Méthodes statistiques de construction de tarif

Méthodes statistiques de construction de tarif

Par Marc Hallin, Université Libre de Bruxelles

Une méthode statistique de construction de tarif a été proposée par *P. Pitkänen* [1976] au récent congrès international de Tokyo. Le problème consiste à sélectionner, pour un risque donné, les variables «descriptives» à introduire dans le tarif. La méthode de Pitkänen tente de généraliser, dans un contexte d'*analyse de la variance*, les méthodes de sélection couramment utilisées dans l'étude des modèles linéaires (et, notamment – mais l'hypothèse de linéarité ne se justifie qu'en première approximation –, par *J. Lemaire* [1977]).

Certains aspects de cette généralisation peuvent cependant être critiqués et améliorés. Ainsi, par exemple, le critère de sélection adopté, fondé sur la comparaison des valeurs observées de variables χ^2 ne présentant pas les mêmes nombres de degrés de liberté, peut conduire, le plus souvent, à un arrêt prématuré de la procédure. En outre, la méthode proposée (cf. aussi [*Pitkänen*, 1974]) se rattache à la méthode dite *de sélection progressive (forward selection)*, et non, ainsi que l'affirme son auteur, à la méthode *de sélection pas à pas (stepwise selection)*.

Ces remarques, ainsi que quelques autres, et qui sont peut-être d'une portée plus théorique que pratique, sont détaillées plus loin; nous donnons également l'extension exacte de la méthode de sélection *pas à pas*.

Malheureusement, tant dans la version de P. Pitkänen que dans celle que nous donnons ici, la méthode repose sur les hypothèses habituelles de l'*analyse de la variance*: l'outil principal est une classique statistique F de Fisher-Snedecor. Ces hypothèses – *normalité* et *homoscédasticité*[1] des observations – sont rarement vérifiées, dans un cadre actuariel, par le nombre ou le montant cumulé des sinistres. Et les transformations non linéaires fréquemment utilisées pour «normaliser» les observations et réduire les inégalités des variances ne laissent pas invariant, généralement, le résultat de la sélection. Or le test F est connu pour peu robuste vis-à-vis des violations de ces hypothèses.

De plus, l'hypothèse même d'homoscédasticité empêche que la méthode soit applicable aux principes de calcul des primes tels que le principe de la variance ou celui de l'écart-type.

Toutes ces raisons nous ont conduit à envisager une nouvelle méthode de sélec-

[1] Soient Y la variable dépendante étudiée (le montant cumulé des sinistres, par exemple), $X_1, \ldots, X_N$ les variables explicatives ou classifiantes (utilisables dans un tarif); rappelons que Y satisfait à l'hypothèse d'homoscédasticité si sa variance conditionnelle ne dépend pas des valeurs prises par $X_1, \ldots, X_N$.

tion des variables, reposant sur des techniques de type non paramétrique. Cette méthode nouvelle présente l'avantage de s'appliquer, sans hypothèses préalables, à tout principe de calcul des primes.
Une application numérique de cette méthode, ainsi qu'une comparaison de ses performances avec celles de la méthode de Pitkänen, sera réalisée, en collaboration avec *J.-F. Ingenbleek* et *J. Lemaire*, à partir de données relatives à l'assurance automobile.

1. La méthode de sélection de Pitkänen

Soit Y la variable aléatoire représentant le montant cumulé des sinistres relatifs à un risque donné. Ce risque est décrit par N variables observables $X_1, \ldots, X_N$, pouvant être, éventuellement, de type nominal ou ordinal, et qui fournissent un maximum de renseignements concernant le risque considéré. Y joue le rôle de variable *dépendante* ou *variable-critère;* $X_1, \ldots, X_N$, considérées ici comme des variables mathématiques, sont utilisées dans la suite comme variables *classifiantes*. Le problème posé par la construction d'un tarif consiste à déterminer la «meilleure» classification possible, dans un sens qui reste à préciser. Il convient pour cela de sélectionner, parmi $X_1, \ldots, X_N$ (nous les appellerons désormais *variables candidates*), les variables les plus *significatives:* ces variables sélectionnées interviendront dans le tarif et seront, en conséquence, appelées *variables de tarif*.

1 a. Définitions et notations

Considérons la variable candidate X_i. Si elle est de type continu, supposons avoir divisé son domaine de variation en N_i classes disjointes; si elle prend un ensemble discret de valeurs, désignons par N_i le nombre de ces valeurs. Dans tous les cas, nous écrirons $X_i = j$ lorsque X_i prend sa $j^{\text{ème}}$ valeur ou appartient à sa $j^{\text{ème}}$ classe.
Supposons disposer d'un échantillon de taille n constitué d'observations du type $(y; x_1, \ldots, x_N)$. Pour une valeur déterminée de $(x_1, \ldots, x_N)$ (c'est-à-dire une cellule de la classification basée sur les variables $X_1, \ldots, X_N$), notons

$$n(x_1, \ldots, x_N) \quad \text{et} \quad y_\alpha(x_1, \ldots, x_N) \quad (\alpha = 1, \ldots, n(x_1, \ldots, x_N)) \tag{1}$$

le nombre de ces observations et les valeurs du montant cumulé des sinistres. Nous ferons l'hypothèse qu'aucune de ces cellules n'est vide d'observations:

$$n(x_1, \ldots, x_N) > 0 \quad \forall\, (x_1, \ldots, x_N) \in \prod_{i=1}^{N} \{1, \ldots, N_i\}.$$

