Introduction

Il est fréquent de s'interroger sur la relation qui peut exister entre deux grandeurs en particulier dans les problèmes de prévision et d'estimation.

On essaie de déterminer la relation statistique qui existe entre les deux grandeurs x et y. Ce type d'analyse s'appelle analyse de régression. On considère que la variation de l'une des deux variables (par exemple x) explique celle de l'autre (par exemple y).

Dans ce type d'analyse, on fixe à priori les valeurs de x, x n'est donc pas une variable aléatoire. Mais la deuxième grandeur y, est une variable aléatoire et sa distribution est influencée par la valeur de x. Dans ce cas, x est dite variable explicative ou variable indépendante, et Y est dite variable expliquée ou variable dépendante. On a alors, du point de vue statistique, une relation de cause à effet. Le problème sera d'identifier cette relation [12].

II.1. Ajustement de courbes régression et corrélation

II.1.1. Ajustement des courbes

En pratique, il arrive très souvent que l'on mette en évidence relation entre deux (ou plus) variables et que l'on souhaite exprimer cette relation sous forme mathématique en déterminant une équation qui relie ces variables.

Une première étape consiste à recueillir les données correspondant aux différentes valeurs des variables supposons, par exemple, que x et y représentent respectivement la taille et le Poids d'un adulte de sexe masculin. L'étude d'un échantillon nous donnerait les tailles $x_1, x_2, \ldots$ xn et les poids $y_1, y_2, \ldots, y_n$ correspondants.

L'étape suivante consiste à représenter graphiquement les points (x_1, y_1), (x_2, y_2), …, (x_n, y_n) dans un system de coordonnées cartésiennes et l'ensemble de ces points est souvent appelé un diagramme de dispersion.

D'après les données du diagramme de dispersion, il est souvent possible de mettre en évidence une courbe continue qui suit approximativement les données. Une telle courbe est appelée courbe d'ajustement sur la figureII.1, par exemple, la courbe d'ajustement est une droite et nous dirons que la relation est linéaire.

Sur la Figure.II.2, il existe, cependant, une relation entre les variables qui n'est pas linéaire et nous l'appellerons non linéaire.

Sur la Figure II.3, il apparait qu'il n'y a pas de relation entre les variables.

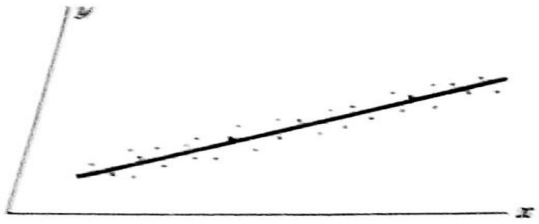

Figure II.1 : Une relation entre les variables linéaire

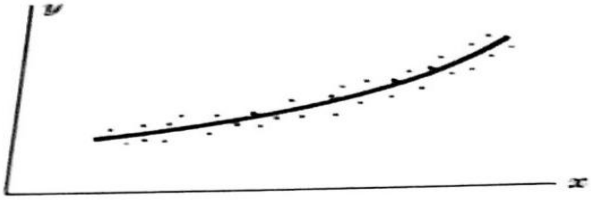

Figure II.2 : Une relation entre les variables non linéaire

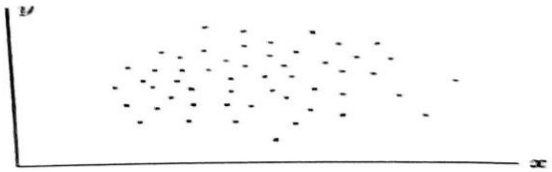

Figure II.3 : aucune relation entre les variables.

Le problème général de trouver des équations d'approximation des courbes qui permettent d'exprimer des données est celui de l'ajustement des courbes pratiquement, le type d'équation est souvent suggéré par le diagramme de dispersion. Ainsi pour la Figure.II.1, nous pourrions écrire l'équation:

$$y = ax + b \qquad \text{II.1}$$

Tandis que pour la FigeurII.2, nous pourrions écrire l'équation:

$$y = a + bx + cx^2 \qquad \text{II.2}$$

Parfois il est intéressant de construire les diagrammes en fonction des variables transformées. Ainsi par exemple, si $logy = f(x)$ conduit à une droite. Nous essaierons $logy = a + bx$ comme équation de la courbe d'approximation.

