STATISTIQUE ET ANALYSE DES DONNÉES

CHRISTIAN DUHAMEL

# Estimation d'une courbe de régression de la moyenne

*Statistique et analyse des données*, tome 3, n° 1 (1978), p. 47-52

<http://www.numdam.org/item?id=SAD_1978__3_1_47_0>

Statistique et Analyse des Données
1 - 1978

# ESTIMATION D'UNE COURBE DE REGRESSION DE LA MOYENNE.
## (Programme REGMOY)

DUHAMEL christian

Mathématique - Université Paris-Sud - 91405 ORSAY
Laboratoire de Biométrie du C.N.R.Z. - 78350 JOUY EN JOSAS

RESUME.

Un programme (Fortran IV) d'estimation non paramètrique d'une courbe de régression de la moyenne $E(Y/X = x)$, où $Y$ et $X$ sont deux variables continues est présenté.
L'estimation est faite par la méthode des noyaux de Parzen suivant les résultats de Nadaraya. Les noyaux sont des gaussiennes centrées sur les observations et les paramètres sont estimés par maximum de vraisemblance. Une comparaison est faite avec la régression linéaire. Des possibilités de généralisation à deux variables explicatives où à l'estimation de coordonnées manquantes sont indiquées.

SUMMARY.

A program (Fortran IV) of non-parametric estimation of a regression curve $E(Y/X = x)$, where $Y$ and $X$ are two continuous variables is presented.
The estimation is made by use of Parzen's Kernels as indicated by Nadaraya. The kernels are gaussian, centered on the observations and the parameters are estimated by maximum likelihood. The comparaison is made with linear regression and possibilities of extension to two explicative variables or to estimation of missing coordinates are indicated.

Soit $(x^i, y^i)_{i=1,\ldots,n}$ un n-échantillon d'observations d'une bivariate $(X, Y)$ de densité $g(x, y)$ dans $\mathbb{R}^2$.
L'étude de la dépendance de $Y$ en $X$ se fait généralement par l'estimation d'un paramètre $\theta$ de $\mathbb{R}^k$ d'une fonction $f(\theta, x)$ censée représenter cette dépendance. Ce qui représente déjà une connaissance a priori, et exclut des comportements très localisés. Ainsi $Y$ peut présenter un pic ou un brusque affaissement autour d'une valeur $x_o$ de $X$.

Le but de cette étude est de donner une estimation graphique de la courbe de régression de la moyenne :

$$y(x) = E(Y/X = x)$$

sans faire d'a priori sur le comportement de cette courbe.

CONSTRUCTION D'UN ESTIMATEUR DE $E(Y/X = x)$.

Soit $\hat{g}_n(x,y)$ un estimateur de densité $g(x, y)$, donné par la moyenne de produits de noyaux $K$, centrés sur les observations $(x^i, y^i)$ :

$$\hat{g}_n(x,y) = \frac{1}{n\alpha_n^2} \sum_{i=1}^{n} K\left(\frac{x-x^i}{\alpha_n}\right) K\left(\frac{y-y^i}{\alpha_n}\right)$$

où $\alpha_n$ est un paramètre de lissage [Parzen (1962), Cacoullos (1966)]. Un estimateur $\hat{y}_n(X)$ de $y(x) = E(Y/X = x)$ est donné par :

$$\hat{y}_n(x) = \frac{\int_R y\, \hat{g}_n(x,y)\, dy}{\int_R \hat{g}_n(x,y)\, dy} = \frac{\sum_{i=1}^{n} y^i K\left(\frac{x-x^i}{\alpha_n}\right)}{\sum_{i=1}^{n} K\left(\frac{x-x^i}{\alpha_n}\right)}$$

Nadaraya (1964) donne les conditions de convergence de cet estimateur vers $E(Y/X = x)$. Soit $f(x)$ et $h(y)$ les densités respectives de $X$ et $Y$ ; et $K(x)$ une densité telle que :

a) $K(x) < C < \infty$

b) $\lim_{x \to \pm\infty} x\, K(x) = 0$

Soit $\{\alpha_n\}$ une suite de nombres positifs tendant vers zéro

a) si $n\alpha_n^2 \to \infty$ et si $\int_R y^2 h(y)\, dy$ est finie, $\hat{y}_n(x)$ est un estimateur convergent de $y(x)$ pour tout $x$ tel que $y(x)$ et $f(x)$ sont continues et $f(x)$ positive.

