Chapitre 4

Ajustement linéaire

Plan

1- Introduction

On considère une population de taille (effectif total) n sur laquelle on définit une statistique double X et Y. (On éffectue sur Ω deux observations portant sur 2 caractères différents).
Le problème qui se pose est celui qui consiste à rechercher s'il existe une relation entre X et Y.
A chaque élement ω_i de l'échantillon, on associe un couple de valeurs (x_i, y_i) qu'on représente graphiquement par un point $M_i(x_i, y_i)$ du plan. Et on obtient ainsi un nuage de n points qui constitue ce qu'on appelle un diagramme de disperssion.
Ajuster un ensemble de points consiste à déterminer une courbe (C) simple aussi proche que possible des points M_i.
L'ajustement linéaire est le cas où (C) est une droite.

2- Définition

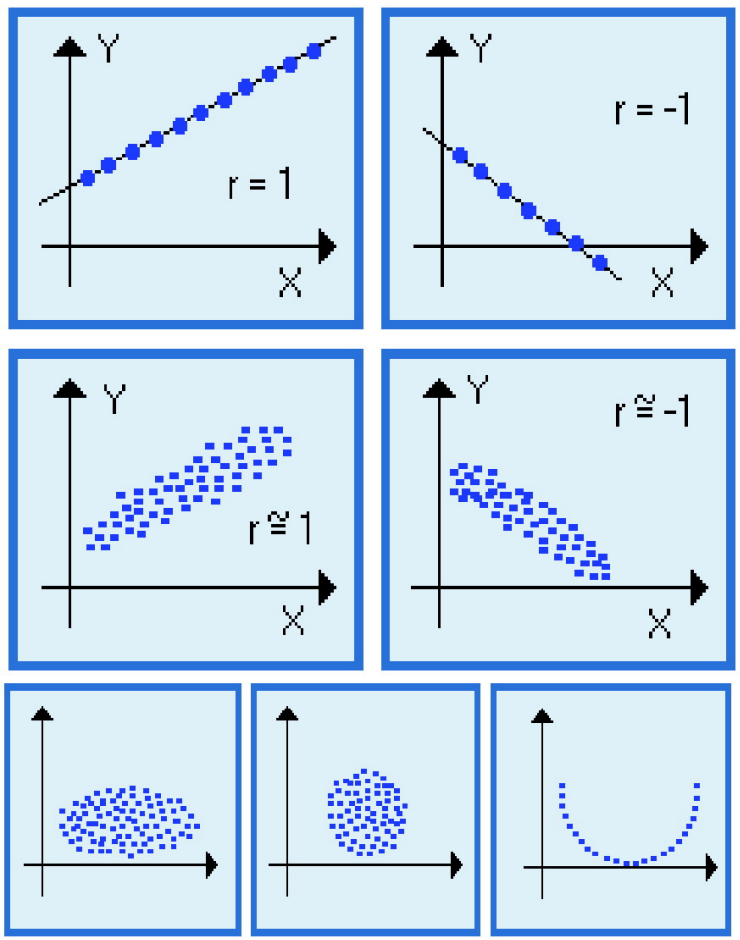
On dit qu'il y a correlation entre deux caractères observés sur une même population lorsque les variations des deux caractères se produisent dans le même sens ou lorsque les variations sont de sens contraires.

Nuage de points : Diagramme de disperssion

L'existence d'une correlation peut-être décelée (détectée) graphiquement. La forme du nuage de points formé par les points $M_i(x_i, y_i)$ nous permettent de constater si les caractères X et Y sont en correlation ou non.

Définition. On dit qu'une correlation (lorsqu'elle existe) qui lié 2 caractères X et Y est positive ou directe si Y croît en même temps que X. Si Y décroît lorque X croît, la correlation est dite inverse ou négative.

Exemples de différents nuages



Coefficient de correlation

Rappel. Le coefficient de correlation du couple (X, Y) la quantité

$$\rho = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n} (y_i - \overline{y})^2}} = \frac{cov(X, Y)}{\sqrt{V(X)}\sqrt{V(Y)}}$$

(n représente le nombre de couple d'observations (x_i, y_i))
ρ est compris entre -1 et 1.

* Si les caractères X et Y sont indépendants alors $\rho = 0$. Cependant la réciproque n'est pas necessairement vraie. Si $\rho = 0$, on dit qu'il y a correlation nulle entre X et Y ; la liaison entre X et Y peut être de forme outre que linéaire.

* Si $0 < \rho < 1$, la correlation est positive (X et Y varient dans le même sens) La valeur $\rho = +1$ indique une relation linéaire parfaite $Y = aX + b$ avec $a > 0$. C'est un cas extrème très peu rencontré en pratique.