II.1.2. Régression

Un des objets d'ajustement des courbes est d'obtenir l'expression de l'une des variables (les variables dépendantes) en fonction de l'autre (les variables indépendantes) cette opération d'estimation s'appelle souvent une régression, si y peut-être estimer en fonction de x par l'intermédiaire d'une équation quelconque, cette équation est dite équation de yen x, et la courbe correspondante, la courbe de régression de yen x.

II.1.3 méthode de moindre carré

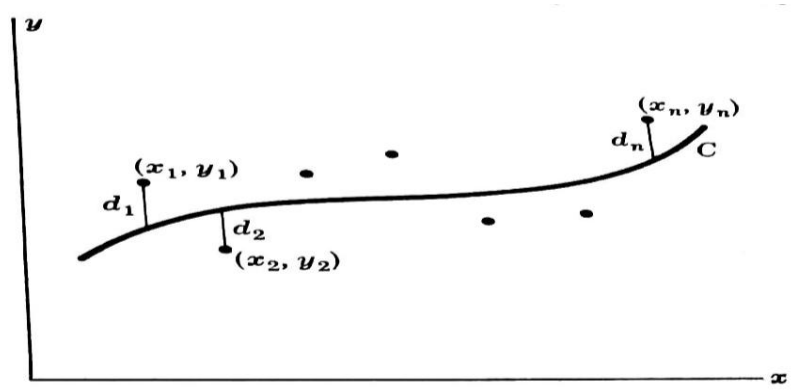

Figure.II.4

Généralement plusieurs courbes sembleront représenter un ensemble des données et pour éviter l'influence du jugement individuel dans la construction de droites paraboles et autre courbe, il est nécessaire de se mettre d'accord sur la meilleure droite, la meilleure parabole etc…

Pour fonder une définition possible considérons la figure. II-4 qui représente les données $(x_1,y_1),\ldots(x_n,y_n)$. Pour une valeur donnée x_1 et x il y aura une différence entre y_1 et la valeur correspondante donnée par la courbe C. cette différence s'écrite d1 qui est aussi appelé erreur, écart ou résidus et qui peut être positive, négative ou nulle.

De même nous obtenons les écarts $d_2,\ldots,d_n$, correspondant aux valeur $x_2,\ldots,x_n$.

La quantité $d_1^2 + d_2^2 + \cdots + d_n^2$ est une mesure de la qualité de l'ajustement de la courbe C aux données. Si elle est faible, l'ajustement est bon. Si elle est grande, celui-ci est mauvais

*II.1.3.1.***Définition**.

Parmi toutes les courbes qui approchent un ensemble donné de points, la courbe présentant la propriété.

$$d_1^2 + d_2^2 + \cdots + d_n^2 = minimum$$

est la meilleure courbe d'ajustement.

Une courbe présentant cette propriété est dite s'ajuster aux donné au sens des moindre carrés et est appelés une courbe de régression des moindres carrés ou plus simplement une courbe des moindres carrés. On a donc des droites des moindres carrés, des paraboles des moindres carrés etc...

Il est habituel d'appliquer la définition ci-dessus quand x est la variable indépendante et y la variable dépendante. Six est la variable dépendante, la définition est modifiée en considérant les écarts horizontaux au lieu des écarts verticaux, ce qui revient à permuter les axes y et x. Ces deux définitions conduisant, en général, à deux courbes des moindre carrés différentes. A moins spécification particulière, nous considérerons que y est la variables dépendante et x la variable indépendante.

Il est également possible de définir un autre type de courbes de moindre carrées en considérant des distances perpendiculaires entre les points et la courbe au lieu des distances horizontale et verticales, mais cette procédure est rarement utilisée.

II.1.4 Droite des moindres carrés

En appliquant la définition présentant, nous pouvons montrer que la droite des moindres carrés approchant l'ensemble des points $(x_1,y_1),\ldots(x_n,y_n)$ a pour équation.

$$y = a + bx \qquad \text{II.3}$$

Ou les constantes *a* et *b* sont déterminées par la résolution simultanée des équations

$$\begin{aligned} \sum y &= an + b\sum x \\ \sum xy &= a\sum x + b\sum x^2 \end{aligned} \qquad \text{II.4}$$

Qui sont dites les équations normales ou équations de normalisation pour droites des moindres Carrés. Remarquons que pour des raisons de brièveté, nous avons employé l'écriture$\sum y$,$\sum xy$au lieu de $\sum_{j=1}^{n} y_j$, $\sum_{j=1}^{n} x_j y_j$Les équations (II.4) sont aisément mémorisées en observant que la première peut-être formellement obtenue par sommation des deux membres de (II.3), tandis que la seconde est obtenue en multipliant d'abord les deux membres de (II.3) par x puis en sommant. Ce n'est pas, bien sûr, une manière de les établir mais, simplement, une manière de s'en souvenir.