De façon plus générale, soit $\{X_{(1)}, \ldots, X_{(k)}\}$ un sous-ensemble quelconque de variables candidates: nous noterons

$$n(x_{(1)}, \ldots, x_{(k)}) \quad \text{et} \quad y_\alpha(x_{(1)}, \ldots, x_{(k)}) \quad (\alpha = 1, \ldots, n(x_{(1)}, \ldots, x_{(k)}))$$

l'effectif observé et la $\alpha^{\text{ème}}$ observation de y dans la cellule $(x_{(1)}, \ldots, x_{(k)})$ de la classification basée sur les variables considérées.

Introduisons encore les moyennes conditionnelles des observations:

$$\bar{y}(x_{(1)}, \ldots, x_{(k)}) = \frac{1}{n(x_{(1)}, \ldots, x_{(k)})} \sum_{\alpha=1}^{n(x_{(1)}, \ldots, x_{(k)})} y_\alpha\,(x_1, \ldots, x_k); \tag{2}$$

pour $k = 0$, on peut considérer que ces notations représentent, respectivement, le nombre total n d'observations et la $\alpha^{\text{ème}}$ valeur observée y_α $(\alpha = 1, \ldots, n)$ de Y, et la moyenne générale $\bar{y}$ de ces valeurs.

1b. Le test F

L'outil principal de la méthode de sélection de Pitkänen est le test F classique de l'*analyse de la variance*. Soit $\mu(x_{(1)}, \ldots, x_{(k-1)}, x_i)$ la moyenne conditionnelle de la variable aléatoire Y pour

$$(X_{(1)}, \ldots, X_{(k-1)}, X_i) = (x_{(1)}, \ldots, x_{(k-1)}, x_i).$$

Si les hypothèses habituelles de l'*analyse de la variance* sont satisfaites, la quantité

$$F(i|(1)\ldots(k-1)) =$$

$$\frac{(n - N_{(1)} \cdot \ldots \cdot N_{(k-1)} \cdot N_i) \sum_{x_{(1)}=1}^{N_{(1)}} \cdots \sum_{x_{(k-1)}=1}^{N_{(k-1)}} \sum_{x_i=1}^{N_i} n(x_{(1)} \ldots x_{(k-1)}, x_i) \left[\bar{y}(x_{(1)} \ldots x_{(k-1)}\ x_i) - \bar{y}(x_{(1)} \ldots x_{(k-1)})\right]^2}{N_{(1)} \cdot \ldots \cdot N_{(k-1)}\,(N_i - 1) \sum_{x_{(1)}=1}^{N_{(1)}} \cdots \sum_{x_{(k-1)}=1}^{N_{(k-1)}} \sum_{x_i=1}^{N_i} \sum_{\alpha=1}^{n(x_{(1)} \ldots x_{(k-1)} x_i)} \left[y_\alpha(x_{(1)} \ldots x_{(k-1)}\ x_i) - \bar{y}(x_{(1)} \ldots x_{(k-1)}\ x_i)\right]^2} \tag{3}$$

est distribuée, sous l'hypothèse d'égalité des moyennes

$$H_0^{(0)}: \mu(x_{(1)}, \ldots, x_{(k-1)}, x_i) = \mu(x_{(1)}, \ldots, x_{(k-1)}, x'_i) \tag{4}$$

$$\forall\, x_i, x_i' \in \{1, \ldots, N_i\},\ \forall\, (x_{(1)}, \ldots, x_{(k-1)}) \in \prod_{i=1}^{k-1} \{1, \ldots, N_{(i)}\},$$

comme une variable F à $N_{(1)} \ldots . N_{(k-1)} \,.\, (N_i - 1)$ et $n - N_{(1)} \ldots . N_{(k-1)} \,.\, N_i$ degrés de liberté.

Ce test est utilisé pour tester $H_0^{(0)}$ contre l'hypothèse $H_1^{(0)}$ que l'une au moins des égalités de moyennes (4) est fausse.

