G. E. P. Box

Du bon et du mauvais usage de la régression

Revue de statistique appliquée, tome 14, nº 3 (1966), p. 25-29
<http://www.numdam.org/item?id=RSA_1966__14_3_25_0>

DU BON ET DU MAUVAIS USAGE DE LA RÉGRESSION (1)

G.E.P. BOX
Department of Statistics
University of Wisconsin, Madison, Wisconsin
texte traduit par A. Vessereau

On rappellera tout d'abord les hypothèses classiques et les conclusions de la méthode des moindres carrés. Gauss a montré que si l'on dispose de n observations $y_1, y_2, \ldots, y_n$ et sous la condition que le modèle approprié pour la $u^{ème}$ observation soit :

$$y_u = \beta_o + \beta_1 x_{1u} + \beta_2 x_{2u} + \ldots + \beta_k x_{ku} + \varepsilon_u \quad (1)$$

les β étant des paramètres inconnus, les x des constantes connues, et les ε des variables aléatoires sans corrélation, de même variance et d'espérance mathématique nulle, les estimations $b_o, b_1, \ldots, b_k$ des β obtenue en minimisant la quantité :

$$\sum(y - \hat{y})^2 \text{ avec } \hat{y} = b_o x_o + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k$$

sont des estimations sans biais et ont la plus petite variance parmi toutes les estimations linéaires sans biais.

La méthode des moindres carrés est utilisée dans l'analyse de données provenant d'expériences organisées, et aussi dans l'analyse de données provenant de simples observations. Le mot de "régression" est le plus souvent utilisé pour décrire l'analyse de simples observations. C'est l'hypothèse tacite que les conditions requises pour la validité de la méthode des moindres carrés sont satisfaites pour des données non planifiées qui est à l'origine du maximum de confusion. Que les données proviennent ou non d'expériences organisées, la quantité ε, dont généralement on oublie vite qu'elle représente une variable aléatoire ayant les propriétés très particulières mentionnées ci-dessus, décrit en réalité l'effet d'un grand nombre de variables "latentes" $x_{k+1}, x_{k+2}, \ldots, x_m$ au sujet desquelles on ne sait rien. Si l'on admet qu'il est suffisant de considérer les effets linéaires de ces variables latentes (ce qui sera souvent convenable pour de petites variations en $x_{k+1}, \ldots, x_m$), on aura :

$$\varepsilon = \beta_{k+1} x_{k+1} + \beta_{k+2} x_{k+2} + \ldots + \beta_m x_m \quad (2)$$

En notation matricielle, la matrice colonne des n observations y s'écrira :

$$\underline{y} = \underline{X}_1 \underline{\beta}_1 + \underline{X}_2 \underline{\beta}_2 \quad (3)$$

----

(1) Exposé préparé pour la Xe Conférence sur "The Design of Experiments in Army Research, Development and Testing" Washington D.C., 5 Novembre 1964.

où la matrice $\underline{X_1}$ a pour éléments les n valeurs des k variables de régression, et la matrice $\underline{X_2}$ les n valeurs inconnues des m – k variables latentes. Cette situation est illustrée dans la figure 1, où les varia-

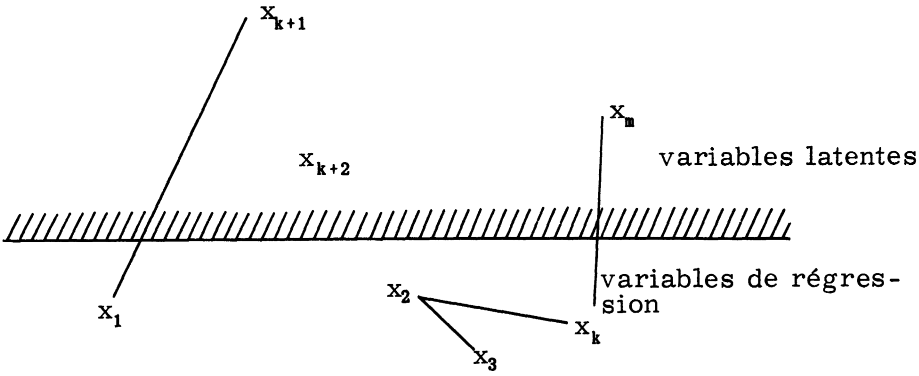

Figure 1 - Variables latentes et variables de régression

bles x_{k+1} . . . . , x_m sont "cachées derrière le mur". Dans la pratique, plusieurs sortes de liaisons peuvent exister entre les variables réunies par des lignes. Ces liaisons peuvent être de nature causale ; par exemple, il peut se faire qu'une augmentation de température produise nécessairement une augmentation de pression ; ou bien elles peuvert traduire de simples corrélations. Ainsi, celui qui conduit une fabrication, peut, comme <u>procédure standard</u>, réduire l'arrivée d'un des réactifs, chaque fois qu'il constate une température élevée.

On doit maintenant se poser la question "A quoi désire-t-on utiliser l'équation de régression ?". On peut :

(i) désirer prévoir la valeur de y dans le futur à partir des observations passives de x_1 . . . . x_k. On admet que l'ensemble des relations causales et corrélatives qui existaient lors de la collecte de ces observations ne s'est pas modifié et continuera à intervenir pendant la période où les prévisions seront faites.