Les valeurs de a et de b, tirées de (II.4) s'expriment

$$a = \frac{(\sum y)(\sum x^2)-(\sum x)(\sum xy)}{n\sum x^2-(\sum x)^2}, \qquad b = \frac{n\sum xy-(\sum x)(\sum y)}{n\sum x^2-(\sum x)^2} \qquad \text{II.5}$$

Le résultat pour *b* dans *(II.5)* peut aussi s'écrite

$$b = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2} \qquad \text{II.6}$$

Ici, comme toujours, la barre supérieure indique une *moyenne*, c'est-à-dire$\bar{x} = (\sum x)/n$. La division des deux membres de la première équation normale (II.4) par *n* donne

$$\bar{y} = a + b\bar{x} \qquad \text{II.7}$$

On peut ainsi, si c'est nécessaire, trouver d'abord *b* à partir de (II.5) ou *(II.6)*, puis utiliser (II.7) pour évaluer $a = \bar{y} - b\bar{x}$. Ce qui est équivalent à écrire la droite des moindres carrés sous la forme

$$y - \bar{y} = b(x-\bar{x}) \text{Ou} y - \bar{y} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2}(x-\bar{x}) \qquad \text{II.8}$$

Le résultat (II.8) montre que la constante b, pente de la droite (II.3), est fondamental dans la détermination de la droite. On voit aussi d'après (II.8) que la droite des moindres carres passe par le point $(\bar{x},\bar{y})$, dit centre de gravité des données (ou barycentre).

La pente *b* de la droite de régression est indépendante de 1'origine des coordonnées, ce qui signifie que la transformation (translation des axes).

$$x = x' + hy = y' + k \qquad \text{II.9}$$

Ou *h* et *k* sont des constantes quelconques, conduit à une valeur de *b*

$$b = \frac{n\sum x'y' - (\sum x')(\sum y')}{n\sum x'^2 - (\sum x')^2} = \frac{\sum(x'-\bar{x}')(y'-\bar{y}')}{\sum(x'-\bar{x}')^2} \qquad \text{II.10}$$

Ou x et *y* sont simplement remplacées par x' et y'(on dit, alors, que *b* est invariante par la transformation (II.3)). II faut remarquer cependant que a, qui détermine l'intersection avec *l'axe* des x, dépend de I 'origine (et n'est pas invariant).

Dans le cas particulier *ou* $h = \bar{x}, k = \bar{y}$*,(II.10)* se simplifie en

$$b = \frac{\sum x'y'}{\sum x'^2} \qquad \text{II.11}$$

Les formules (II.10) ou (II.11) sont souvent utiles car ils simplifient les calculs nécessaires pour

obtenir la droite des moindres carrés.

Les remarques ci-dessus sont également valables pour la droite de régression de y sur x.il suffit de permuter x et y. ainsi, la droite de régression des moindres carrés de x sur y est

$$x - \bar{x} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(y-\bar{y})^2}(y-\bar{y}) \qquad \text{II.12}$$

En général, (II.12) n'est pas la même droite que (II.8).

II.1.5 droite des moindres carrés en fonction des variances et de la covariance des échantillons

Les variances et covariance d'échantillons de x et y s'expriment

$$s_x^2 = \frac{\sum(x-\bar{x})^2}{n}, s_y^2 = \frac{\sum(y-\bar{y})^2}{n}, s_{xy} = \frac{\sum(x-\bar{x})(y-\bar{y})}{n} \qquad \text{II.13}$$

En fonction de ces valeurs, les droites des moindres carrés pour y sur x et x sur y s'expriment, respectivement

$$y - \bar{y} = \frac{s_{xy}}{s_x^2}(x-\bar{x}) x - \bar{x} = \frac{s_{xy}}{s_y^2}(y-\bar{y}) \qquad \text{II.14}$$

Et si nous définissons le coefficient de corrélation d'échantillon par :

$$r = \frac{s_{xy}}{s_x s_y} \qquad \text{II.15}$$

(II.14) peut s'écrire

$$\frac{y-\bar{y}}{s_y} = r\left(\frac{x-\bar{x}}{s_x}\right)\frac{x-\bar{x}}{s_x} = r\left(\frac{y-\bar{y}}{s_y}\right) \qquad \text{II.16}$$

Puisque $(x-\bar{x})/s_x$ et $(y-\bar{y})/s_y$ sont les valeurs d'échantillon réduites, les formules en (II.16) peuvent permettre de mémoriser simplement les droites de régression. Il est clair que les deux droites de (II.16) différentes, à moins que r = ±1, auquel cas les points d'échantillonnage sont sur une droite (ce qui sera montré en (II.26) et il y a, alors, régression et corrélation linéaires parfaites).

