

Leçon 7 : Régression non linéaire et ajustement par quantiles

Régression non linéaire par MCO La méthode des MCO qui a été utilisée pour calculer l'équation de la droite des moindres carrés peut s'étendre théoriquement à la recherche d'autres courbes, non nécessairement linéaires qui s'ajusteraient mieux au nuage de points considéré. Ainsi pour ajuster une courbe exponentielle $y(x) = Ae^{Bx}$ (si le nuage a une allure exponentielle), il suffit de calculer les nombre A et B qui rendent minimale la somme E des carrés des écarts

$$E = \sum_{i=0}^n \varepsilon_i^2 = \sum_{i=0}^n (y_i - (Ae^{Bx_i}))^2$$

et cela peut se faire de façon tout-à-fait générale pour toute fonction $y = f(x)$. Cette méthode présente cependant des défauts graves. Le premier, qui n'est pas le plus sérieux, est que l'on ne dispose plus, en général, de formules permettant de calculer les coefficients inconnus qui minimise E et qu'on a donc recours à des calculs approchés dont la qualité n'est pas toujours garantie. Mais c'est le choix de la fonction f qui pose le plus sérieux problème. En effet, chercher parmi toutes les fonctions possibles et imaginables une fonction qui réaliserait le minimum de la somme des carrés des écarts est une question sans intérêt car on sait qu'il existe au moins une fonction qui rende $E = 0$, c'est-à-dire une fonction dont le graphe passe par tous les points du nuage : si le nuage a n points, il existe un polynôme de degré n pour lequel $E = 0$ qui s'appelle le polynôme d'interpolation de Lagrange. Mais ce polynôme n'ayant rien à voir avec les données formant le nuage, est ici absolument sans intérêt. La difficulté n'est donc pas d'ajuster une fonction plus ou moins arbitraire qui minimise E mais plutôt de trouver une fonction f qui soit un *modèle pertinent des données* et qui fournisse en quelque sorte la loi du phénomène étudié, que l'on serait ainsi parvenu à séparer du *bruit* inclu dans les données. Il est alors évident que le choix de f n'est pas une question purement mathématique.

Régression non linéaire par linéarisation de la dépendance Une autre façon de choisir un modèle de dépendance $y=f(x)$ non linéaire est d'effectuer une transformation des variables (x, y) en $(\tilde{x}, \tilde{y})$ qui rende linéaire la dépendance entre les nouvelles variables $\tilde{x}$ et $\tilde{y}$. On peut alors déterminer cette dépendance par simple régression linéaire.

Le tableau suivant (emprunté à l'ouvrage "Computer simulation in Biology : a basic introduction", de R.E. Kean et J.D. Spain) indique quelques une des transformations parmi les plus utilisées en biologie. La figure ci-dessous indique l'allure des courbes concernées qui fournissent une sorte de menu auquel on peut comparer le nuage de données étudié afin de choisir la transformation la plus raisonnable. Quoiqu'il en soit, on n'oubliera pas qu'un calcul du R^2 (ou du coefficient de corrélation linéaire) accompagné d'un examen soigneux des résidus ε_i est indispensable avant toute validation du modèle.

