LES CAHIERS DE L'ANALYSE DES DONNÉES

J. P. BENZÉCRI

Contre-exemple à l'estimation des coefficients dans une régression linéaire

Les cahiers de l'analyse des données, tome 10, n° 3 (1985), p. 303-304

<http://www.numdam.org/item?id=CAD_1985__10_3_303_0>

Les Cahiers de l'Analyse des Données
Vol. X - 1985 - n° 3 - pp. 303-304

CONTRE-EXEMPLE A L'ESTIMATION DES COEFFICIENTS DANS UNE RÉGRESSION LINÉAIRE

[CONTRE REG.]

par J.P. Benzécri *

1 *Le modèle* : On désigne par x et y deux variables explicatives décrivant un objet aléatoire ; et par z une variable à expliquer (relative au même objet). On suppose que x et y sont très étroitement liées par une relation quasi linéaire qu'on note :

$$y = x + \varepsilon x^2 + \text{bruit1} \; ; \qquad (1) \; ;$$

on peut, pour fixer les idées, postuler que x est distribué uniformément entre 0 et 19.

On suppose de plus que z est une fonction presque linéaire de x et y :

$$z = x + y + \eta x^2 + \text{bruit2} \; ; \qquad (2).$$

On cherche une formule de régression linéaire :

$$z = ax + bx + \text{bruit3} \; ; \qquad (3)$$

bruit3 étant à minimiser.

2 *Etude du cas limite où les relations algébriques sont strictement vérifiées* : Faisons l'hypothèse complémentaire que les quantités aléatoires notées bruit1 et bruit2 sont strictement nulles, en sorte que les relations (1) et (2) sont des identités algébriques. Il est alors facile de déterminer les coefficients de régression a et b. On a :

$$ax + by = (a + b)x + b\,\varepsilon x^2 \; ;$$

$$z - (ax + by) = (x + x + \varepsilon x^2 + \eta x^2) - (a + b)x - b\,\varepsilon x^2$$

$$= (2 - (a + b))x + (\varepsilon + \eta - b\varepsilon)x^2 = \text{bruit3}.$$

On annule bruit3 en posant :

$$a + b = 2 \quad ; \quad b = 1 + (\eta/\varepsilon).$$

Si, par exmple η et ε sont opposés ($\eta = -\varepsilon$), on aboutit à :

$$(a, b) = (2, 0) \; ;$$

Et il est facile d'imaginer les diverses issues possibles, selon que l'on impose ou non à (a, b) la contrainte de positivité.

(*) *Professeur de statistique. Université Pierre et Marie Curie.*

3 *Conséquences pour le cas général où les relations ne sont qu'approchées* : En reprenant le calcul du § 2, il vient :

$$ax + by = (a + b)x + b\,\varepsilon x^2 + b\ \text{bruit1} ;$$

$$z = 2x + (\varepsilon+\eta)x^2 + \text{bruit1}' + \text{bruit2} ;$$

$$\text{bruit3} = z-(ax + by)$$

$$= (2-(a+b)x + (\varepsilon+\eta - b\varepsilon)x^2 + (1-b)\text{bruit1} + \text{bruit2}.$$

Si bruit1 l'emporte sur ηx^2, on aboutira à une estimation de b voisine de 1, avec a voisin de 1 également. Si, au contraire ηx^2 l'emporte sur bruit1, on est ramené au cas précédent (§ 2) où bruit1 = 0 ; (bruit2 ne jouant, quant à lui aucun rôle, s'il est indépendant de x et de bruit1). Si bruit1 et ηx^2 sont du même ordre, on peut avoir pour (a ,b) une estimation différant notablement de (1 ,1).

4 *Commentaires* : Même si (a ,b) ≈ (2 ,0), la formule de régression obtenue est satisfaisante en ce qu'elle minimise véritablement l'erreur bruit3 ; il ne s'agit donc pas d'une régression illusoire comme celles obtenues quand le nombre de variables explicatives dépasse le nombre des individus de l'échantillon (cf. [*REGR. GEOM.*] CAD Vol III n° 2, 1978).

Cependant, si l'on s'intéresse aux coefficients a, b eux-mêmes, le résultat (b = 0), peut être absurde. C'est en particulier le cas si z est le coût global de l'objet, et x et y sont les deux principaux postes de dépense (e.g. la matière et la main d'oeuvre...) d'après lesquels on peut prétendre estimer l'ensemble des postes.

Un problème réaliste est celui de l'estimation du prix de revient r, des actes élémentaires (qu'on supposera pour fixer les notations être de trois types 1, 2, 3) d'après le nombre ceux-ci rentrant dans des opérations complexes dont on connaît le coût global z . Dans ce cas, il est naturel de postuler une relation telle que :

$$z = r1\ n1 + r2\ n2 + r3\ n3 + \text{bruit} \quad ; \qquad (4) ;$$

dans laquelle n1, n2, n3 sont des variables explicatives connues, et les prix de revient r1, r2, r3 jouent le rôle de coefficients de régression. D'après le modèle étudié ci-dessus (§§ 1, 2, 3), on voit que si n1, n2, n3 sont fortement liés entre eux (éventualité qui n'est nullement à écarter : cf. A. Skalli Vol 10 n°3 pp 305-310) on ne peut se fier aux méthodes de régression pour estimer les prix de revient r1, r2, r3 des actes élémentaires.

Dans le langage des spécialistes de la régression, le résultat b = 0 obtenu au § 2 peut s'interpréter comme une régression par choix de variable pas à pas ; la variable x étant ici la seule choisie ! Ce résultat ne suffira pas à réconcilier avec les méthodes de pas à pas ceux qui comme nous demandent d'abord à l'analyse des données une vision synthétique de l'ensemble des variables. Si l'on veut une formule simple mettant en jeu peu de variables, on pourra la construire ensuite, en connaissance de cause. Dans la pratique, le choix des variables explicatives retenues requiert d'autant plus une vision globale, que ce choix résulte d'un compromis entre la qualité de l'approximation atteinte, et le coût des variables ; (coût en argent ; ou coût en souffrance ; s'il sagit d'un problème médical). Or à notre connaissance, les programmes de régression pas à pas ne tiennent pas compte de tels coûts. En tiendraient-ils compte (par exèmple en introduisant comme critère de choix le rapport "information acquise"/coût), qu'il nous paraîtrait encore imprudent de livrer un choix essentiel au déroulement aveugle d'un programme.