REVUE FRANÇAISE D'INFORMATIQUE ET DE RECHERCHE OPÉRATIONNELLE. SÉRIE VERTE

DAN ZORILESCO

Calcul des régressions à l'aide de l'algorithme simplex

Revue française d'informatique et de recherche opérationnelle. Série verte, tome 3, n° V2 (1969), p. 113-116

<http://www.numdam.org/item?id=RO_1969__3_2_113_0>

R.I.R.O.
(3e année, nº V-5, 1969, p. 113-116)

CALCUL DES REGRESSIONS A L'AIDE DE L'ALGORITHME SIMPLEX

par Dan Zorilesco (1)

Résumé. — *On présente dans cet article le procédé de la détermination des régressions de variables aléatoires à l'aide de l'algorithme Simplex de la théorie de la programmation linéaire.*

1. ETABLISSEMENT DU PROGRAMME LINEAIRE

On suppose le cas général d'une variable aléatoire « y » qui dépend du point de vue statistique d'un nombre borné $\{ x_i \}_{i \in N}$ où $N = \{ 1, 2 \ldots n \}$, de paramètres aléatoires.

La forme générale de la régression de y par rapport aux paramètres $\{ x_i \}_{i \in N}$ peut être formulée comme suit :

$$\tilde{y} = \sum_{i \in N} a_i f(x_i) \tag{1}$$

où les a_i sont les coefficients de régression, et les

$f(x_i)$ peuvent être les fonctions élémentaires de l'x_i (exponentielles, logarithmes, fonctions trigonométriques, etc.).

Par conséquent on ne demande pas la linéarité par rapport à x_i mais seulement par rapport aux coefficients de régression a_i.

Usuellement la détermination des coefficients de régression s'effectue à l'aide de la méthode des moindres carrés sur la base d'une sélection $\{ y_j ; x_{1j} ; x_{2j} ; \ldots x_{nj} \}_{j \in M}$ où $M = \{ 1, 2 \ldots m \}$.

Entre les valeurs mesurées $\{ y_j \}_{j \in M}$ et celles calculées par la relation (1) se trouvent les déviations

$$D_j = y_j - \tilde{y}_j \qquad \text{pour } j \in M \tag{2}$$

(1) Institut des Recherches Minières, Bucarest.

Pour une série des problèmes pratiques les déviations D_j ne sont pas à tout prix sous-unitaires (ainsi qu'il advient dans de nombreux problèmes géologiques-géophysiques, où D_j peut désigner les déviations locales par rapport à une tendance générale, régionale [2], [4]).

Mais le fait qu'une déviation surunitaire influence beaucoup plus une relation du type (1) par la méthode des moindres carrés, que si on se proposait de réduire au minimum la somme des déviations absolues, est universellement connu.

De plus, pour de nombreux problèmes pratiques la détermination des déviations est également nécessaire, soit pour déterminer l'exactitude de l'estimation à l'aide de la ligne de régression, soit à cause du fait que ces déviations ont un caractère physique qui doit être mis en évidence.

Dans le but de déterminer les déviations D_j en même temps que les coefficients de régression, nous proposons ci-dessous une résolution à l'aide de la théorie de la programmation linéaire.

La relation (2) peut être écrite aussi sous la forme suivante :

$$y_j = \sum_{i \in N} a_i f(x_{ij}) + D_j \qquad \text{pour tout } j \in M$$

c'est-à-dire un système de m équations linéaires de a_i ayant $n + m$ inconnues (n — inconnues a_i et m — inconnues D_j). Nous allons déterminer le vecteur des inconnues $(a_1, \ldots a_n, D_1 \ldots D_m)$ de telle sorte que la fonction linéaire

$$z = \sum_{j \in M} |D_j| \tag{4}$$

soit réduite au minimum.