Si on note $\Theta^2_{y\,.\,x_{(1)}, \ldots, x_{(k)}}$ le rapport de corrélation observé de la variable Y en les variables $X_{(1)}, \ldots, X_{(k)}$, la statistique $F\big(i \mid (1), \ldots, (k-1)\big)$ peut encore s'écrire, sous une forme moins encombrante,

$$F\big(i \mid (1), \ldots, (k-1)\big) = \frac{(n - N_{(1)} \,.\, \ldots \,.\, N_{(k-1)} \,.\, N_i)}{N_{(1)} \,.\, \ldots \,.\, N_{(k-1)} \,.\, (N_i - 1)} \, \frac{\Theta^2_{y\,.\,x_{(1)} \ldots x_{(k-1)} x_i} - \Theta^2_{y\,.\,x_{(1)} \ldots x_{(k-1)}}}{1 - \Theta^2_{y\,.\,x_{(1)} \ldots x_{(k-1)}\, x_i}}.$$

Son interprétation intuitive, sous cette forme, se fait aisément. Si on se rappelle que le rapport de corrélation $\Theta^2_{y\,.\,x_{(1)} \ldots x_{(k)}}$ représente la proportion de la variance de Y expliquée par le «tarif» (basé, ici, sur $X_{(1)}, \ldots, X_{(k)}$), ce rapport, compris entre 0 et 1, constitue une mesure de l'*efficacité* de ce «tarif». La statistique (3) évalue donc l'amélioration de cette efficacité lors de l'introduction de la variable X_i dans le tarif basé sur $X_{(1)}, \ldots, X_{(k-1)}$.

1 c. Méthode de sélection pas à pas (stepwise selection)

La méthode suivante est une transposition, dans un contexte d'*analyse de la variance*, de la méthode dite de *régression pas à pas (stepwise regression)* utilisée dans l'étude des modèles linéaires [*Draper* and *Smith*, 1966, et *Lemaire*, 1977]. Les variables sont sélectionnées une à une selon la méthode récurrente décrite ci-dessous. Tous les tests sont effectués à un niveau de probabilité α.

Etape 1. Notons $F(i)$ la valeur de la statistique (3) calculée pour $k = 1$.
A chacune de ces quantités correspond un niveau de signification q_i, valeur en $F(i)$ de la fonction de répartition de la variable F à $N_i - 1$ et $n - N_i$ degrés de liberté.
Soit $X_{(1)}$ la variable à laquelle correspond le niveau q_i le plus élevé: $X_{(1)}$ peut être considérée comme la variable candidate la plus significative *individuellement*, c'est-à-dire celle qui mène le plus nettement au rejet de l'hypothèse (4)

$$H_0^{(0)}: \mu(x_i) = \mu(x_i') \quad \forall\, x_i, x_i' \in \{1, \ldots, N_i\}.$$

- Si $q_{(1)}$ est strictement supérieur à $1 - \alpha$, $H_0^{(0)}$ est rejetée au niveau α, et $X_{(1)}$ devient, *provisoirement*, la première variable de tarif, celle qui contribue le plus à expliquer les variations des moyennes conditionnelles $\bar{y}(x_i)$.

– Sinon, aucune des variables candidates ne peut être considérée comme significative, et une prime uniforme, indépendante des valeurs prises par $X_1, \ldots, X_n$, semble devoir convenir pour tous les risques examinés.

Etape k. Notons $X_{(1)}, \ldots, X_{(k-1)}$ les variables de tarif obtenues à la fin de l'étape précédente[2]. Pour chaque variable X_i restée hors-tarif, considérons la valeur $F\big(i|(1), \ldots, (k-1)\big)$ prise par la statistique (3); à chacune de ces quantités correspond un niveau de signification q_i, valeur en $F\big(i|(1), \ldots, (k-1)\big)$ de la fonction de répartition de la variable F à $N_{(1)} \ldots . N_{(k-1)} . (N_i - 1)$ et $n - N_{(1)} . \ldots . N_{(k-1)} . N_i$ degrés de liberté.
Soit $q_{(k)}$ le plus élevé de ces niveaux de signification: $X_{(k)}$ est, provisoirement, considérée comme la $k^{\text{ème}}$ variable de tarif.
Considérons ensuite, pour chacune des variables de tarif $X_{(l)}$ ($X_{(k)}$ comprise), la valeur $F\big((l)\,|\,(1), \ldots, (l-1), (l+1), \ldots, (k)\big)$ de la statistique (3) (pour $(l) = (k)$, c'est la quantité qui vient d'être calculée). A chacune de ces valeurs correspond à nouveau un niveau de signification $q_{(l)}$ (à calculer sur la base de $N_{(1)} . \ldots . (N_{(k)} - 1) / N_{(l)}$ et $(n - N_{(1)} . \ldots . N_{(k)} / N_{(l)}$ degrés de liberté). Soit $q_{(m)}$ le plus bas de ces niveaux.
– Si $q_{(m)}$ est strictement supérieur à $1-\alpha$, on passe à l'étape suivante avec $\{X_{(1)}, \ldots, X_{(k)}\}$ pour nouvel ensemble de variables de tarif.
– Si $q_{(m)}$ est inférieur ou égal à $1-\alpha$, l'hypothèse $H_0^{(0)}$ obtenue en remplaçant i par (m) dans (4) ne peut être rejetée au niveau de probabilité α. Si $(m) \neq (k)$, on passe à l'étape suivante en prenant pour variables de tarif les variables $X_{(1)}, \ldots, X_{(m)-1}, X_{(m)+1}, \ldots, X_{(k)}$; si $(m) = (k)$, la procédure de sélection s'arrête, l'ensemble de variables de tarif final étant $\{X_{(1)}, \ldots, X_{(k-1)}\}$.

