}{w_\bullet} (Y_i - Y_p)(Y_i - Y_q) \\ &= \sum_{ipq} c_i \frac{w_p}{w_\bullet} \frac{w_q}{w_\bullet} (Y_i^2 - Y_i Y_p - Y_\imath Y_q + Y_p Y_q) \\ &= \sum_\imath c_i Y_i^2 - 2 \sum_{ip} c_i \frac{w_p}{w_\bullet} Y_i Y_p + c_\bullet \sum_{pq} \frac{w_p}{w_\bullet} \frac{w_q}{w_\bullet} Y_p Y_q. \end{aligned}$$

D'où

$$S^2 = \sum_{ij} c_{\imath\jmath} Y_\imath Y_\jmath \quad \text{où} \quad c_{\imath\jmath} = c_\imath \delta_{ij} - 2c_\imath \frac{w_\jmath}{w_\bullet} + c_\bullet \frac{w_\imath}{w_\bullet} \frac{w_\jmath}{w_\bullet},$$

$$\text{et} \quad c_\bullet = \sum c_i.$$

La contrainte

De cette expression résulte que

$$ES^2 = \sum_\imath c_{\imath\imath} EY_i^2 = s^2 \sum c_{\imath\imath} u_\imath.$$

L'absence de biais se traduit donc par la contrainte linéaire en les inconnues $c_\imath$, $\sum c_{\imath\imath} u_\imath = 1$, soit

$$\sum c_\imath u_i - \frac{c_\bullet}{w_\bullet} = 1.$$

Variance de S^2

Par la formule $(**)$ du théorème 1, on a

$$\begin{aligned} ES^4 &= \sum_{\imath\jmath pq} c_{\imath\jmath} c_{pq} E(Y_\imath Y_\jmath Y_p Y_q) \\ &= s^4 \sum_{\imath\jmath} (c_{\imath\imath} c_{\jmath\jmath} + c_{\imath\jmath} c_{\imath\jmath} + c_{\imath\jmath} c_{\jmath\imath}) u_\imath u_\jmath + s^4 \sum_\imath c_{\imath\imath}^2 u_\imath^2 e_\imath \end{aligned}$$

Puisque

$$E^2S^2 = s^4\left(\sum_i c_{ii}u_i\right)^2 = s^4\sum_{ij} c_{ii}c_{jj}u_iu_j,$$

on a

$$\text{Var}\, S^2 = s^4(v + v_e)$$

où

$$\begin{aligned} v &= \sum_{ij}(c_{ij}c_{ij} + c_{ij}c_{ji})u_iu_j \\ &= 2\sum_i c_i^2u_i^2 - \frac{4}{w_\bullet}\sum_i c_i^2u_i + 2\,\frac{c_\bullet^2}{w_\bullet^2} \end{aligned}$$

et

$$\begin{aligned} v_e &= \sum_i c_{ii}^2u_i^2e_i \\ &= \sum_i \left(c_iu_i - 2\frac{c_i}{w_\bullet} + c_\bullet\frac{w_i}{w_\bullet^2}\right)^2 e_i. \end{aligned}$$

Estimateur optimal analytique de s^2

A partir de cette expression de Var S^2 on obtient facilement par la méthode du multiplicateur de Lagrange, le système linéaire en les c_i et en λ fournissant le minimum de Var S^2 sous la contrainte d'absence de biais. La solution unique de ce système se laisse alors expliciter au moyen de la dérivation matricielle par exemple. L'estimateur optimal ainsi obtenu est peu intéressant parce que trop compliqué.

Estimateur optimal numérique de s^2

Si les w_i et les e_i sont connus numériquement, on peut employer un algorithme numérique quelconque pour minimiser Var S^2 sous la contrainte d'absence de biais. En particulier, on pourrait résoudre numériquement le système linéaire mentionné ci-dessus, mais il existe des routines numériques générales qui évitent la considération de ce système et le calcul des dérivées partielles en les c_i.

En pratique, si n n'est pas trop grand, on peut donc déterminer l'estimateur optimal de s^2 dans chaque cas particulier si les w_i et les e_i sont connus.

