Régressions linéaires. (Document provisoire)

Le terme régression désigne l'opération permettant de trouver une fonction qui satisfasse au mieux un ensemble de valeurs. Typiquement, on dispose d'un certain nombre de points, qu'on appelle quelque fois "nuage de points", le but étant de trouver une fonction unique acceptable pour touts ces points. On utilise la méthode des moindres carrés, la justification de son emploi dépasse le cadre de ce papier.
Cette régression est linéaire parce qu'elle se calcule par la résolution d'un système linéaire de N équations à N inconnues.

Approche élémentaire.

On dispose d'un certain nombre, supérieur à deux, de couples Xi, Yi et l'on conjecture qu'il existe une fonction $y = a + bx$ qui permettrait de satisfaire correctement les relations.
Les paramètres a et b étant fixés, pour tout X on pourra calculer $Y = a + bX$.

Les erreurs commises sur les valeurs Yi sont $ei=Yi - (a + bXi)$.
On sait que le résultat le plus probable est celui qui minimise la somme des carrés des erreurs.
$S = \Sigma(ei^2) = \Sigma [(Yi - (a + bXi))^2]$
$S = \Sigma [Yi^2 - 2Yi(a + bXi) + (a + bXi)^2]$
$S = \Sigma [Yi^2 - 2aYi - 2b XiYi + a^2 + 2abXi + b^2Xi^2]$ pour i de 1 à n.
S sera minimale pour les valeurs qui annulent sa dérivée.
$Sa' = \Sigma[- 2Yi + 2a + 2bXi]$
$Sb' = \Sigma[- 2 XiYi + 2aXi + 2bXi^2]$

D'où le système linéaire en A et B
$An + 2B\Sigma Yi = \Sigma Yi$
$A\Sigma Xi + B\Sigma Xi^2 = \Sigma XiYi$

Le coefficient de détermination s'écrit
$R^2 = [A\Sigma(Yi) + B\Sigma(XiYi) - 1/n (\Sigma Yi)^2] / [\Sigma(Yi^2) - 1/n (\Sigma Yi)^2]$

Généralisation de la fonction.

Il y deux types de généralisation, par changement de variable et par multi-paramétrage.

Changement de variable.

Cette méthode est très utilisée puisqu'elle permet d'obtenir quatre nouvelles fonctions à moindre coût.

1- Exponentielle : on fait le changement de variable yi -> ln(yi)
2- Logarithmique : on fait le changement de variable xi -> ln(xi)
3- Puissance : on fait le changement de variable xi -> ln(xi) et yi -> ln(yi)
4- Composite : on fait le changement de variable xi -> ln(xi)/xi

Cette liste n'est pas limitative.

Multi-paramétrage.

La fonction de la forme y = A + Bx comporte deux paramètres, A et B, quels que soient les changements de variable éventuels.
Il peut être nécessaire de trouver une fonction polynomiale de degré plus grand. Alors, il suffit d'écrire la fonction avec les paramètres supplémentaires. Par exemple la fonction de degré 4 aura 5 paramètres, A, B, C, D, E. Les calcul est le même et le système à résoudre aura 5 équations à 5 inconnues.

Autres généralisations

Généralisation en 3D. De nombreux phénomènes se modélisent par des fonctions à trois variables. Elles s'écrivent sous la forme Z = f(X, Y). Ce type de phénomène se représente souvent à l'aide d'abaque, c'est à dire la lecture se fait graphiquement. Il y a plusieurs types d'abaque, les plus connus sont des courbes correspondant à certaines valeurs rondes de l'une des variables et les deux autres variables sont cotées sur l'axe des abscisses et sur l'axe des ordonnées.
Certaines de ces abaques ont été réalisées à partir d'observations et graphiquement. Il peut être intéressant de trouver la fonction numérique qui donne le résultat.
Plus généralement, si on dispose de triplets XYZ et non plus de couples, on procédera de la même façon et la fonction aura trois paramètres A, B, C.
Etant donné les changements de base avec le logarithme, il y a huit fonctions de base possibles. On choisira naturellement la meilleure.

Si le nombre de variables est supérieur à trois, sauf cas particulier, il peut devenir très compliqué d'envisager tous les cas possibles.
La formule à préférer est du type
$Y = K . X_1^a . X_2^b . X_3^c \ldots X_n^n$
La méthode de calcul est toujours la même.