L. V. De Carolis

La méthode des moindres volumes dans les prévisions économiques

Statistique et analyse des données, tome 2, nº 2 (1977), p. 73-86.
<http://www.numdam.org/item?id=SAD_1977__2_2_73_0>

LA METHODE DES MOINDRES VOLUMES DANS LES PREVISIONS ECONOMIQUES*

L.V. DE CAROLIS

Université de Rome

I - INTRODUCTION

Ce qui suit se réfère au problème mis en evidence par G. Pompilj (1) et par T. Salvemini (2), d'introduire dans l'étude de la variabilité et de la régression d'une variable réelle à n composantes des mesures qui generalisent la distance euclidienne. Chaque mesure est associée à un ensemble de n points linéairement indépendants (les n points de l'ensemble appartiennent à un espace ayant dimensions $>$ n-2). De cette façon la distance euclidienne est un caractère métrique associé à chaque couple de points (linéairement indépendants) à chaque triplet de points est associée la surface du triangle qui a pour sommets les trois points, à chaque quaterne est associé le volume du tétraèdre qui a pour sommets les quatre points et finalement à chaque n-uple de points est associé le volume du simplexe qui a pour sommets les n points. On peut appliquer ces reflexions au problème de l'ajustement d'une fonction mathématique à une distribution observée. Particulièrement considérons des observations dans l'espace euclidien à n dimensions auxquelles il est possible d'ajuster un hyperplan π à n-1 dimensions. Pour determiner les coefficients de l'equation de

* Conférence faite au Colloque International "Structures Economiques et Econometriques" Lyon 22-24 Avril 1976.

l'hyperplan la méthode la plus généralement employée est celle des <u>moindres distances</u>. Selon ce crithère on minimise une fonction des distances comptées parallèlement à une direction, entre points observés P_i $(x_{i1}, x_{i2}, \ldots x_{in})$ et points correspondants $\bar{P}_i$ de π. Comme cas particulier on obtient <u>la méthode des moindres carrés</u>. Mais on peut remarquer que chaque point P_i est l'un des n+1 sommets d'un simplexe dont les autres n sommets sont les intersections avec π des droites passant par P_i et parallèles aux axes des coordonnées. On peut donc ajuster à la distribution observée l'hyperplan qui minimise une fonction des volumes des simplexes correspondants à chaque point P_i. Sur la base de ces considérations j'ai introduit <u>la méthode des moindres volumes</u> (3) que j'ésporrai brièvement en II. Ce crithère employé, comme je propose ici, dans le domaine des prévisions économiques peut apporter une amélioration de la prévision lorsque certaines conditions sont verifiées.

II - LES HYPERPLANS DES MOINDRES VOLUMES

Exposons brièvement la méthode des moindres volumes que j'ai introduite en 1973 (3). Considérons des observations dans l'espace euclidien à n dimensions auxquelles il est possible d'ajuster un hyperplan à n-1 dimensions. Pour déterminer les coefficients de l'équation de l'hyperplan on peut employer la méthode des moindres volumes. Par ce procédé, on minimise la valeur moyenne $M(|V|^{\frac{2}{n}})$ des volumes pris en valeur absolue et élevés à la puissance $\frac{2}{n}$, de simplexes de l'espace à n dimensions. L'un des n+1 sommets de ces simplexes est chacun des points P_i observés et les autres n (sommets) sont les intersections avec un hyperplan à n-1 dimensions des droites passant par chaque point P_i et paralleles aux axes des coordonnées. Il y a 2^{n-1} hyperplans réels des moindres volumes dont les coefficients sont les solutions d'un système de n-1 équations du deuxième degré. Les coefficients de chaque hyperplan diffèrent des coefficients des autres en valeur et signe. Parmi ces hyperplans nous ajusterons aux observations celui auquel correspond le minimum absolu de la $M(|V|^{\frac{2}{n}})$; cet hyperplan a les coefficients du même signe que l'hyperplan des moindres carrés.

Je traiterai d'abord des advantages que le procedé exposé peut apporter dans les prévisions économiques par rapport à la méthode des moindres carrés en me référant à des distribution à deux variables. Je ferai après quelque généralisation en me référant à des distributions à trois variables.

III - LES DROITES DES MOINDRES SURFACES ET LES PREVISIONS ECONOMIQUES.