Estimateur analytique approximativement optimal

A partir de poids relatifs connus p_i quelconques on obtient l'estimateur sans biais suivant de s^2 :

$$S^2 = c\sum_i p_i(X_i - \hat{m})^2 \quad \text{où} \quad \frac{1}{c} = \sum_i \frac{p_i}{w_i} - \sum_i \frac{p_i}{w_\bullet}$$

Tenant compte de ce que nous a appris le paragraphe précédent dans le cas m connu, nous prendrons $p_i = w_i/(2 + e_i)$.

L'estimateur approximativement optimal de s^2, mais exactement sans biais, ainsi obtenu est

$$\tilde{s}^2 = c \sum_i \frac{w_i}{2 + e_i} (X_i - \hat{m})^2 \text{ où } \frac{1}{c} = \sum_i \frac{1}{2 + e_i} - \frac{1}{w_\bullet} \sum_i \frac{w_i}{2 + e_i}.$$

Cas des coefficients d'aplatissement nuls

Si les e_i sont tous nuls, l'estimateur optimal analytique se laisse expliciter facilement. En effet, par la méthode du multiplicateur de Lagrange, on a alors à minimiser

$$F = w_\bullet^2 \sum u_i^2 c_i^2 - 2w_\bullet \sum u_i c_i^2 + \left(\sum c_i\right)^2 - 2\lambda(w_\bullet \sum c_i u_i - \sum c_i - w_\bullet).$$

L'annulation de la dérivée partielle en c_i fournit le système linéaire

$$w_\bullet^2 u_i^2 c_i - 2w_\bullet u_i c_i + c_\bullet - \lambda(w_\bullet u_i - 1) = 0 \quad (i = 1, \ldots, n).$$

On vérifie que ce système et la contrainte de non-biais sont satisfaits pour

$$\lambda = \frac{w_\bullet}{n - 1}, \quad c_i = \frac{w_i}{n - 1} \quad (i = 1, \ldots, n).$$

Cette solution fournit le minimum absolu cherché. Nous avons donc démontré la théorème suivant.

Théorème 3

Si $e_1 = \ldots = e_n = 0$, l'estimateur optimal de s^2 dans la classe

$$\left\{\sum c_i (X_i - \hat{m})^2 / c_1, \ldots, c_n \in \mathbb{R}\right\}$$

est l'estimateur classique

$$\hat{s}_{cl}^2 = \frac{1}{n - 1} \sum w_i (X_i - \hat{m})^2.$$

Cet estimateur est alors identique à l'estimateur approximativement optimal $\tilde{s}^2$ considéré ci-dessus.

Remarque

Dans le cas des coefficients d'aplatissement nuls, on peut montrer que $\widehat{s}^2_{cl}$ est optimal dans la classe, beaucoup plus large, des estimateurs de la forme

$$\sum_{ijpq} a_{ijpq}(X_i - X_j)(X_p - X_q).$$

Ce résultat semble difficile à obtenir par la méthode des multiplicateurs de Lagrange, car les n^4 inconnues a_{ijpq} ne sont pas seulement liées par la contrainte d'absence de biais, mais par des contraintes de symétrie telles que $a_{ijpq} = a_{jiqp}$ par exemple.

Le résultat indiqué est obtenu dans De Vylder & Goovaerts (April 1992) par des méthodes de projection orthogonale dans l'espace de Hilbert des variables de carré intégrable sur un espace de probabilité fixé.

6. Exemple d'application

Coefficient d'aplatissement d'une moyenne arithmétique

Soit

$$X = \frac{1}{k}\,(U_1 + \ldots + U_k)$$

où les variables U_i sont i.i.d., chacune avec espérance m, variance s^2 et coefficient d'aplatissement e. Alors

$$EX = m, \quad \text{Var}\, X = \frac{s^2}{k}, \quad e(X) = \frac{e}{k}.$$

En effet, pour démontrer la dernière relation, posons $Y = X - EX$, $V_i = U_i - EU_i$ $(i = 1, \ldots, k)$. Par la formule $(**)$ du théorème 1, on a alors

$$\begin{aligned} EY^4 &= \frac{1}{k^4} \sum_{ijpq} E(V_i V_j V_p V_q) \\ &= \frac{1}{k^4} \left(\sum_{ij} 3s^4 + \sum_{i} s^4 e \right) = \frac{1}{k^4}(3k^2 + ke)s^4. \end{aligned}$$

Puisque $EY^2 = s^2/k$, on a la formule annoncée.