1 d. Remarques

i) Sélection *progressive (forward)* et sélection *pas à pas (stepwise)*

Comme on a pu le voir, chaque étape de sélection comporte deux parties:
– introduction dans le tarif de la variable candidate $X_{(k)}$ la plus significative;
– exclusion éventuelle, au cas oú elle serait devenue (après introduction de $X_{(k)}$) non significative au niveau α, de la variable de tarif la moins significative.
Cette seconde partie, qui est cependant caractéristique de la méthode *pas à pas*, est absente chez *Pitkänen:* seule est testée, à chaque étape, la variable qui vient

[2] Remarquons que leur nombre peut être strictement inférieur à k–1, et que, éventuellement, et en dépit de la notation, $X_{(1)}$ peut ne plus en faire partie.

d'être introduite. Sa méthode n'est pas, en fait, comme il l'annonce, une méthode du type *pas à pas:* il s'agit plutôt d'une méthode de sélection *progressive* (*forward* selection), beaucoup moins satisfaisante en pratique (cf. [*Draper* and *Smith*, 1966]).

ii) Degrés de liberté et niveau de signification

Il peut être dangereux de comparer entre elles des statistiques ne présentant pas le même nombre de degrés de liberté; la sélection de $X_{(k)}$ se fait, chez *Pitkänen*, en choisissant, parmi toutes les variables candidates X_i, celle qui réalise le maximum observé d'une forme quadratique à $N_{(1)} \cdot \ldots \cdot N_{(k-1)} \cdot (N_i - 1)$ degrés de liberté. Si les valeurs des N_i sont très différentes d'une variable à l'autre, cela peut conduire à des choix peu cohérents, favorisant systématiquement la sélection des variables X_i présentant un N_i élevé, c'est-à-dire un grand nombre de classes ou de niveaux.
En outre, la statistique (de type χ^2) sur laquelle est basée la sélection n'est pas la même que celle (F) qui est utilisée pour règle d'arrêt: la variable $X_{(k)}$ sélectionnée n'étant pas, en général, la plus significative des variables candidates, la procédure s'arrête souvent prématurément.
Ce problème n'apparaissait pas dans le cadre des modèles de régression, toutes les statistiques F rencontrées à une étape k présentant les mêmes degrés de liberté. (La sélection peut, alors, également s'effectuer sur la base des coefficients de corrélation partielle.)

iii) Sélection globale et sélection hiérarchisée

L'hypothèse $H_0^{(0)}(4)$ dont le test constitue la règle d'arrêt de la procédure de sélection postule l'égalité des N_i moyennes conditionnelles de Y dans *chacune* des $N_{(1)} \cdot \ldots \cdot N_{(k-1)}$ cellules résultant des variables de tarif $X_{(1)} \ldots X_{(k-1)}$; le rejet de cette hypothèse peut être dû à sa fausseté dans l'une des cellules seulement. On peut par conséquent désirer n'introduire la subdivision associée à $X_{(k)}$ *que* dans quelques cellules, non dans toutes.
En effectuant séparément les tests dans certains sous-ensembles de cellules, on peut ainsi obtenir une classification finale totalement ou partiellement hiérarchisée. La procédure de sélection s'adapte aisément; la seconde partie de chaque étape (élimination rétrospective de certaines variables) peut, éventuellement, demander certaines précautions, toutes les combinaisons de toutes les variables n'étant pas à envisager.

iv) Choix du découpage en classes

Le découpage en classes du domaine de variation des variables X_i a été considéré comme une des données du problème, mais influence de façon prépondérante le résultat final, et donc la forme du tarif. Une valeur trop petite de N_i ou un regroupement maladroit peuvent conduire à un arrêt prématuré de la procédure. Un nombre trop élevé de classes se traduit par une amélioration factice (à cause de la diminution du nombre de degrés de liberté) du rapport de corrélation, et à un gonflement inutile du volume du tarif. Dans un exemple numérique [1974], *Pitkänen* introduit, par exemple, une variable représentant le poids des véhicules assurés, une autre donnant les caractéristiques géographiques du contrat considéré. Les poids sont répartis en 12 classes: pourquoi pas 11? pourquoi pas 4? Faut-il distinguer seulement *villes/campagne*? ou *grandes agglomérations/moyennes agglomérations/petites villes/campagne*? ou regrouper *petites villes et campagne*?
La qualité du tarif obtenu – et sa maniabilité – dépendant autant du découpage en classes que de la procédure de sélection utilisée, est-il bien utile d'appliquer une méthode raffinée après un découpage arbitraire?
Grâce à l'emploi de variables dichotomiques, la méthode de sélection pas à pas permet de déterminer simultanément les variables et les classes à prendre en considération. Soit X_i la $i^{\text{ème}}$ variable candidate (N_i classes, fixées a priori, qui ne seront pas forcément toutes retenues pour le tarif), que nous supposerons au moins de type ordinal (pour un ordre noté $\leq$). Soit $x^+_{i\,;\,j}$ la plus grande valeur possible pour X_i dans la $j^{\text{ème}}$ classe; posons