b) si la fonction caractéristique de $K(x)$ est absolument intégrable et si les conditions suivantes sont satisfaites :

$y(x)$ et $f(x)$ continues sur $[a,b]$ borné : $\min_{[a,b]} f(x) > 0$ ;

$$\int y^4 h(y) < \infty \; ; \quad \sum_{n=1}^{\infty} \frac{1}{n^2 \alpha_n^4} < \infty \; ; \text{ alors :}$$

$$\Pr\left[ \sup_{[a,b]} |\hat{y}_n(x) - g(x)| \to 0 \right] = 1 \qquad n \to \infty$$

Remarquons que Nadaraya ne fait pas intervenir le fait que $K(x)$ soit centré.

Le noyau $K(x)$ choisi est une gaussienne centrée de variance la variance estimée $\sigma_x$ de $X$. C'est-à-dire :

$$\hat{g}_n(x,y) = \frac{1}{n\alpha_n^2} \cdot \frac{1}{2\pi\, \sigma_x \cdot \sigma_y} \sum_{i=1}^{n} e^{-1/2\left(\frac{x-x^i}{\sigma_x \alpha_n}\right)^2} e^{-1/2\left(\frac{y-y^i}{\sigma_y \alpha_n}\right)^2}$$

qui est un estimateur convergent de $g(x, y)$ sous les conditions ci-dessus.

CHOIX DU PARAMETRE DE LISSAGE $\alpha_n$ .

En fait, il n'est pas arbitraire sous les seules contraintes indiquées. Dans leur utilisation des estimateurs de densités de Parzen étendus au cas multivariate pour une analyse discriminante multigroupe sur des données continues, Hermans et Habbema (1976) donnent une estimation de maximum de vraisemblance de $\alpha_n$.
Soit $g^r_{n-1}(\alpha, x, y)$ la densité estimée par l'expression ci-dessus, avec la valeur $\alpha$ du paramètre de lissage, à l'aide des points de l'échantillon sauf $(x^r, y^r)$. $\alpha_n$ est choisi de façon à maximiser la vraisemblance :

$$G(\alpha) = \prod_{r=1}^{n} g^r_{n-1}(\alpha, x^r, y^r)$$

Si on prenait $\prod_{r=1}^{n} g_n(\alpha, x^r, y^r)$, le Sup serait obtenu pour $\alpha = 0$, ce qui correspondrait à une dégénérescence vers une distribution discrète sur les observations.
La réduction des variables X et Y permet de rechercher le maximum de $G(\alpha)$ en partant de $\alpha = 1$ et jusqu'à obtenir :

$$g(\alpha_n) > g(\alpha_n + 0.01) \quad \text{et} \quad g(\alpha_n) \geq g(\alpha_n - 0.01)$$

Les valeurs obtenues au cours des différentes études sur des données réelles semblent bien vérifier :

$\alpha_n$ croît et $n\,\alpha_n^2$ décroit lorsque $n$ augmente.

Le programme REGMOY calcule $\alpha_n$ par la méthode du maximum de vraisemblance, et en déduit pour des valeurs fixées de $x$ les $\hat{y}_n(x)$ donnés par :

$$\hat{y}_n(x) = \frac{\sum_{i=1}^{n} y^i \exp\left[ -\frac{1}{2} \frac{(x-x^i)^2}{\sigma_x^2 \alpha_n^2} \right]}{\sum_{i=1}^{n} \exp\left[ -\frac{1}{2} \frac{(x-x^i)^2}{\sigma_x^2 \alpha_n^2} \right]}$$

Dans le plan $(x, y)$ sont alors projetés : les points observations, la courbe de régression de la moyenne estimée et la droite de régression.
Le programme permet aussi la représentation de la courbe estimée du maximum de vraisemblance :

$$y(x) = \operatorname{Sup}_y \hat{g}_n(x,y)$$

Le rapport de corrélation $R_c$ est défini par :

$$R_c^2 = 1 - \frac{\text{var}(Y-E(Y/X))}{\text{var } Y}$$

ou, ce qui revient au même puisque $E(Y/X)$ et $Y-E(Y/X)$ sont orthogonaux dans $L^2$ :

$$R_c^2 = \frac{\text{var } E(Y/X)}{\text{var } Y}$$

Dans le programme, $R_c$ est calculé suivant :