Définition du modèle considéré

Supposons que les variables $X_1, \ldots, X_n$ soient telles que

$$X_i = \frac{1}{k_i}\,(U_{i1} + U_{i2} + \ldots + U_{ik_i}) \quad (i = 1, \ldots, n)$$

où les U_{ij} sont i.i.d., chacune avec espérance m, variance s^2 et coefficient d'aplatissement e.

Alors les variables $X_1, \ldots, X_n$ sont indépendantes et

$$EX_i = m, \quad \text{Var}\, X_i = \frac{s^2}{k_i}, \quad e(X_i) = \frac{e}{k_i} \quad i = 1, \ldots, n).$$

Nous supposons que les variables X_i sont observables et qu'elles ont la structure indiquée, mais que les réalisations des composantes U_{ij} sont ignorées.

Estimation de s^2

Les considérations précédentes d'appliquent maintenant avec $w_i = k_i$, $e_i = e/k_i$. On pourra estimer s^2 par

$$\widetilde{s}^2 = c \sum \frac{k_i}{2 + e_i} (X_i - \widehat{m})^2 = c \sum \frac{k_i^2}{e + 2k_i} (X_i - \widehat{m})^2$$

où c est la constante convenable assurant l'absence de biais.

Si on sait, par exemple, que e est grand comparé aux k_i, *sans connaître la valeur exacte* de e, on pourra adopter l'estimateur

$$\widetilde{s}_1^2 = c_1 \sum k_i^2 (X_i - \widehat{m})^2$$

où c_1 est la constante convenable assurant l'absence de biais. Cet estimateur sera encore optimal, en bonne approximation, puisque $2k_i$ est négligeable devant e par hypothèse.

Les hypothèses faites justifient donc l'emploi des poids relatifs k_i^2, alors que l'estimateur classique emploierait les poids relatifs k_i.

7. Application à la théorie de la crédibilité

Problème préliminaire

Notre problème préliminaire est la détermination du coefficient d'aplatissement $e(M)$, dans les hypothèses qui suivent, de la variable

$$M = \sum_{i=1}^{n} \frac{w_i}{w_\bullet} X_i, \quad \text{où} \quad w_\bullet = w_1 + \ldots + w_n.$$

Nous supposons les variables $X_1, \ldots, X_n$ indépendantes, avec

$$EX_i = m, \quad \text{Var}\, X_i = s_i^2, \quad e(X_i) = e_i \quad (i = 1, \ldots, n).$$

Par des calculs directs basés sur la formule $(**)$ du théorème 1, on obtient

$$EM = m, \quad \text{Var}\, M = \sum \frac{w_i^2}{w_\bullet^2} s_i^2, \quad e(M) = \frac{\sum w_i^4 s_i^4 e_i}{\left(\sum w_i^2 s_i^2\right)^2}.$$

Supposons de plus que chaque variable X_i soit moyenne arithmétique de w_i variables indépendantes, dites *composantes élémentaires de* X_i, chacune avec espérance m, variance σ^2 et coefficient d'aplatissement e. Alors, par les formules du début du paragraphe 6, $EX_i = m$, $\text{Var}\, X_i = s_i^2 = \frac{\sigma^2}{w_i}$, $e(X_i) = \frac{e}{w_i}$.

Alors $\text{Var}\, M = \frac{\sigma^2}{w_\bullet}$, $e(M) = \frac{e}{w_\bullet}$.

Le modèle de crédibilité de Bühlmann-Straub

Ce modèle est décrit en détail dans chacun des articles indiqués dans les références, le troisième excepté.

Nous considérons le portefeuille avec *variables observables* X_{ji} et *poids naturels* w_{ji} correspondants $(j = 1, \ldots, k; i = 1, \ldots, n)$. La variable X_{ji} est l'observation du contrat j et de l'année i.