Type de courbes	Fonction	Transformation	Forme affine
droite	$y = A + Bx$	-	$y = A + Bx$
hyperbole	$y = \frac{Ax}{B+x}$	$\tilde{y} = \frac{x}{y}$	$\tilde{y} = \frac{B}{A} + \frac{1}{A}x$
inverse modifié	$y = \frac{A}{B+x}$	$\tilde{y} = \frac{1}{x}$	$\tilde{y} = \frac{B}{A} + \frac{1}{A}x$
exponentiel	$y = Ae^{Bx}$	$\tilde{y} = \ln(y)$	$\tilde{y} = \ln(A) + Bx$
exponentiel réciproque	$y = Ae^{\frac{B}{x}}$	$\tilde{y} = \ln(y) \quad \tilde{x} = \frac{1}{x}$	$\tilde{y} = \ln(A) + B\tilde{x}$
maxima	$y = Axe^{Bx}$	$\tilde{y} = \ln(\frac{y}{x})$	$\tilde{y} = \ln(A) + Bx$
exponentiel saturé	$y = A(1 - e^{-Bx})$	$\tilde{y} = \ln(A - y)$	$\tilde{y} = \ln(A) + Bx$
logistique	$y = \frac{K}{1 + Ae^{-Bx}}$	$\tilde{y} = \ln(\frac{K}{y-1})$	$\tilde{y} = \ln(A) + Bx$
logarithmique	$y = K + Ax^B$	$\tilde{y} = \ln(y) \quad \tilde{x} = \ln(x)$	$\tilde{y} = \ln(A) + B\tilde{x}$
Sigmoïde	$y = \frac{K}{1 + Ae^{-Bx}}$	$\tilde{y} = \ln(\frac{K}{y-1}) \quad \tilde{x} = \ln(x)$	$\tilde{y} = \ln(A) + B\tilde{x}$

Variable aléatoire distribuée selon une loi normale (ou gaussienne) : On appelle courbe de Gauss (ou *cloche de Gauss*) le graphe de la fonction $g(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$. Cette fonction est paire (graphe symétrique par rapport à l'axe des y) et on peut montrer que l'on a $\int_{-\infty}^{+\infty} g(x)dx = 1$. On dit qu'une v.a. d'espérance 0 et de variance 1 a une *distribution normale* (ou *gaussienne*), ou encore qu'elle est *distribuée selon une loi normale*, si pour tout x , on a :

$$P(X \leq x) = \int_{-\infty}^x g(u)du.$$

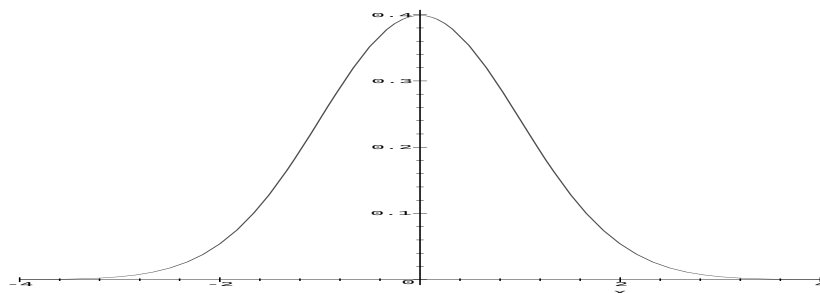


FIG. 1 – La cloche de Gauss : densité d’une loi normale

FIG. 2 – Papier gaussio-arithmétique

En pratique, les v.a. que l’on étudie prenant en général un nombre fini de valeurs, elles ne peuvent vérifier cette équation exactement. Cette propriété se traduit alors par le fait que leur histogramme a une allure de cloche de Gauss. Nous allons voir ci-dessous comment déterminer, pour un histogramme donné si son approximation par une cloche de Gauss est acceptable ou non.

On appelle *quartiles* d’une v.a. X distribuée selon une loi normale les trois nombres q_1 , q_2 et q_3 vérifiant :

$$P(X \leq q_1) = 0,25 \quad , \quad P(X \leq q_2) = 0,50 \quad , \quad P(X \leq q_3) = 0,75$$

le quartile q_2 étant aussi appelé *médiane*. Ce sont les 3 valeurs de x qui permettent de découper la cloche de gauss en 4 parties de même surface. On définit de la même façon les *déciles* qui sont les 9 valeurs de x qui permettent de découper la cloche de gauss en 10 parties de même surface, et on définirait de même les 99 *centiles*. L’ensemble de ces nombres sont désignés sous le nom de *quantiles* de la v.a. X .

Ajustement par quantiles : l’exemple de la droite de Henri

Exercices : 1.

2.