$$\hat{R}_c^2 = 1 - \frac{\sum_{i=1}^{n} (y^i - \hat{y}_n(x^i))^2}{\sum_{i=1}^{n} (y^i - \bar{y})^2}$$

les $\hat{y}_n(x^i)$ sont obtenus par interpolation.
$R_c$ peut être comparé au coefficient $r_L$ de corrélation linéaire :

$$(R_c^2 - r_L^2) \cdot \text{var } Y = E(Y_{\text{linéaire}} - E(Y/X))^2$$

Il est à noter que, lorsque la courbe de régression $\hat{y}_n(x)$ et la droite de régression coïncident à peu près, $\hat{R}_c^2$ est parfois légèrement inférieur à $r_L^2$. Cela semble dû au fait que $\hat{y}_n(x^i)$ ne représente pas la moyenne des $y$ d'abscisse $x^i$.

GENERALISATIONS POSSIBLES.

1) Les résultats précédents peuvent aisément s'étendre au cas de deux variables explicatives $X$ et $Z$ et permettre d'estimer $E(Y/X = x, Z = z)$ par

$$\hat{y}_n(x, z) = \frac{\sum_{i=1}^{n} y^i K\left(\frac{x-x^i}{\alpha_n}\right) K\left(\frac{z-z^i}{\alpha_n}\right)}{\sum_{i=1}^{n} K\left(\frac{x-x^i}{\alpha_n}\right) K\left(\frac{z-z^i}{\alpha_n}\right)}$$

La représentation graphique se faisant à l'aide de courbes de niveaux. (Le programme REGMOY se limite actuellement à une seule variable explicative).

2) Dans le cas de plusieurs variables explicatives, on obtient une possibilité d'estimation de coordonnées manquantes de vecteurs de données quantitatives. Soit sous forme d'espérances conditionnelles :