Notons, aussi, que si les deux droites de régression (II.16) sont écrites $y = a + bx,\ x = c + dx$, respectivement,

$$bd = r^2 \qquad \text{II.17}$$

jusqu'ici nous n'avons pas considéré la signification précise du coefficient de corrélation, nous contentant seulement de l'exprimer de façon formelle en fonction des variances et de la covariance..

II.1.6 parabole des moindres carrés

Les raisonnements ci-dessus sont facilement généralisés. Par exemple, l'expression de la parabole des moindres carrés pour un ensemble de points donnés est :

$$y = a + bx + cx^2 \qquad \text{II.18}$$

Ou a, b et c sont déterminées par les équations normales

$$\begin{aligned} \sum y &= na + b\sum x + c\sum x^2 \\ \sum xy &= a\sum x + b\sum x^2 + c\sum x^3 \\ \sum x^2 y &= a\sum x^2 + b\sum x^3 + c\sum x^4 \end{aligned} \qquad \text{II.19}$$

Qui sont obtenues formellement en sommant les deux membres de *(II.18)* après les avoir multiplié par 1, x et x^2, respectivement.

II.1.7. Régression multiple

Les raisonnements précédents peuvent également être généralises à un plus grand de variables. Si nous sentons, par exemple, qu'il existe une relation linéaire entre une variable dépendante z et deux variables indépendantes x et y, nous chercherons une équation de forme.

$$z = a + bx + cy \qquad \text{II.20}$$

Qui est une équation de régression de z sur x et y. Si c'est x qui est la variable indépendante, une équation du mère type sera dite équation de régression de x sur y et z.

Comme *(II.20)* représente 1'équation d'un plan dans un espace a trois dimensions, ce plan est souvent appelé *plan de régression*. Pour trouver son expression, on détermine a, b et c de l'expression *(II.20)* de sorte que

$$\sum z = na + b\sum x + c\sum y$$
$$\sum xz = a\sum x + b\sum x^2 + c\sum xy \qquad \text{II.21}$$
$$\sum yz = a\sum y + b\sum xy + c\sum y^2$$

Qui sont les équations normales correspondant à*(II.20)* et qui sont obtenues en appliquant une définition analogue**.** Notons qu'elles peuvent être tirées formellement de *(II.20)* en multipliant par 1, *x* et x^2, puis en sommant.

Des généralisations a plus de variables encore peuvent conduire à des surfaces de régression d'ordres supérieurs.

II.1.8 Erreur type d'estimation

Soit y_{est} la valeur estimée de *y* pour une valeur donnée de *x,* obtenue à partir de la courbe de régression de *y* sur *x*, la quantité.

$$s_{yx} = \sqrt{\frac{\sum(y - y_{est})^2}{n}} \qquad \text{II.22}$$

Donne, alors, une mesure de la dispersion par rapport à la courbe de régression et s'appelé erreur-type d'estimation de y sur x. Puisque $\sum(y - y_{est})^2 = \sum d^2$, nous voyons que parmi toutes les courbes de régression possibles, celle de moindres carres présente la plus faible erreur-type d'estimation [12].

Dans le cas d'une droite de régression$y_{est} = a + bx$,avec a et b donnés par (II.4), nous avons.

$$s_{yx}^2 = \sum y^2 - a\sum y - b\sum xy \qquad \text{II.23}$$

$$s_{yx}^2 = \sum(y - \bar{y}) = b\sum(x - \bar{x})(y - \bar{y}) \qquad \text{II.24}$$

Nous pouvons également exprimer s_{yx}^2 en fonction de la variance et du coefficient de corrélation sous la forme,

$$s_{yx}^2 = s_y^2(1 - r^2) \qquad \text{II.25}$$

Dont un corollaire est $r^2<1$, c'est-à-dire -1< r <1

L'erreur type d'estimation a des propriétés analogues à celles de l'écart-type il nous construisons par exemple, des paires de droit parallèles a la droite de régression de y sur x situées à des distances verticales s_{xx}, $2s_{yx}$ et $3s_{yx}$ de celle-ci pour n suffisamment posé nous trouverons 68%,95% et 99.7% des point d'échantillonnage inclus dans les intervalles d'espace.

