Leçon 4 : La régression

NUAMA Ekou

24 / 03 / 2017
1.0

Table des matières

I - Objectifs **3**

II - Introduction **4**

III - Notion de courbe de régression **5**

IV - Ajustement linéaire d'une fonction exponentielle **7**

V - Ajustement linéaire par une fonction puissance **8**

VI - Corrélation simple **9**

Objectifs

faire des prévisions à partir d'un modèle simple

Introduction

Il s'agira d'élaborer une relation entre deux variable décrites unité par unité ; et d'élaborer la corrélation qui existe entre les deux variables ; par exemple le revenu des ménages et les dépenses de consommation de ceux-ci. Quand le revenu augmente les dépenses de consommation augmentent (corrélation positive) ; autre corrélation (prix et quantité) quand le prix augmente, la quantité diminue (corrélation négative).

Notion de courbe de régression

L'ajustement linéaire consiste à minimiser la somme des carrés des écarts. Ce principe est connu sous le nom « Méthode des moindres carrés ordinaires (mco) ».

$$d_i = Y_i - (a\,x_i + b)$$

$$\sum d_i^2 = \sum \left[Y_i - (a x_i + b) \right]^2$$

$$P(a, b) = \sum d_i^2 = \sum \left[Y_i - (a x_i + b) \right]^2$$

P(a, b) admet un minimum.

Condition nécessaire de premier ordre

$$\frac{\partial^2 P}{\partial a^2} > 0 \qquad \text{et} \qquad \frac{\partial^2 P}{\partial b^2} > 0$$

Condition nécessaire de second ordre

$$\frac{\partial^2 P}{\partial a^2} > 0 \qquad \text{et} \qquad \frac{\partial^2 P}{\partial b^2} > 0$$

Le principe

Il s'agit de déterminer deux droites d'équation $Y = ax + b$ et $X = a' y + b'$ telles que pour chacune d'elles, les distances prises entre chaque point du nuage et la droite soient les plus petites possibles. Pour la droite de régression Dy/x les écarts (d_i) sont parallèles à OY et pour la droite de régression Dx/y les écarts d_i, sont parallèles à OX

Les deux droites passent par le point moyen du nuage $(\overline{X} ; \overline{Y})$

La droite Dy/x a pour équation : $Y = ax + b$ a est la pente de la droite Dy/x

La droite Dx/y a pour équation $X = a\ ' y + b$ est la pente de la droite Dx/y et b' l'ordonnée à l'origine.

$$a = \frac{Cov(x,y)}{V(x)} \qquad a' = \frac{Cov(x,y)}{V(y)}$$

Ajustement linéaire d'une fonction exponentielle

II

Le nuage de points peu t avoir l'allure d'une fonction exponentielle. Dans ce cas, les n points Mi de coordonnées (Xi, Yi) avec i =1.....n, peuvent avoir la relation :

$Y = exp\ (aX + b)$, relation en base e

$Y = 10^{(aX+b)}$, relation en base 10.

On peut rendre linéaire la fonction en lui appliquant la fonction logarithmique c'est-à-dire en écrivant :

- en base e :

$$Ln(Y) = Ln\left[\exp(aX+b)\right], \text{ soit Ln (Y)= a X +b}$$

- en base 10 :

$$\log(Y) = \log\left[10^{aX+b}\right]$$

Posons les changements de variables suivants :

$$\text{Ln (Y)} = Z \text{ ou } \log(Y) = Z$$

On a alors : Z= aX +b. Ainsi, le nouveau nuage de points Mi de coordonnées (X_i, Ln Y_i) avec i = 1....., n est linéaire grâce à la transformation semi-logarithmique.

Ajustement linéaire par une fonction puissance

III

Nous pouvons penser que la liaison entre les variables statistiques X et Y est de la forme :

$Y = bX^a$ On peut rendre linéaire la fonction en lui appliquant la fonction logarithmique (Ln représente le logarithme népérien), mais, il est possible d'utiliser aussi le logarithme décimal. On écrit :

$$Ln(Y) = Ln(bX^a)$$

$$Ln(Y) = Ln(b) + aLn(X)$$

Posons les changements de variables $Ln(Y) = Z$, $Ln(b) = \beta$, $Ln(X) = W$

On a alors : $Z = \beta + aW$

Ainsi, un nouveau nuage de n points Mi, de coordonnées (Ln X_i, Ln Y_i) avec i= 1..... n, est linéaire par la transformation logarithmique en base e ou en base 10 si l'on utilise le logarithme décimal.

Corrélation simple

IV

Le coefficient de détermination linéaire (R^2)

Il mesure l'intensité de la liaison linéaire entre X et Y.

$$R^2 = \frac{Cov(x,y)^2}{V(x)V(y)} \qquad a = \frac{Cov(x,y)}{V(x)}$$

$$a' = \frac{Cov(x,y)}{V(y)}$$

$$R^2 = a\, a'$$

Le rapport de corrélation

$$(Y_i - \bar{Y}) = (Y - Y_i) + (Y_i - \bar{Y})$$

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(Y_i - Y_i)^2 + \sum_{i=1}^{n}(Y - \bar{Y})^2 + 2\sum_{i=1}^{n}(Y - Y)(Y_i - \bar{Y})$$

$2\sum_{i=1}^{n}(Y_i - Y_i)(Y_i - \bar{Y}) = 0$, car la somme des écarts par rapport à la moyenne est toujours nulle.

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(Y_i - Y)^2 + \sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

$$\frac{1}{N}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \frac{1}{N}\sum_{i=1}^{n}(Y - Y)^2 + \frac{1}{N}\sum_{i=1}^{n}(Y - \bar{Y})^2$$

On a : $V(Y) = V(e) + V(Y)$

Variance totale de Y : $Y : \frac{1}{N}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$

Variance de l'erreur ou variance résiduelle : $\frac{1}{N}\sum_{i=1}^{n}(Y_i - Y_i)^2$

Variance expliquée par la régression : $\frac{1}{N}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$

$$\eta^2_{Y/X} = 1 - \frac{V(e)}{V(Y)}$$

$$\eta_{X/Y}{}^2 = 1 - \frac{V(e)}{V(X)}$$