La *moyenne observée* du contrat j est

$$\widehat{m}_j = \sum_{i=1}^{n} \frac{w_{ji}}{w_{j\bullet}} X_{ji} \quad \text{où} \quad w_{j\bullet} = w_{j1} + \ldots + w_{jn}.$$

Le *poids de crédibilité* du contrat j est

$$z_j = a w_{j\bullet}/(s^2 + a w_{j\bullet}),$$

où a est le *paramètre d'hétérogénéité* du portefeuille et s^2 le *paramètre de variabilité des observations dans le temps.*

L'estimation de s^2 ne pose généralement pas de problème pratique. Elle n'est pas considérée ci-dessous.

L'estimateur de Bichsel-Straub

Les variables $\widehat{m}_j$ $(j = 1, \ldots, k)$ sont indépendantes, telles que

$$E\widehat{m}_j = m, \quad \text{Var}\, \widehat{m}_j = \frac{a}{z_j}.$$

La situation semble classique. Le *pseudo-estimateur de Bichsel-Straub* pour a est celui défini par

$$\widehat{a}_{BS} = \frac{1}{k-1} \sum z_j (\widehat{m}_j - \widehat{m})^2, \quad \text{où} \quad \widehat{m} = \sum \frac{z_j}{z_\bullet} \widehat{m}_j, \quad z_\bullet = \sum z_j.$$

En fait, le paramètre à estimer a figure dans les z_j et $\widehat{a}_{BS}$ n'est pas un estimateur au sens propre du terme. Il est pourtant employé avec succès par les praticiens, de la manière itérative suivante. On fixe une valeur initiale a_0 à partir de laquelle on calcule, par $\widehat{a}_{BS}$, un premier itéré a_1. On recommence avec a_1 pour trouver a_2,... On démontre que la suite a_0, a_1, a_2,... converge vers une valeur a_∞ ne dépendant pas de la valeur initiale $a_0 > 0$. L'estimation adoptée est alors a_∞. En pratique, la convergence est très rapide.

Nous ne discuterons pas ici la méthodologie des pseudo-estimateurs. Il se fait que $\widehat{a}_{BS}$ a pratiquement poussé dans l'oubli les autres estimateurs, de type classique, de a.

Le problème qui nous préoccupe est que $\widehat{a}_{BS}$ ne tient nullement compte des coefficients d'aplatissement des $\widehat{m}_j$. Dans certains portefeuilles à sinistres rares, les coefficients d'aplatissement sont loin d'être négligeables et l'optimalité de $\widehat{a}_{BS}$ est fort douteuse dans ces cas.

Estimateur tenant compte des coefficients d'aplatissement

Considérons un contrat fixé, le j-ième, du portefeuille. La loi du vecteur $(X_{j1}, \ldots, X_{jn})$ de ses variables observables dépend d'un paramètre inconnu θ_j (le plus souvent multidimensionnel), traité en variable aléatoire Θ_j, dite la *variable de structure* du contrat. Dans une multitude de portefeuilles pratiques, on peut admettre que les hypothèses faites sur le vecteur $(X_1, \ldots, X_n)$ du problème préliminaire, restent valables, pour Θ_j fixé, pour le vecteur conditionnel $(X_{j1}, \ldots, X_{jn}/\Theta_j)$.

On aboutit ainsi à un coefficient d'aplatissement, toujours pour Θ_j fixé,

$$e(\widehat{m}_j/\Theta_j) = \frac{e_j(\Theta_j)}{w_{j\bullet}},$$

mais qui est inutilisable pour la dérivation de $e(\widehat{m}_j)$.

En effet, on ignore le plus souvent la loi de Θ_j. De plus, les observations sont généralement trop rares pour estimer les $e(\Theta_j)$. Par exemple, il est utopique de vouloir estimer, pour un conducteur automobile, à partir d'observations concernant seulement ce conducteur, le paramètre λ de son processus d'arrivées de sinistres, admettant que ce soit un processus de Poisson homogène.

L'hypothèse simplificatrice que nous ferons pour sortir de cette impasse est que

$$e_j = e(\widehat{m}_j) \cong \frac{e}{w_{j\bullet}}$$

pour un certain paramètre e interprété comme étant un coefficient d'aplatissement moyen des composantes élémentaires des variables observables. Travailler avec cet e, même interprété et estimé grossièrement, vaudra sans doute mieux qu'adopter la valeur $e = 0$, implicitement admise par Bichsel et Straub dans la construction de leur estimateur $\widehat{a}_{BS}$.