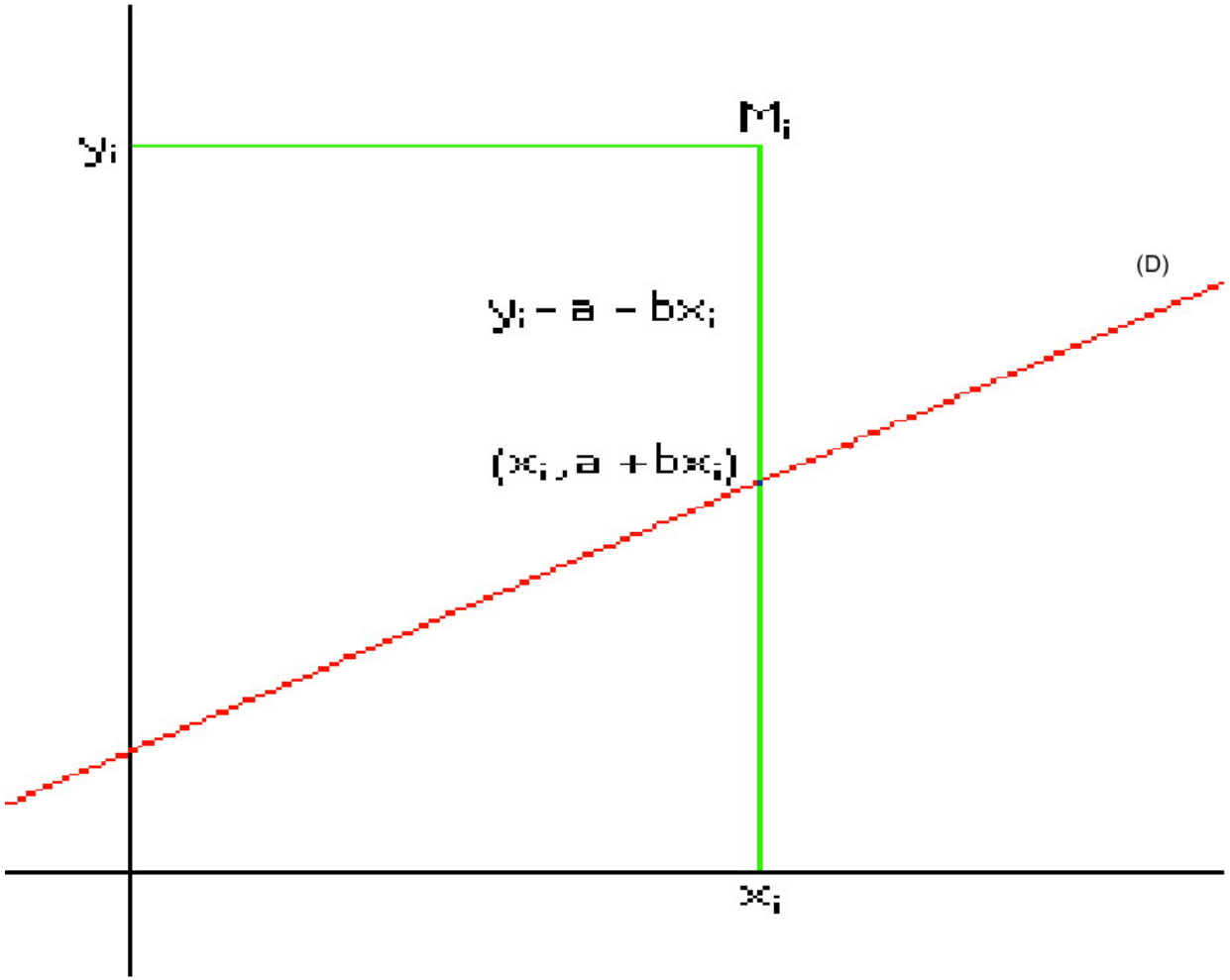
* Si $-1 < \rho < 0$, la correlation est négative (X et Y varient dans le sens contraire) La valeur $\rho = -1$ indique une relation linéaire parfaite $Y = aX + b$ avec $a < 0$. C'est un cas extrème très peu rencontré en pratique.

Formule pratique de ρ pour le calcul

$$\rho = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}}{\sqrt{\left[\left(\sum_{i=1}^{n} x_i\right)^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right]\left[\left(\sum_{i=1}^{n} y_i\right)^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right]}}.$$

3-Méthode des moindres carrés:

Soit M_i un point de cordonnées (x_i, y_i) On appelle distance de M_i parallélernent à l'axe (oy) à la droite $(\triangle)$ d'équation $y = ax + b$, le réel positif $d_i = |y_i - ax_i - b|$ (Attention ! il ne s'agit pas des distances des points M_i à la droite $(\triangle)$.)
La méthode des moindres carrés consiste à chercher les valeures de a et b (c.à.d trouver une droite $(\triangle)$ d'équation $y = ax + b$)qui mininisent $\delta(a, b) = \sum_i (y_i - ax_i - b)^2 = \sum_i d_i^2$.