La condition de la non-négativité des inconnues s'obtient aisément, [3] en remarquant qu'il est possible d'écrire chaque inconnue comme différence des parties positives et négatives, car l'algorithme Simplex est de telle nature que les deux parties ne sont jamais en même temps non-nulles. Par conséquent on peut noter :

$$a_i = a_i^+ - a_i^- \text{ ayant } a_i^+ \geqslant 0 \qquad \text{et} \qquad a_i^- \geqslant 0 \text{ pour tout } i \in N, \text{ et}$$

$$D_j = D_j^+ - D_j^- \text{ ayant } D_j^+ \geqslant 0 \text{ et } D_j^- \geqslant 0 \text{ pour tout } j \in M \tag{5}$$

Les relations (3) et (4) deviennent dans ce cas :

$$y_j = \sum_{i \in N} a_i^+ f(x_{ij}) - \sum_{i \in N} a_i^- f(x_{ij}) + D_j^+ - D_j^- \tag{6}$$

$$Z = \sum_{j \in M} D_j^+ + \sum_{j \in M} D_j^- \tag{7}$$

c'est-à-dire un problème de programmation linéaire standard ayant un nombre double d'inconnues $(2n + 2m)$.

Des programmes linéaires semblables ont été obtenus en résolvant certains problèmes de géologie et de géophysique dans les ouvrages [2], [4].

2. ESTIMATION DES COEFFICIENTS DE LA REGRESSION PAR LES SOUS-ENSEMBLES DES SELECTIONS ALEATOIRES

En pratique il est souvent possible qu'm, et par conséquent $m + n$, soient tellement grands qu'ils arrivent à dépasser la capacité de la mémoire des machines à calculer existantes.

Dans ce cas on prendra de l'ensemble total des sélections $\{ y_j ; x_{1j} ; x_{2j} ; \ldots x_{nj} \}$ certains sous-ensembles de sélection choisis aléatoirement et on résoudra un problème de programmation linéaire pour chaque sous-ensemble séparément.

On obtiendra alors pour les coefficients de régression a_i des valeurs différentes.

On désigne ces valeurs $a_i^{[K]}$ pour $K \in L = \{ 1, 2, \ldots l \}$ et $i \in N$.

Les coefficients de régression réels seront estimés à l'aide de la moyenne arithmétique :

$$\bar{a}_i = \frac{\sum_{K \in L} a_i^{[K]}}{l}.$$

De même on peut déterminer un intervalle de confiance pour les valeurs réelles des coefficients de régression de la forme :

$$\bar{a}_i \pm t_q \frac{S}{\sqrt{l}}$$

où S^2 est l'estimation de la variance :

$$S^2 = \frac{\sum_{K \in L} (a_i^{[K]} - \bar{a}_i)^2}{l - 1}$$

et t_q est un coefficient catalogué en fonction du seuil de signification q et du nombre des degrés de liberté l, dans le tableau des répartitions de « Student ».

REMARQUE

La détermination des coefficients de régression, à condition que la réduction au minimum de la somme des déviations quadratiques engendre un programme qui peut être résolu à l'aide des algorithmes de la programmation quadratique.

3. CONCLUSIONS

Pour la résolution du problème des régressions on est arrivé au programme linéaire, programme qui dans certains cas pratiques convient mieux que le programme obtenu par la réduction au minimum de la somme des déviations quadratiques.

La résolution peut être aisément effectuée à l'aide d'un calculateur électronique.

L'avantage résulte du fait qu'on obtient les déviations des valeurs, calculées par rapport à celles mesurées, en même temps que les coefficients de régression.

De plus, le calcul des éléments de la matrice du problème de la programmation linéaire est beaucoup plus facile que celui des éléments de la matrice du système d'équations normales obtenu par la méthode des moindres carrés.

BIBLIOGRAPHIE

[1] H. CRAMER, *Mathematical methods of statistics*, Princeton University Press, 1946.

[2] M. IANAS et D. ZORILESCU, *Some application of the linear programming within gravimetry*. Pure and applied Geophysics, vl. 72, I, 1969.

[3] M. SIMONNARD, *Programmation linéaire*, Dunod, Paris, 1962.

[4] D. ZORILESCU, La représentation mathématique des phénomènes géologiques. *Annales des Mines*, septembre 1968.

Le directeur de la Publication : Georges DUNOD. — *Imprimé en France.*
Dépot légal : 4e trimestre 1969. N° 6359.
IMPRIMERIE NOUVELLE, ORLÉANS. — N° 6056.