III.1. Les droites des moindres surfaces

Supposons que l'on ait n observations (x_i, y_i) des variables x et y et que l'on s'interesse aux variations de y par rapport à x. On se propose d'ajuster aux n points observés $P_i(x_i, y_i)$ une droite. La méthode qui vient d'être exposée, conduit à la determination de la droite qui minimise la valeur moyenne des surfaces, prises en valeur absolue et élevées à la puissance $\frac{2}{2}$, de triangles tels que chacun des points P_i observés est l'un des trois sommets, les autres sont les intersections, avec une droite, des droites passant par chaque point P_i, parallèles aux axes des x et des y (figure 1)

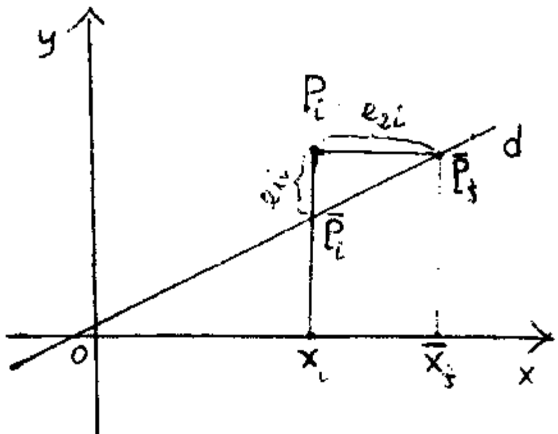

figure 1

De cette façon on considère, en déterminant les surfaces des triangles $\bar{P}_i\ P_i\ \bar{P}_j$, les deux erreurs e_{1i} et e_{2i}. L'erreur e_{1i} utilisée dans la méthode des moindres carrés est l'erreur que l'on commet en remplaçant le point $P_i(x_i, y_i)$ par le point $\bar{P}_i(x_i, \bar{y}_i)$ sur la droite d, c'est à dire en remplaçant l'ordonnée y_i de P_i par l'ordonnée $\bar{y}_i$ de $\bar{P}_i$. L'erreur e_{2i} est celle que l'on commet en ne remplaçant pas le point P_i par le point $\bar{P}_j$ sur d, c'est à dire en considérant l'abscisse x_i de P_i au lieu de l'abscisse $\bar{x}_j$ de $\bar{P}_j$ qui a l'ordonnée $\bar{y}_j = y_i$. On a ainsi les droites des moindres surfaces d_s(4). Ces droites ont pour équations:

$$d_s) \qquad y = \pm \frac{\sigma_y}{\sigma_x} (x - m_x) + m_y$$

où m_x, σ_x, m_y et σ_y sont les moyennes et les écarts types des observations x_i et y_i. Les droites d_s ont un coefficient angulaire superieur, en valeur absolue, à celui de la droite des moindrés carrés d_c d'équation:

$$d_c) \qquad y = r_{xy} \frac{\sigma_y}{\sigma_x} (x - m_x) + m_y$$

où r_{xy} est le coefficient de correlation entre les observations x_i et y_i. Des deux droites d_s, il faut ajuster à la distribution observée celle qui a un coefficient angulaire positif si $r_{xy} > 0$, negatif si $r_{xy} < 0$, l'une ou l'autre si $r_{xy} = 0$.

III.2. <u>Prévisions économiques</u>

Passons maintenant aux applications dans le domaine des prévisions économiques.

Prenons des prévisions économiques relatives à une distribution à deux variables x, y (on suppose que seuls les y sont atteints d'erreurs) à laquelle on peut ajuster une droite; la méthode la plus comunément utilisée est celle des moindres carrés. Mais il convient d'avoir à côté de la valeur fournie par cette technique, d'autre valeurs fournies par des ajustements basées sur diverses hypothèses. On obtient ainsi un ensemble de valeurs entre lesquelles on peut choisir la valeur moyenne comme approximation de la valeur à prévoir.

Examinons, en nous référant aux prévisions à court terme (cinq ans au plus) quelques hypothèses qui, si elles sont verifiées, amènent à considérer, en plus de la valeur qui nous est donnée par la droite des moindres carrés, celle fournie par la droite des moindres surfaces, pour former un intervalle qui généralement, apporte une amélioration de la prévision. Ajoutons que, parmi les diverses droites ajustées que nous pourrions considérer sur la base des hypothèses que nous formulerons.ci-aprés, celle des moindres surfaces semble être la plus adaptée à la formation de cet intervalle, justement à cause du critère d'ajustement employé pour la détermination de la droite.

Examinons donc certains cas (α_1, α_2, β_1, β_2) dans lesquels il est possible de démontrer que les valeurs données dans les prévisions économiques par la droite d_s sont préférable à celle données par la droite d_c.