$$X_{i;j} = \begin{cases} 0 & si \quad X_i \leq x^+_{i;j} \\ 1 & \text{sinon} \end{cases} \qquad j = 1, \ldots, N_i - 1.$$

Chaque variable candidate donne ainsi naissance à $N_i - 1$ variables indicatrices. La procédure de sélection peut alors être appliquée aux $N' = \sum_{i=1}^{N} (N_i - 1)$ variables obtenues. Ces variables ne sont pas, bien entendu, indépendantes, et l'hypothèse qu'aucune des cellules n'est vide ne peut être vérifiée: une certaine hiérarchisation s'imposera naturellement. Moyennant quelques précautions, la procédure s'applique encore. Si *toutes* les variables candidates sont ainsi dichotomisées, la statistique (3) prend, pour k variables de tarif, une forme simplifiée à 2^{k-1} et $n - 2^k$ degrés de liberté, et il est inutile de recourir aux niveaux de signification q_i.

2. Vérification des hypothèses. Principes de calcul des primes

2a. Robustesse du test F utilisé

La procédure de sélection décrite au paragraphe précédent n'est une méthode exacte que si les hypothèses habituelles de l'*analyse de la variance* sont remplies:

i) normalité des observations Y;

ii) homoscédasticité des observations: la variance de Y doit être la même, quelles que soient les valeurs des variables candidates;

iii) indépendance des observations entre elles.

Si la troisième de ces hypothèses peut, en général, être considérée comme satisfaite, les deux premières ne le sont sûrement pas:

i) les statistiques de sinistres montrent en général, pour Y, une distribution asymétrique et leptocurtique («fat-tailed»); de surcroît, la probabilité de non-sinistre p_0 n'est en général pas nulle, et 0 constitue donc un atome de cette distribution;

ii) l'homoscédasticité n'est *évidemment* pas satisfaite dans le cas des montants de sinistres; c'est la raison pour laquelle ont été introduits les principes de la variance, de l'écart-type et les fonctions d'utilité. Cette hypothèse anéantit tout espoir d'appliquer la méthode de sélection de *Pitkänen* à un tarif tenant compte de la variabilité du risque.

La robustesse du test F est relativement satisfaisante (cf. [*Scheffé*, 1959, chap. 10]) par rapport aux violations de l'hypothèse de normalité; il est beaucoup plus sensible, en revanche, aux inégalités des variances, surtout lorsque les fréquences dans les cellules ne sont pas les mêmes; il est très difficile d'évaluer le comportement de la statistique (3) lorsque les deux hypothèses sont violées simultanément.

2b. Principes de calcul des primes

Les trois principes de calcul des primes considérés ici sont ceux qui satisfont au critère d'additivité [*Gerber*, 1974, a]. Notons $P(Y)$ la prime associée à un risque Y.

i) *Principe de l'espérance mathématique:* $P(Y) = E(Y)$. Pour une valeur $(x_{(1)}, \ldots, x_{(k)})$ des variables de tarif, on a donc

$$P(Y|x_{(1)}, \ldots, x_{(k)}) = \mu(x_{(1)}, \ldots, x_{(k)});$$

ce cas est en fait le seul considéré par *Pitkänen*, sa méthode reposant sur des tests d'égalité de ces moyennes conditionnelles.

ii) *Principe de la variance:* $P(Y) = E(Y) + \lambda \operatorname{var}(Y) \quad (\lambda > 0)$.

On a donc

$$P(Y|x_{(1)}, \ldots, x_{(k)}) = \mu(x_{(1)}, \ldots, x_{(k)}) + \lambda\sigma^2(x_{(1)}, \ldots, x_{(k)}),$$

$\sigma^2(...)$ représentant la variance conditionnelle de Y. Pour permettre l'application de la méthode de sélection, cette variance doit être supposée constante, ce qui ôte presque tout son intérêt au principe de la variance.

iii) *Principe de l'utilité nulle:* $P(Y)$ est solution de $E[u(P(Y)-Y)] = 0$, u étant une fonction d'utilité de classe C^2 et à dérivée strictement positive. *Gerber* ayant montré l'intérêt des fonctions d'utilité exponentielles (cf. aussi [*Lemaire*, 1975]), nous nous restreignons à ce cas. La prime est alors, pour

$$u(x) = \frac{1}{a}(1 - e^{-ax}) \quad (a > 0),$$

$$P(Y) = \frac{1}{a}\log_e E(e^{aY}),$$

et

$$P(Y|x_{(1)}, \ldots, x_{(k)}) = \frac{1}{a}\log_e E(e^{aY}|x_{(1)}, \ldots, x_{(k)}).$$