(ii) chercher à savoir comment des <u>modifications</u> volontaires dans les x_1 . . . x_k affectent y, dans l'intention de <u>modifier</u> réellement le système de façon à obtenir une meilleure valeur pour y .

La position est tout à fait différente, suivant qu'on a l'intention de prévoir à partir d'observations passives, ou d'améliorer par une intervention active. Ceci va être éclairé par l'exemple suivant

Supposons que, dans un processus chimique, on a constaté que la production d'une mousse indésirable peut être réduite par une augmentation de la pression. La procédure opératoire standard sera d'augmenter la pression chaque fois que de la mousse apparaîtra. Supposons que la mousse se produise en fait à cause d'une impureté insoupconnée x_2 (qui, naturellement, n'est pas mesurée, puisqu'elle est inconnue). Supposons enfin qu'une valeur élevée de l'impureté x_2, non seulement produit de la mousse, mais aussi diminue le rendement, alors que le rendement n'est pas directement modifié par un changement de pression.

Désignant maintenant par x et y les écarts par rapport aux moyen-

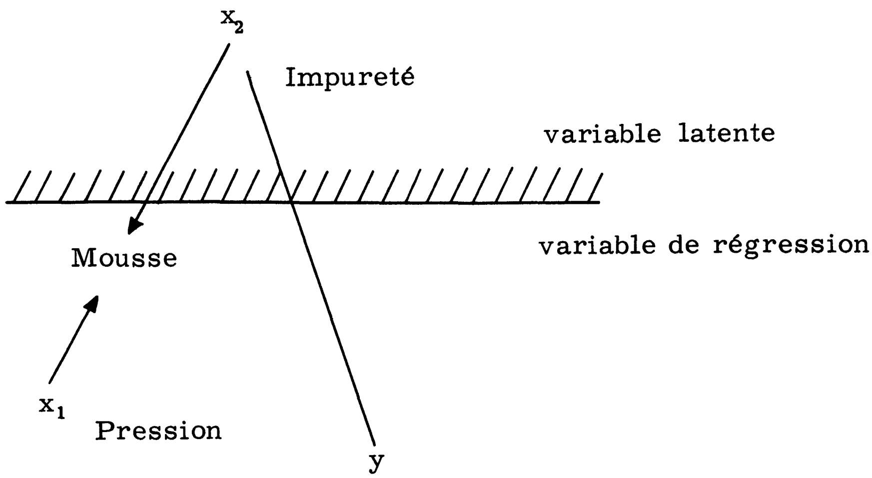

Figure 2 - Relations entre rendement, impureté et pression

nes, si une "équation de régression du rendement par rapport à la pression" $\hat{y} = b_1x_1$ a été obtenue par la méthode habituelle des moindres carrés, il se peut fort bien qu'on trouve un coefficient b_1 hautement significatif.

Le phénomène bien connu de corrélation dépourvue de sens qui apparaît dans cet exemple mérite d'être étudié plus à fond. La vraie relation $y = \beta_1x_1 + \beta_2x_2$ qui relie y aux deux variables x_1 et x_2 est en fait telle que $\beta_1 = 0$. Naturellement, les niveaux réels de x_2 sont inconnus, mais supposons que l'estimation de x_2 par les moindres carrés que l'on obtiendrait en ajustant x_2 par rapport à x_1 soit $\hat{x}_2 = a\, x_1$. Alors, on peut montrer facilement que :

$$b_1 = \beta_1 + a\,\beta_2 \tag{4}$$

Dans cette expression β_1 est égal à 0, et si l'on constate un effet réel, c'est uniquement à cause de l'existence du biais ($a\,\beta_2$). D'autre part, utilisant la relation (4) on voit que l'équation d'ajustement $\hat{y} = b_1x_1$ où x_2 est ignoré peut s'écrire $\hat{y} = \beta_1x_1 + \beta_2\, a\, x_1$ ou encore

$$\hat{y} = \beta_1x_1 + \beta_2\,\hat{x}_2 \tag{5}$$

Cette équation, dans laquelle x_2 est remplacé par $\hat{x}_2$ fournit la meilleure estimation de y que l'on peut espérer obtenir en observant x_1 seul. A condition que le système <u>continue à évoluer de la même façon que lorsque les données ont été recueillies</u>, on peut utiliser la pression pour obtenir le niveau de y. Généralement, il est vrai, la prévision ne sera pas aussi bonne que si l'on avait mesuré x_2 ; mais, ne connaissant pas l'existence de x_2 (ou dans d'autres cas son importance), il est néanmoins possible de l'utiliser.

Par contre, la valeur de b_1 serait complètement erronée, si on l'interprétait comme l'effet sur la variable y d'une unité de <u>modification</u> sur x_1. Si nous souhaitons augmenter le rendement en augmentant la pression, nous courons à une déception.