M_i
y_i
(D)
$y_i - a - bx_i$
$(x_i, a + bx_i)$
x_i

Le problème admet une solution unique solution du système linéaire issue de l'annulation des dérivées partielles premières de la fonction $\delta(a, b)$. Il s'agit de résoudre

$$\begin{cases} \dfrac{\partial \delta}{\partial a} = -2\displaystyle\sum_{i=1}^{n} x_i \left(y_i - a x_i - b\right) = 0 \\ \dfrac{\partial \delta}{\partial b} = -2\displaystyle\sum_{i=1}^{n} \left(y_i - a x_i - b\right) = 0 \end{cases}$$

Ce système en a et b admet pour solution unique

$$a = \frac{cov(X, Y)}{V(X)} \text{ et } b = \overline{y} - a\overline{x}.$$

Théorème

Soit $M_i\,(x_i, y_i)_{1\leq i\leq n}$ *un ensemble fini de points fixes du plan euclidien où* x_i *sont les modalités d'un caractère* X *et* y_i *celles d'un autre caractère* Y *définis sur une même population* Ω.

Soient $\overline{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$, $\overline{y} = \dfrac{1}{n}\sum_{i=1}^{n} y_i$, $V(X) = \dfrac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^2$ *et*

$$cov(X, Y) = \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})\,(y_i - \overline{y})\,.$$

La droite d'équation $y - \overline{y} = a\,(x - \overline{x})$ *où* $a = \dfrac{cov(X, Y)}{V(X)}$ *est la droite de régression de* y *en* x *et est notée* $D_{y/x}$.

Remarque

i) On définit de même la droite de régression de x en y notée $D_{x/y}$ *et d'équation*

$$x - \overline{x} = a'(y - \overline{y}) \text{ avec } a' = \frac{cov(X, Y)}{V(Y)}.$$

ii) Les droites $D_{y/x}$ *et* $D_{x/y}$ *passent par le point* $G(\overline{x}, \overline{y})$.

Ajustement et corrélation

Les droites de regression $(\triangle)$ et $(\triangle')$ ayant pour équations: $y = ax + b$, $x = a'y + b'$ ont les propriétés suivantes:

- elles passent toutes les deux par le point $G(\overline{X}, \overline{Y})$ appelé point moyen de la statistique.
- les pentes des deux droites sont de même signe celui de la covariance et sont respectivement a et $\frac{1}{a'}$
- $aa' = \frac{cov(x,y)^2}{v(x)v(y)} = (\rho(x,y))^2$

- $(\triangle)$ et $(\triangle')$ sont confondues si elles ont la même pente (car elles passent toutes les deux par $G(\overline{X}, \overline{Y})$).
- dans ce cas $a = \frac{1}{a'}$ c.à.d. $aa' = 1$ donc $\rho(x, y) = \pm 1$. les points M_{ij} sont alors alignés.
- La corrélation linéaire est d'autant bonne (ou forte) que le coefficient de corrélation ρ est proche en valeur absolue de 1 ($\rho \simeq 1 \Longleftrightarrow a \simeq \frac{1}{a'}$).
- si ρ est proche de zéro, on dit qu'il y a corrélation linéaire très mauvaise entre X et Y. il faudrait alors approcher le nuage des points M_{ij} par une courbe.

Remarque

i) La corrélation est dite forte lorsque $R^2 > 0,75$, et dans ce cas on estime qu'on peut approcher le nuage de points par une droite (la méthode des moindres carrés s'applique).
ii) Dans le cas contraire, on ne peut pas approcher le nuage de points par une droite (l'approximation serait trop mauvaise) mais il se peut que d'autre courbe permette un bon ajustement (ajustement exponentiel par exemple)
En général, si le coefficient de corrélation est fort, on peut conclure à une corrélation entre les deux séries statistiques, mais ce n'est pas toujours vari

Exemple

La statistique suivante indique l'évolution de la consommation d'énergie éléctrique dans un pays exprimée en TWh

Année	*Consommation*
1949	30
1953	41
1957	56
1961	73
1965	97
1969	123
1973	165
1977	207

La relation qui lie la consommation au temps (année) est de type exponentiel.
Déterminons la droite de régression de $Y = \log y$ *en* x

Exemple (suite)

On effectue le changement de variable $X = \dfrac{x - 1961}{4}$, *on a*

X_k	$Y_k = \log y_k$
-3	*1.417*
-2	*1.613*
-1	*1.748*
0	*1.863*
1	*1.987*
2	*2.019*
3	*2.217*
4	*2.316*

On en déduit $\overline{X} = \dfrac{1}{2}$, $\overline{Y} = \dfrac{15.18}{8} \simeq 1.8975$

$$\sigma_X^2 = 5.25 \quad \sigma_Y^2 \simeq 0.0794$$

Exemple (suite)

La droite $Y = AX + B$ est définie par

$$\begin{aligned} A &= \frac{\sigma_{XY}}{\sigma_X^2} \simeq \frac{0.64}{5.25} \simeq 0.12 \\ B &= \overline{Y} - A\overline{X} \simeq 1.836. \end{aligned}$$

Remarquons que l'on a :

$$\rho_{XY} \simeq \frac{0.64}{\sqrt{5.25 \times 0.0794}} \simeq 0.99$$

ce qui justifie la recherche d'un ajustement exponentiel.

Conclusion :

L'étude des séries statistiques à deux variables permet de mettre en rapport deux caractères afin de pouvoir déterminer une valeur manquante ou de prévoir une tendance.
Néanmoins, deux caractères peuvent avoir un très fort coefficient de corrélation sans pour autant être réellement liés.