Leçon 4 : Notion de régression (l'ajustement linéaire) : la droite des moindres carrés ordinaires

Prof NUAMA Ekou

12-03-19
2.0

Table des matières

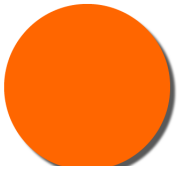
Introduction 3

I - I- Le principe 4

II - La qualité de la régression 5

III - Liaison entre les variables qualitatives ordinales : la corrélation des rangs 7

Introduction



L'objectif de ce chapitre est de déterminer les paramètres d'une droite. Les données statistiques sont décrites unité par unité, on n'est pas dans le cas d'un tableau de contingence, mais un tableau à deux variables où celles-ci sont indicées par i. Il existe plusieurs méthodes d'estimation de la droite de régression. Ces méthodes sont: la méthode de Mayer, la méthode du maximum de vraisemblance et la méthode des moindres carrés ordinaires.

La méthode de Mayer

Elle consiste à diviser le nuage de points en deux parties égales. Dans chaque sous partie, il faut déterminer un point moyen, on a deux points moyens du nuage. De ces deux points moyens, on fait passer une droite et une seule. Cette droite est appelée: la droite de Mayer.

La méthode du maximum de vraisemblance

Il faut écrire la fonction de vraisemblance et chercher les conditions de la maximisation de cette fonction de vraisemblance.

La méthode des moindres carrés ordinaires

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

$e_i = Y_i - (a\,Xi + b)$

e_i représente l'écart

$$\sum e_i^2 = \sum \left[Y_i - \left(aX_{i+b} \right) \right]^2$$

$$P(a, b) = \sum e^2 = \sum \left[Y_i - \left(aX_{i+b} \right) \right]^2$$

P(a, b) admet un minimum.

Condition nécessaire de premier ordre

$$\frac{\partial P}{\partial a} = 0 \qquad \text{et} \qquad \frac{\partial P}{\partial b} = 0$$

Condition nécessaire de second ordre

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

I- 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 (e_i) sont parallèles à OY et pour la droite de régression Dx/y les écarts 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*avec *a* est la pente de la droite DY/X et b est l'ordonnée à l'origine.

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

$$a = \frac{Cov(X,Y)}{V(X)} \quad a' = \frac{Cov(X,Y)}{V(Y)}$$

La qualité de la régression

II

Le coefficient de détermination linéaire (R^2)

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

$$R^2=\frac{Cov(X,Y)^2}{V(X)V(Y)} \quad a=\frac{Cov(X,Y)}{V(X)} \quad a'=\frac{Cov(X,Y)}{V(Y)}$$

$R^2= a\ a'$

$$\varphi=\frac{COV(X,Y)}{\sigma_X\sigma_Y}$$

Il est le coefficient de corrélation linéaire. Il permet de mesurer la dépendance statistique.

On a également la relation suivante : $COV(X, Y)^2 < V(X)\ V(Y)$.

Ainsi, on a la relation :

$$-1\le\varphi\le+1$$

$$0\le R^2\le1$$

Lorsque l'ajustement n'est pas linéaire, on utilise le rapport de corrélation

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

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

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

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

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

Liaison entre les variables qualitatives ordinales : la corrélation des rangs

III

Le coefficient de corrélation des rangs de Spearman

Le coefficient de corrélation linéaire doit être utilisé pour mesurer la dépendance statistique de deux variables quantitatives, mais, il ne peut pas être utilisé si les variables sont qualitatives. Lorsque les variables sont qualitatives ordinales, la caractéristique de rang de Spearman peut être utilisé pour comparer le classement des séries statistiques et mesurer la similitude plus ou moins grande de ces classements. Sur n individus sont observées deux variables qualitatives ordinales et à chaque individu i, on associe le couple (Ri, Si) où Ri est son rang de classement selon la première variable et Si son rang de classement selon la deuxième variable. Soit Di = Ri- Si Où Di est la différence des rangs des deux variables. Le coefficient de corrélation des rangs de Spearman est :

$$r_s = 1 - \frac{6\sum_{i=1}^{n} di^2}{n(n^2-1)}$$

Où r_S est le coefficient de rang de Spearman, di est la différence de classement, n est la taille de l'échantillon ou de la population.

Le coefficient r_S est compris entre -1 et 1 ; S'il est égal à 1, les deux classements sont identiques. S'il est proche de 1, leur similitude est autant plus grande. En revanche, s'il est égal à -1, les deux classements sont opposés. La dissemblance entre les deux classements est d'autant plus grande que r_S est proche de -1. Si r_S =0, les deux classements sont indépendants.