α) Supposons qu'aux n observations (x_i, y_i) il soit possible d'ajuster une droite de coefficient angulaire positif. Par conséquent les droites d_c et d_s seront du type de celles de la figure numero 2 suivante où ont été tracés des points et des segments que nous allons utiliser.

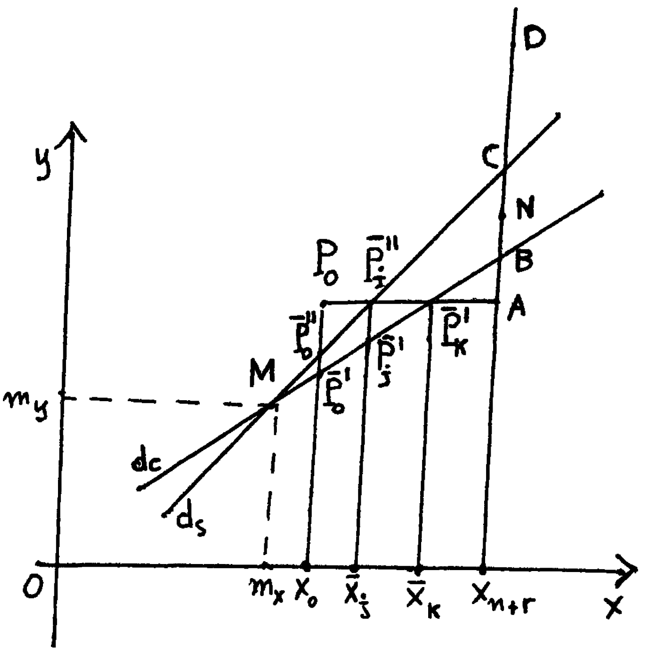

figure 2

On peut avoir deux cas:

α_1) La distribution observée montre que les y_i augmentent en augmentant les x_i. En ce cas, par rapport à une valeur $x_{n+r} > m_x$ nous voulons faire une prévision sur y_{n+r} inconnu au moyen de la valeur donnée par l'une des deux droites. Dans ces hypothèses supposons qu'il existe une point P_o satisfaisant aux conditions:

1) P_o est le dernier point appartenant à la distribution tel que
$x_o < x_{n+r}$; $y_o > m_y$;

2) ($y_o > y'_o$ et y''_o) où y'_o et y''_o sont les ordonnées des points $\bar{P}'_o$ et $\bar{P}''_o$ d'abscisse x_o respectivement sur d_c et d_s;

3) les écarts de y_o par rapport à y'_o et y''_o sont petits (l'on peut employer pour les mesurer les rapports $i_1 = \frac{|y_o - y'_o|}{y_o} \times 100$, $i_2 = \frac{|y_o - y''_o|}{y_o} \times 100$);

4) les causes fondamentales qui agissent sur le phénomène conservent la même téndance, ce qui permet raisonnablement de prévoir $y_{n+r} > y_o$.

En ce cas, la droite d_s donne plus probablement une valeur plus voisine de y_{n+r} que la droite d_c. En effet, l'on peut remarquer en observant la figure numero 2 que si $x_o < x_{n+r} \leq \bar{x}_j$ les deux droites donnent des valeurs y'_{n+r} et y''_{n+r} inacceptables parce qu'inférieures ou égales à y_o. Si $\bar{x}_j < x_{n+r} \leq \bar{x}_k$ des deux droites la d_s est préférable vu que la droite d_c donne une valeur inférieure ou égale à y_o. Finalement si $x_{n+r} > \bar{x}_k$, nous pouvons observer que si le point $P_{n+r}(x_{n+r}, y_{n+r})$ inconnu appartient au segment AN, où N est le point moyen du segment BC, y_{n+r} est mieux estimé par y'_{n+r}; si P_{n+r} appartient au segment ND, y_{n+r} est mieux estimé par y''_{n+r}; mais puisque nous pouvons toujours prendre ND $>$ AN, concluons qu'il est plus probable que la droite d_s donne une valeur y''_{n+r} plus voisine de y_{n+r} que la droite d_c. Si, cependant, nous voulons effectuer une analyse plus détaillée nous pouvons observer qu'il est toujours plus probable que la droite d_s donne une valeur plus voisine de y_{n+r} que la droite d_c dans la sous-estimation de y_{n+r}, car il est suffisant de prendre CD $>$ BN. Dans la sur-estimation de y_{n+r}, il faut comparer le segment AB avec le segment NC. A ce propos on peut, par exemple, calculer le rapport:

$$Q = \frac{AB}{NC} = \frac{a_1(x_{n+r} - m_x) + m_y - y_o (= AB)}{(x_{n+r} - m_x)(a_2 - a_1)/2 \ (=NC)}$$