$$\hat{E}(X_1/X_2 = x_2, \ldots, X_p = x_p)$$

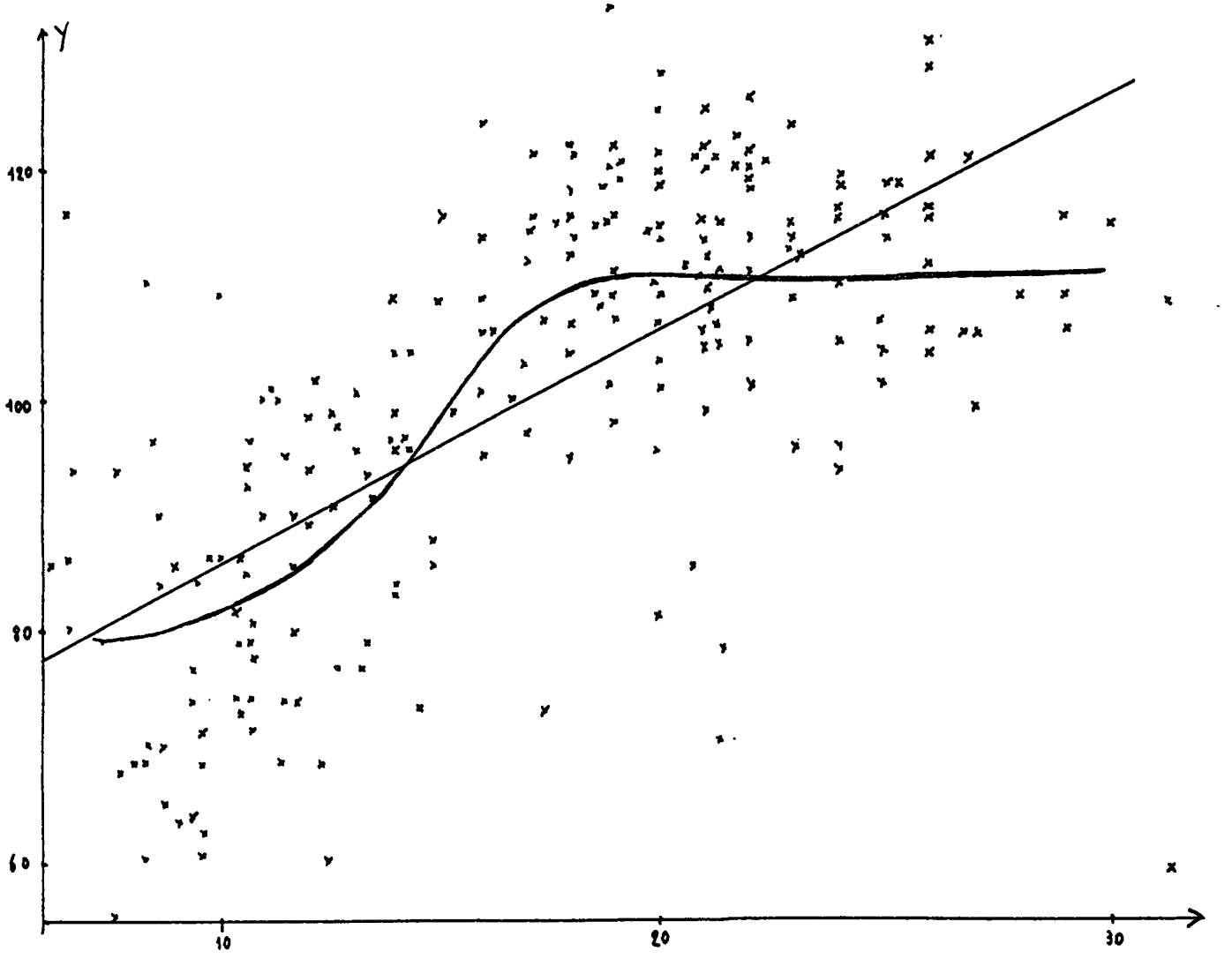
Soit par maximum de vraisemblance

$$\operatorname{Sup}_x \hat{g}(x, x_2, \ldots, x_p)$$

Le programme et sa notice sont disponibles au laboratoire de Biométrie du C.N.R.Z. 78350 Jouy-en-Josas.

APPLICATION.

Le programme REGMOY a été utilisé pour étudier la dépendance de différentes variables mesurées sur des animaux. Il apparaissait clairement que la seule connaissance du coefficient de corrélation linéaire était insuffisante.
On voit apparaître des dépendances différentes suivant l'abscisse de la variable explicative $X$, ou bien une coïncidence entre la droite de régression et la courbe de régression estimée.
La courbe $\hat{E}(Y/X)$ représentée ci-dessous est déterminée à partir de 220 observations.




Etude de la dépendance de la concentration de calcium Y en fonction de la concentration de magnésium dans le plasma.

La droite de régression est également tracée. On étudie la concentration de calcium Y dans le plasma de porcs en fonction du taux de magnésium X. La courbe de régression estimée fait apparaître, sur cette expérience, que Y croît avec X, jusqu'à une certaine valeur de X, puis, au delà de cette valeur, reste à peu près constante.

paramètre de lissage $\alpha_n = 0,32$

coefficient de corrélation linéaire $r_L = 0,683$

rapport de corrélation $R_c = 0,760$

BIBLIOGRAPHIE.

BROWN et al. (1975) : Techniques for testing the constancy of regression relationships over time. J.R.S.S. 37 B, 149-163.

BREIMAN L. et al. (1977) : Variable Kernel estimates of multivariate densities. Technometrics 19(2), 135-144

CACOULLOS Th. (1966) : Estimation a multivariate density. Ann. Inst. Stat. Math. 18(2) 179-189.

DUHAMEL C. (1977) : Les programmes ALLOC d'analyse discriminante (d'après Hermans J. et Habbema J.D.F.) Laboratoire de Biométrie du C.N.R.Z. note 77/04.

HERMANS J. et HABBEMA J.D.F. (1976) : The Alloc package multigroup discriminant analysis programs based on direct density estimation in Compstat 1976, Proc. in Comp. Stat. J. Gordesch and P. Naeve (des) Physica Verlag, Wien.

NADARAYA E.A. (1964) : On estimating regression $Th^y$ of Prob. and Appl. 141-2

NADARAYA E.A. (1965) : On parametric estimates of density function and regression curves. $Th^y$ of Prob. and Appl. 186-196.

PARZEN E. (1962) : On estimation of a probability density function and mode. Ann. Math. Statis. 33(3), 1065-1076.

ROSENBLATT M. (1969) : Conditional Probability Density and Regression Estimator. Krishnaiah R. (ed) : in Multivariate Analysis II Academic Press, New-York, Lon don, 25-31.

SCHMERLING S. and PEIL J. (1977) : Zur Schätzung der Regression aus Beobachtungswerten stetiger Zufallsgrössen. Biometrical Journal 19(4), 291-301.