Un argument analogue s'applique à un nombre quelconque de variables. le modèle vrai est :

$$\underline{y} = X_1\beta_1 + X_2\beta_2 \tag{6}$$

En ne faisant figurer que les variables X_1 dans l'équation de régression, l'équation de prévision pour y devient :

$$\hat{\underline{y}} = \underline{X}_1\underline{b}_1 = \underline{X}_1(\underline{X}_1'\underline{X}_1)^{-1}\ \underline{X}_1'\ y \tag{7}$$

$$= \underline{X}_1(\underline{X}_1'\underline{X}_1)^{-1}\ \underline{X}_1'(\underline{X}_1\underline{\beta}_1 + \underline{X}_2\underline{\beta}_2) \tag{8}$$

c'est-à-dire :

$$\hat{\underline{y}} = \underline{X}_1\underline{\beta}_1 + \underline{\hat{X}}_2\underline{\beta}_2 \tag{9}$$

où $\underline{\hat{X}}_2 = \underline{X}_1\underline{A}$ et $\underline{A} = (\underline{X}_1'\underline{X}_1)^{-1}\ \underline{X}_1'\underline{X}_2$ est la matrice (k + 1, m – k) des coefficients de régression des variables latentes sur les variables de régression. On voit encore que, dans la mesure où l'on ne s'intéresse qu'à la prévision passive de $\hat{y}$, la régression sur les seules variables connues X_1 a pour effet de remplacer les $\underline{X_2}$ inconnus par $\underline{\hat{X}_2}$.

Par ailleurs, les coefficients de régression $\underline{b_1} = \beta_1 + A\ \beta_2$ représentent des combinaisons d'effets dues aux variables de régression et aux variables latentes, et, comme précédemment, il est impossible de prévoir valablement comment une intervention sur les niveaux des variables de régression affectera le système.

Dans une expérience planifiée, la situation est entièrement différente. C'est, on le sait, pour surmonter les difficultés qui ont été décrites ci-dessus que Fisher a introduit l'idée de plan d'expérience, et en particulier de randomisation. Lorsque les niveaux des variables de régression sont fixés par un procédé délibérativement aléatoire, il est impossible que les niveaux d'une variable de régression soient affectés par le niveau d'une variable latente. La seule raison pour laquelle les variables de régression ont telle ou telle valeur particulière à l'intérieur de l'organisation du plan d'expérience, est le résultat du jet d'un dé non biaisé, ou de toute autre opération aléatoire. Fisher a rendu possible l'analyse des données <u>comme si</u> les hypothèses gaussiennes étaient vraies, en transformant $\underline{X_1}$ en une variable aléatoire. Les variables de régression peuvent, naturellement, affecter encore les variables latentes, et celles-ci, en retour, affecter y. A condition, cependant, que les résultats de l'expérience soient appliqués au même système que celui à partir duquel les données ont été obtenues, cela ne cause pas de problème. Il sera absolument vrai que, sous réserve de l'erreur expérimentale, la <u>modification</u> imposée aux variables de régression produira les changements prévus sur y, même si cela s'effectue par le canal de quelque variable latente.

La difficulté fondamentale mentionnée ci-dessus est loin d'être la seule que l'on rencontre dans l'analyse des observations non planifiées. Dans un processus industriel, l'expérience passée montre souvent que certaines variables ont une importance majeure. Par suite, dans le but de contrôler les fluctuations du processus, on prend soin de maintenir ces variables très près de valeurs fixées. Comme la "signification statistique" d'une variable est grandement affectée par l'étendue des valeurs qu'elle recouvre, il y a une forte probabilité que les variables les plus importantes soient baptisées "non significatives" lors d'une analyse de régression standard. Une autre difficulté qui se présente avec des observations non planifiées, est que les variables de régression seront souvent hautement corrélées du seul fait de la règle opératoire adoptée. L'opérateur a pour mission de réduire x_2 chaque fois que x_1 devient élevé. Dans une telle circonstance, même si les difficultés provenant de variables latentes sont absentes, il peut être presque impossible de savoir si un changement d'un y est associé à x_1, à x_2, ou au deux. Naturellement, dans les expériences planifiées on s'arrange normalement pour que x_1 et x_2 soient sans corrélation en utilisant un plan orthogonal.

En résumé, l'analyse de régression de données non planifiées est une technique qui doit être utilisée avec beaucoup de précaution. Cependant

(i) elle peut fournir une prévision utile de y, dans un système déterminé observé passivement, même s'il existe des variables latentes de quelque importance. Pour cette application, des programmes de calcul, avec addition ou suppression progressive de variables peuvent avoir une signification ;

(ii) elle fait partie des nombreux outils qui peuvent être utilisés pour indiquer les variables qui devraient être prises en considération dans un plan d'expérience ultérieur (où la randomisation constituera, naturellement, une partie essentielle du plan). Elle ne devrait jamais être utilisée pour décider des variables qui devraient être exclues des recherches futures, pour les raisons qui sont évidentes d'après ce qui précède.

Pour savoir ce qui se produira lorsqu'on intervient dans un système, il faut intervenir dans le système (et ne pas se borner à l'observer passivement).

REFERENCES

[1] FISHER, R.A. - Design of Experiments, (1937) Oliver and Boyd.