où a_1 et a_2 sont les coefficients angulaires des droites d_c et d_s. Si $Q < 1$ la droite d_s est préférable dans la sur-estimation; si $Q > 1$ la droite d_c est préférable. En tout cas il est à conseiller de considérer l'intervalle entre y'_{n+r} et y''_{n+r} auquel probablement y_{n+r} appartient ou, encore mieux, l'intervalle entre y_o et y''_{n+r} + CD où nous avons posé CD = AB et considérer pour la prévision l'ordonnée y_N du point N. En faisant des considérations graphiques du type qui vient d'être exposé, l'on peut démontrer que la droite d_s donne aussi des valeurs préférables à celles données par la droite d_c dans le cas α_2, β_1 et β_2 suivants:

α_2) La distribution observée montre que le y_i décroissent lorsque les x_i décroissent. En ce cas, par rapport à une valeur $x_{n+r} <$ m_x, nous voulons faire une prévision sur y_{n+r} inconnu au moyen de la valeur donnée par l'une des deux droites. Dans cette hypothèse, supposons qu'il existe un point $P_o(x_o, y_o)$ satisfaisant aux conditions suivantes:

1) P_o est le dernier point appartenant à la distribution tel que $x_o > x_{n+r}$; $y_o < m_y$;

2) $y_o < y'_o$ et y''_o;

3) les écarts de y_o par rapport à y'_o et y''_o sont petits;

4) on peut raisonnablement supposer $y_{n+r} < y_o$

β) Supposons qu'aux n observations (x_i, y_i), il soit possible d'ajuster une droite de coefficient augulaire négatif:

β_1) La distribution observée montre que les y_i décroissent lorsque les x_i augmentent. En ces cas - là par rapport à une valeur $x_{n+r} > m_x$, nous voulons faire une prévision sur y_{n+r} au moyen de la valeur donnée par l'une des deux droites. Dans cette hypothèse supposons qu'il existe un point $P_o(x_o, y_o)$ satisfaisant aux conditions suivantes:

1) P_o est le dernier point appartenant à la distribution tel que $x_o < x_{n+r}$; $y_o < m_y$;
2) $y_o < y'_o$ et y''_o;
3) les écarts de y_o par rapport à y'_o et y''_o sont petits;
4) on peut raisonnablement supposer $y_{n+r} < y_o$

β2) La distribution observée montre que les y_i augmentent lorsque le x_i décroissent. En ce cas-là, par rapport à une valeur $x_{n+r} < m_x$, nous voulons faire une prévision sur y_{n+r} au moyen de la valeur donnée par l'une des deux droites. Dans cette hypothèse, supposons qu'il existe un point $P_o(x_o, y_o)$ satisfaisant aux conditions suivantes:

1) P_o est le dernier point appartenant à la distribution tel que $x_o > x_{n+r}$; $y_o > m_y$;
2) $y_o > y'_o$ et y''_o;
3) les écarts de y_o par rapport à y'_o et y''_o sont petits
4) on peut raisonnablement supposer $y_{n+r} > y_o$

III.3. Application numérique

Considérons dans le tableau ci-aprés le développement de 1948 à 1956 de la valeur ajoutée nette italienne, Administration publique exclue, des activités tertiaires, c'est à dire de toutes les activités, celles agricoles et industrielles exclues

(5). Il faut exprimer les observations en lires ayant un pouvoir d'achat inchangé; on a donc choisi le pouvoir d'achat de l'année 1953.

Tableau - Valeur ajoutée nette italienne en milliards de 1948 à 1956 des activités tertiaires

Années x_i	Renevu en lires 1953 y_i	Valeurs ajustées par d_c : y'_i	Valeurs ajustées par d_s : y''_i
(1)	(2)	(3)	(4)
1948(0)	1568	1583,964	1579,879
1949(1)	1706	1693,928	1691,205
1950(2)	1864	1803,893	1802,531
1951(3)	1850	1913,857	1913,857
1952(4)	2020	2023,821	2025,183
1953(5)	2128	2133,786	2136,509
1954(6)	2261	2243,750	2247,835
1955(7)	2397	2353,714	2359,161
1956(8)	2530	2463,679	2470,487

$m_x = 3$; $m_y = 1913,85$

Valeurs calculées sur les sept premières observations

En considérant lessept premièrs couples observés on veut faire une prévision sur le revenu des années 1955 et 1956 en supposant que les causes structurelles qui agissent sur le phénomène, restentinchangées.