La méthode de sélection peut, a priori, s'appliquer encore dans ce cas, à condition de prendre comme variable dépendante $Z = e^{aY}$.
Malheureusement, si Y ne satisfait pas aux hypothèses, Z y satisfait encore bien moins. L'exponentielle accentue la non-normalité, l'asymétrie, l'aplatissement; si Y_1 et Y_2 ont, pour moyenne et variance respectives, (m, σ^2) et $(m, \sigma^2 + \Delta)$, la différence entre les variances de e^{Y_1} et e^{Y_2} est de l'ordre de

$$e^{2m+\sigma^2}(2e^{\sigma^2} - 1)\,\Delta;$$

en outre, des moyennes différentes pour Y_1 et Y_2 impliquent des variances différentes pour e^{aY_1} et e^{aY_2}!
Peu sûre dans le cadre du principe de l'espérance mathématique, la méthode de *Pitkänen* devient tout à fait inapplicable si on s'intéresse au principe de la variance ou à celui de l'utilité nulle. On est ainsi conduit à rechercher une procé-

dure à la fois plus robuste et plus souple; les méthodes de rangs semblent, à cet égard, l'outil le plus approprié.

3. Une méthode de sélection basée sur les statistiques de rangs

La procédure de sélection que nous proposons ci-dessous s'appuie sur deux hypothèses seulement.

i) Les observations de la variable Y sont mutuellement indépendantes.

ii) La variable Y est absolument continue.

Pour que cette seconde hypothèse soit vérifiée, il faut traiter séparément les cas de non-sinistre et les éliminer des observations. Y représente donc dans la suite le montant cumulé des sinistres d'un risque ayant donné lieu à un sinistre au moins pendant la période d'observation. Les cas de non-sinistre peuvent être examinés à travers la probabilité p_0 de ne pas avoir de sinistre; le lien entre p_0 et les variables candidates X_i peut s'étudier au moyen de statistiques χ^2 d'homogénéité de type classique.

3a. Tests de comparaisons multiples basés sur les rangs

Considérons N populations, et un ensemble

$$\{y_\alpha(j) | \alpha = 1, \ldots, n_j;\ j = 1, \ldots, N\}$$

d'observations indépendantes de Y, $y_\alpha(j)$ représentant la $\alpha^{\text{ème}}$ observation provenant de la $j^{\text{ème}}$ population. En regroupant ces $n = \sum_{j=1}^{N} n_j$ observations, on peut leur attribuer un rang $R_\alpha(j)$ compris entre 1 et n.

Soient $F_1, \ldots, F_N$ les fonctions de répartition, supposées continues, de chacune des populations. Sous l'hypothèse nulle

$$H_0: F_1 \equiv F_2 \equiv \ldots \equiv F_j \equiv \ldots \equiv F_N, \tag{5}$$

ces n rangs peuvent être considérés comme une permutation aléatoire de $\{1, \ldots, n\}$. Toute statistique construite à partir de ces rangs possède donc, sous H_0, une distribution entièrement spécifiée, et indépendante de la distribution F commune aux populations.

Afin de construire une telle statistique, introduisons un ensemble de fonctions de ces rangs, de la forme

$$E_R = J_n\left(\frac{R}{n+1}\right) \quad 1 \leq R \leq n; \tag{6}$$

on obtient ainsi *n scores* $E^n_{R_\alpha(j)}$. A chacune des N populations correspond un score moyen:

$$\bar{E}^n(j) = \frac{1}{n_j}\sum_{\alpha=1}^{n_j} E^n_{R_\alpha(j)} \quad 1 \leq j \leq N. \tag{7}$$

Considérons encore la moyenne et la variance générales des $E^n_{R_\alpha(j)}$:

$$\bar{E}^n = \frac{1}{n}\sum_{j=1}^{N} n_j\,\bar{E}^n(j), \tag{8}$$

$$V_n = \frac{1}{n}\sum_{j=1}^{N}\sum_{\alpha=1}^{n_j}\left(E^n_{R_\alpha(j)}\right)^2 - \left(\bar{E}^n\right)^2. \tag{9}$$

Remarquons que $\bar{E}^n$ et V_n dépendent uniquement du choix des scores utilisés et de la taille n de l'échantillon.
Sous des conditions assez générales, et qui sont satisfaites par tous les scores utilisés en pratique, on peut montrer que la statistique

$$L_n = \frac{1}{V_n}\sum_{j=1}^{N} n_j\left(\bar{E}^n(j) - \bar{E}^n\right)^2 \tag{10}$$

est asymptotiquement distribuée, sous H_0, comme une χ^2 à $N-1$ degrés de liberté. Utiliser cette statistique L_n revient, en fait, à appliquer aux scores les méthodes traditionnelles de l'Analyse de la Variance, la variance de la population étant connue (ce qui permet d'utiliser une variable χ^2 au lieu d'une variable F).
L'hypothèse H_0 (identité des N populations) peut être testée, au moyen de L_n, contre tois types d'hypothèse adverse:
$H_1^{(1)}$ non-homogénéité des moyennes (si on suppose l'homogénéité des variances);
$H_1^{(2)}$ non-homogénéité des variances (si on suppose l'homogénéité des moyennes);
$H_1^{(3)}$ non-homogénéité des moyennes *ou* des variances.