Par exemple, dans un portefeuille en assurance automobile, le λ moyen s'estime trivialement et il suffit alors de prendre pour e l'inverse de ce λ moyen. On trouve

ainsi pour e des valeurs qui peuvent être de l'ordre de grandeur de 10, qui sont loin d'être négligeables dans certains cas. L'exemple numérique qui suit le prouvera.

En définitive, les considérations du paragraphe 5 nous conduisent au pseudo-estimateur

$$\widehat{a} = c \sum p_j (\widehat{m}_j - \widehat{m})^2 \quad \text{avec} \quad p_j = \frac{z_j}{2 + e_j}$$

où e_j est approximé par $e/w_{j\bullet}$ et c est la constante de non-biais.

Explicitement, on obtient

$$\widehat{a} = c \sum \frac{z_j w_{j\bullet}}{e + 2w_{j\bullet}} (\widehat{m}_j - \widehat{m})^2 \quad \text{où} \quad \frac{1}{c} = \sum \frac{w_{j\bullet}}{e + 2w_{j\bullet}} \left(1 - \frac{z_j}{z_\bullet}\right).$$

Bonus-malus en assurance automobile dans le cas de contrats couvrant plusieurs véhicules

Nous allons appliquer ce qui précède à un portefeuille en assurance automobile composé de 996 contrats couvrant chacun de un à quatre véhicules. Les véhicules couverts par un même contrat sont indiscernables par hypothèse. Pour chaque contrat nous disposons des couples

$$(w_{j1}, N_{j1}), (w_{j2}, N_{j2}), (w_{j3}, N_{j3})$$

où w_{ji} $(i = 1, 2, 3)$ est le nombre de véhicules assurés durant l'année i et N_{ji} le nombre de sinistres survenus durant cette année. Nous avons travaillé sur statistiques réelles.

Le modèle de Bühlmann-Straub ne s'applique pas aux nombres de sinistres N_{ji}, mais bien, aux fréquences $X_{ji} = N_{ji}/w_{ji}$.

En vue d'élaborer une tarification bonus-malus, basée seulement sur le nombre de sinistres survenus (et non sur leurs coûts), il est nécessaire d'estimer la fréquence annuelle future, soit Y_j, dans chaque contrat j. Elle résulte de la formule de Bühlmann-Straub

$$Y_j = z_j \widehat{m}_j + (1 - z_j)\widehat{m}.$$

Ci-dessous nous comparons quelques estimations obtenues d'une part à partir de $\widetilde{a}$, d'autre part à partir de $\widehat{a}_{BS}$.

Nous reproduisons les résultats relatifs à quatre contrats particuliers, numérotés de 1 à 4. Il s'avère que la considération des coefficients d'aplatissement est loin d'être négligeable.

Estimations résultant de l'emploi de $\widehat{a}_{BS}$.

$a = 0.00109$

Contrat	z_j	Y_j
1	0.20092	0.09307
2	0.14356	0.08181
3	0.25108	0.08723
4	0.07733	0.08814

Estimations résultant de l'emploi de $\widehat{a}_{BS}$.

$a = 0.00073$

Contrat	z_j	Y_j
1	0.14454	0.08202
2	0.10124	0.07351
3	0.18356	0.07825
4	0.05332	0.07743

Références

[1] BÜHLMANN H. and STRAUB E., (1970), Glaubwürdigkeit für Schadensätze. (Mitteilungen Schweizerische Vereinigung der Versicherungsmathematiker 70, n° 1, 111-133)

[2] DE VYLDER F. and GOOVAERTS M., (January 1992), Estimation of the heterogeneity parameter in the Bühlmann-Straub credibility theory model. (Insurance : Mathematics & Economics, Vol. 10 n° 4, 233-238)

[3] DE VYLDER F. and GOOVAERTS M., (April 1992), Optimal parameter estimation under zero-excess assumptions in a classical model. (Insurance : Mathematics & Economics, Vol. 11, n° 1, 1-6)

[4] DUBEY A. and GISLER A., (1981), On parameter estimation in credibility. (Mitteilungen Schweizerische Vereinigung der Versicherungsmathematiker, Vol. 81, n°2, 187-212).