On a donc ajusté aux sept premières observations, les droites d_c et d_s d'équations:

d_c) $y = 109,96\ (x - 3) + 1913,85$

d_s) $y = 111,33\ (x - 3) + 1913,85$

Les valeurs y'_i fournies par d_c et les y''_i fournies par d_s se trouvent dans les colonnes (3) et (4) du tableau. Observons que les deux droites s'approchent bien de la distribution, comme on peut le contrôler au moyen, par exemple, des rapports

$$I_1 = \frac{\sum |y_i - y'_i|}{\sum y_i}, \quad I_2 = \frac{\sum |y_i - y''_i|}{\sum y_i}$$

On a en effet $I_1 = 0,01335 = I_2$. Par conséquent en nous référant au cas α_1 observons qu'il existe un point $P_o = P_{1954}$ appartenant à la distribution dont les coordonnées (6,2261) satisfont les conditions 1-2-3-4. Par suite, concluons que pour la prévision du revenu des années 1955 et 1956, il est préférable de considérer les ordonnées fournies par les points de d_s. Comme l'on peut noter dans le tableau, la droite d_s donne effectivement, par rapport à ces deux années, des valeurs plus voisines de celles observées que la droite d_c. Si en outre, parce que $Q = 23,6 > 1$, nous voulons déterminer, en nous référant par exemple à l'année 1955, un intervalle dans lequel il est probable que $y_{n+r} = y_{1955}$ soit compris, nous pouvons considérer l'intervalle entre y_o et $y''_{n+r} + AB$. L'on a $y_o = 2261$, $y''_{n+r} + AB = 2359,161 + 92,70 = 2451,861$. Remarquons ainsi que cet intervalle contient la valeur 2397 de y_{n+r} et que la ordonnée $y_N = 2356,43$ du point moyen N s'approche de y_{n+r} mieux que la valeur donnée par la droite d_c.

IV. LES PLANS DES MOINDRES VOLUMES

Concluons en exposant quelques considérations analogues au sujet de distributions à trois variables; une analyse plus détaillée est developpée en (3).

Supposons d'avoir n observations (x_i, y_i, z_i) des variables x, y, z. On s'intéresse, par exemple, aux variations de z par rapport à (x, y) et supposons que l'on ajuste un plan aux n point observés $P_i(x_i, y_i, z_i)$ de l'espace à trois dimensions. La méthode exposée dans le paragraphe II conduit à déterminer $2^{3-1}=4$ plans π_V des moindres volumes ayant pour équations:

$$z = a_{1i}(x - m_x) + a_{2i}(y - m_y) + m_z$$

où les a_{1i} et a_{2i} $(i = 1,2,3,4)$ sont les solutions des deux équations du deuxième dégré ci-après:

$$a_1^2 \sigma_x^2 - a_2^2 \sigma_y^2 - a_1 \sigma_{xz} + a_2 \sigma_{yz} = 0$$

$$a_1^2 \sigma_x^2 + a_1 a_2 \sigma_{xy} + a_2 \sigma_{yz} - \sigma_z^2 = 0$$

(σ_{xy}, σ_{xz} et σ_{yz} étant les covariances empyriques entre couples de phénomènes).

Les solutions du système sont réelles et ont pour signes (+, +), (+, -), (-, +) (-, -). Parmi les quatre plans nous ajusterons à la distribution observée celui dont les coefficients ont les mêmes signes que ceux du plan des moindres carrés π_c.

Supposons que par rapport au couple (x_{n+r}, y_{n+r}), nous voulons faire une prévision sur z_{n+r} inconnu au moyen de la valeur donnée par l'un des deux plans. Nous pouvons conclure, en généralisant à l'espace à trois dimensions les considérations exposées dans le paragraphe III.2 que, sous certaines conditions il est preferable de tenir compte de l'intervalle entre la valeur z'_{n+r} fournie par π_c et la valeur z''_{n+r} fournie par π_V.

REFERENCES

(1) POMPILJ G. Calcolo delle probabilità, University of Pittsburg, a.a. 1963-64.

On the average of multiple random variables, Actes de la 35e session de l'Institut International de Statistique, Belgrade 1965.

(2) SALVEMINI T. Lezioni di statistica metodologica, Università di Roma, a.a. 1967-68.

(3) DE CAROLIS, L.V. Il metodo del minimo volume nell'interpolazione, Metron vol. XXXI, N. 1-4, 1973.

(4) GALLO BALDESSARI F. Rette d'area minima, La scuola in azione, Anno 1963-64 n. 8.

(5) LIVI L. Previsioni economiche, Edizioni scientifiche, Einaudi, 1958.