3b. Sélection non paramétrique des variables de tarif

La procédure de sélection se déroule exactement comme dans le cas paramétrique (1 b), mais les hypothèses testées et les statistiques utilisées ne sont plus les mêmes. Considérons les N_i populations caractérisées par

$$x = (x_{(1)}, \ldots, x_{(k-1)}, x_i), \quad 1 \leq x_i \leq N_i,$$

$x_{(1)}, \ldots, x_{(k-1)}$ étant des valeurs fixées des variables $X_{(1)}, \ldots, X_{(k-1)}$; soit $L^i_{n(x_{(1)}, \ldots, x_{(k-1)})}$ la statistique (10) obtenue en rangeant les $n(x_{(1)}, \ldots, x_{(k-1)})$ observations correspondantes de Y et en leur appliquant les formules (6) à (10) (pour $N = N_i$ et $n = n(x_{(1)}, \ldots, x_{(k-1)})$). Le rôle joué dans la procédure (1.b) par la statistique $F(i|(1), \ldots, (k-1))$ (3) est tenu ici par

$$L(i|(1), \ldots, (k-1)) = \sum_{x_{(1)}=1}^{N_{(1)}} \ldots \sum_{x_{(k-1)}=1}^{N_{(k-1)}} L^i_{n(x_{(1)}, \ldots, x_{(k-1)})}. \tag{11}$$

Notons $F(y|x_{(1)}, \ldots, x_{(k-1)}, x_i)$ la fonction de répartition conditionnelle de Y étant donné $(x_{(1)}, \ldots, x_{(k-1)}, x_i)$; sous l'hypothèse

$$H_0: F(y|x_{(1)}, \ldots, x_{(k-1)}, j) \equiv F(y|x_{(1)}, \ldots, x_{(k-1)}, j')$$

$$\forall j, j' \in \{1, \ldots, N_i\}, \ \forall (x_{(1)}, \ldots, x_{(k-1)}) \in \prod_{i=1}^{k-1} \{1, \ldots, N_{(i)}\}, \tag{12}$$

$L(i|(1), \ldots, (k-1))$ est asymptotiquement distribuée comme une variable χ^2 à $N_{(1)} \cdot \ldots \cdot N_{(k-1)} \cdot (N_i - 1)$ degrés de liberté. L'hypothèse (12) peut ainsi être testée contre

– $H_1^{(3)}: F(y(x_{(1)}, \ldots, x_{(k-1)}, j)|x_{(1)}, \ldots, x_{(k-1)}, j)$

$\equiv F(y(x_{(1)}, \ldots, x_{(k-1)}, j')|x_{(1)}, \ldots, x_{(k-1)}, j')$ (mêmes quantificateurs qu'en (12)),

où

$$y(x_{(1)}, \ldots, x_{(k-1)}, j) = (y + \Theta(x_{(1)}, \ldots, x_{(k-1)}, j))(1 + \delta(x_{(1)}, \ldots, x_{(k-1)}, j))$$

et

$$\Theta(x_{(1)}, \ldots, x_{(k-1)}, j) \neq \Theta(x_{(1)}, \ldots, x_{(k-1)}, j')$$

ou $\quad \delta(x_{(1)}, \ldots, x_{(k-1)}, j) \neq \delta(x_{(1)}, \ldots, x_{(k-1)}, j')$

pour un couple (j, j') et une valeur de $(x_{(1)}, \ldots, x_{(k-1)})$ au moins, (13)

ou

– $H_1^{(1)}$: idem que $H_1^{(3)}$, mais en imposant $\delta(\ldots) = 0$,

ou

– $H_1^{(2)}$: idem que $H_1^{(3)}$, mais en imposant $\Theta(\ldots) = 0$.

H_0 peut se mettre, bien entendu, sous la forme

$$H_0\colon \Theta(x_{(1)}, \ldots, x_{(k-1)}, j) = \Theta(x_{(1)}, \ldots, x_{(k-1)}, j')$$
$$\delta\,(x_{(1)}, \ldots, x_{(k-1)}, j) = \delta\,(x_{(1)}, \ldots, x_{(k-1)}, j')$$

$$\forall j, j' \in \{1, \ldots, N_i\},\ \forall (x_{(1)}, \ldots, x_{(k-1)}) \in \prod_{i=1}^{k-1} \{1, \ldots, N_{(i)}\}.$$

Selon le cas, on a affaire à un test portant sur les moyennes, les variances, ou sur ces deux paramètres simultanément.

3c. Puissance des tests – Scores optimaux

A tout choix d'une fonction E_R (6) correspond un certain type de test de rangs. La puissance de ces tests dépend, bien entendu, des distributions de Y dans les populations (c'est-à-dire des F_i, des Θ et des δ); elle dépend également du choix des scores utilisés.

On peut porter son choix sur des scores de type «classique» (pour le test de H_0 contre $H_1^{(1)}$, par exemple, les scores de Fisher-Yates-Terry-Hoeffding ou *scores normaux*, les scores de van der Waerden, ceux de Kruskal-Wallis...). Si on a une certaine idée de la forme de la distribution de Y, il est préférable de choisir les scores qui maximisent, localement, la puissance: les scores dits *optimaux* (pour leur définition, cf., par exemple, [*Hájek* et *Šidák*, 1967]). Ainsi, pour le test de H_0 contre $H_1^{(1)}$, les scores optimaux sont donnés par

$$E_R = n\binom{n-1}{R-1}\int_0^1 \varphi(u,f)u^{R-1}(1-u)^{n-R}\,\mathrm{du}$$
$$\simeq \varphi\left(\frac{R}{n+1}, f\right),$$

où

$$\varphi(u,f) = \frac{f'(F^{-1}(u))}{f(F^{-1}(u))} \qquad 0 < u < 1,$$

f et F étant la densité de probabilité et la fonction de répartition de Y sous H_0.

Pour le test de H_0 contre $H_3^{(1)}$, la notion de puissance localement maximum est plus délicate: pour $N = 2$, à chaque direction du plan (Θ,δ) correspond un

ensemble de scores optimaux. Si on s'intéresse au principe de la variance, on peut choisir la direction du gradient de $E(y) + \lambda \text{ Var } (y)$ et s'appuyer sur les résultats de [*Beran*, 1970] (cf. [*Hallin* et *Ingenbleek*]).
Signalons enfin que la distribution asymptotique des statistiques L sous les hypothèses adverses peut être obtenue (cf. [*Puri* et *Sen*, 1971], p. 205).

Références

Beran, R. J.: Linear rank statistics under alternatives indexed by a vector parameter. The Annals of Math. Stat., 1970, 41, pp. 1896–1905.
De Frenne, A. et *Ingenbleek, J.-F.:* Inférence non paramétrique univariée et multivariée. Université Libre de Bruxelles, stencilé (1977).
Draper, N. et *Smith, H.:* Applied Regression Analysis. Wiley, New York, 1966.
Gerber, H. U.: On Iterative Premium Calculation Principles. Bull. Ass. Act. Suisses, 1974, 2, pp. 163–172.
Hájek, J. et *Šidák, Z.:* Theory of Rank Tests, Academic Press, New York, 1967.
Hallin, M. et *Ingenbleek, J.-F.:* Sélection de variables en présence d'hétéroscédasticité. A paraître.
Lemaire, J.: Sur l'emploi des fonctions d'utilité en assurance. Bull. ARAB, Vol. 70, 1975, pp. 64–73.
– Selection Procedures of Regression Analysis Applied to Automobile Insurance. Ce volume.
Pitkänen, P.: Tariff theory. ASTIN, Colloquium in Turku, 1974, pp. 204–228.
– A theoretical approach to premium rating. Int.Congress of Actuaries in Tokyo, pp. 247–252.
Puri, M. et *Sen, P.:* Nonparametric Methods in Multivariate Analysis. J. Wiley, New York, 1971.
Scheffé, H.: The Analysis of Variance. J. Wiley, New York 1959.
Nous tenons à remercier A. De Frenne et J.-F. Ingenbleek, dont les connaissances dans le domaine des méthodes non paramétriques nous ont été précieuses.

Marc Hallin
Institut de Statistique
Université Libre de Bruxelles
CP. 210 Campus de la Plaine
Boulevard du Triomphe
B-1050 Bruxelles

Résumé

Une méthode statistique de construction de tarif a été proposée par P. Pitkänen au récent Congrès International de Tokyo. Nous montrons comment cette méthode peut être améliorée à partir d'une généralisation des méthodes de sélection de variables utilisées en analyse des modèles linéaires. Ces méthodes, cependant, reposent sur des hypothèses assez restrictives, et qui ne semblent pas réalisées dans un cadre actuariel. C'est pourquoi nous proposons une méthode nouvelle basée sur les techniques de rangs, et qui s'applique, sans hypothèses, quelles que soient les distributions considérées.

Abstract

A statistical method of tariff construction was proposed by P. Pitkänen at the last International Congress of Actuaries. We show how that method can be improved on the basis of extended multidimensional selection procedures. These methods, however, rely on strong distributional assumptions which do not seem to be fulfilled in actuarial problems. We therefore suggest a new, non-parametric, selection procedure which applies without any distributional assumption.