De même qu'il y a une estimation non baisse de la variance d'échantillon donnée par $\bar{x}^2 = nx^2/(n-1)$, il existe une estimation non basée du carré de l'erreur type qui s'exprime $s_{y.x}^2 = ns_{y.x}^2/(n-2)$ c'est pour cette raison que certaines statisticiennes prirent donnée l'expression (II.22) avec n-2 au dénominateur au lieu de n.

Les remarque ci-dessus sont aisément modifiées pour le cas de la régression de x sur y (ou l'erreur-type est notée $s_{x.y}$ ou pour les régressions non linéaires ou multiple.

II.1.9 coefficient de corrélation linéaire

Nous allons maintenant examiner la signification du coefficient de corrélation dont nous n'avons donné jusqu'ici que l'expression (II.15). Pour ce faire, notons que, d'après (II.25) et les définitions de $s_{y.x}$ et s_y nous avons [12] :

$$r^2 = 1 - \frac{\sum(y - y_{est})^2}{\sum(y - \bar{y})^2} \qquad \text{II.26}$$

Nous pouvons maintenant montrer que :

$$\sum(y - \bar{y})^2 = \sum(y - y_{est})^2 + \sum(y_{est} - \bar{y})^2 \qquad \text{II.27}$$

Le membre de gauche de (II.27) représente l'écart total.la première somme dans le membre de droite est l'écart inexpliqué ou écart résiduel ou écart aléatoire, l'écart expliqué. Cette terminologie provient du fait que $y - y_{est}$ se comporte aléatoirement ou de manière imprévisible, tandis que les écarts $y_{est} - \bar{y}$ s'expliquent par les droites de régression de moindres carrés et à suivre un modèle défini.il résulte de (II.26) et (II.27) que.

$$r^2 = \frac{\sum(y_{est} - \bar{y})^2}{\sum(y - \bar{y})^2} = \frac{Ecart\ expliqué}{Ecart\ total} \qquad \text{II.28}$$

Ainsi r^2peut-êtreinterprété comme la fonction de l'écart total qui peut être interprétée par la droite de régression. En d'autre termes, r mesure la qualité d'ajustement des données d'échantillonnage par la régression des moindres carrés. Si la variation totale est complément expliquée par la droite de régression, c'est-à-dire que r= ±1, nous disons qu'il y a corrélation linéaire parfaite (et dans ce cas également régression linéaire parfaite). Par ailleurs, si l'écart total est complètement inexpliqué, l'écart expliqué est nul et r=0. En pratique, le coefficient r, coefficient de détermination, est compris entre 0 et 1.

Le coefficient de corrélation peut être tiré de l'une des expressions

$$r = \frac{s_{xy}}{s_x s_y} = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2}\sqrt{\sum(y - \bar{y})^2}} \qquad \text{II.29}$$

$$r^2 = \frac{Ecart\ expliqué}{Ecart\ total} = \frac{\sum(y_{est} - \bar{y})^2}{\sum(y - \bar{y})^2} \qquad \text{II.30}$$

Qui sont équivalentes dans le cas de la régression linéaire. L'expression (II.27) est souvent dite formule du moment-produit pour la corrélation linéaire.

Les formules suivantes, équivalentes aux expressions précédentes, sont souvent utilisées,

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2\ (\sum y)^2]}} \qquad \text{II.31}$$

Et

$$r = \frac{\overline{xy} - \bar{x}\bar{y}}{\sqrt{(\overline{x^2} - \bar{x}^2)(\overline{y^2} - \bar{y}^2)}} \qquad \text{II.32}$$

En appliquant la transformation (II.9) il vient :

$$r = \frac{n\sum x'y' - (\sum x')(\sum y')}{\sqrt{[n\sum x'^2 - (\sum x')^2][n\sum y'^2 - (\sum y')^2]}} \qquad \text{II.33}$$

Qui montre que r est invariant dans un changement d'axes par translation. En particulier si $h = \bar{x}, k = \bar{y}$.

$$r = \frac{\sum x' y'}{\sqrt{\sum (x'^2)\,(\sum y'^2)}} \qquad \text{II.34}$$

Est une autre expression de (II.33), souvent utile pour les calculs.

La corrélation linéaire peut-être positive ou négative. Si r>0.y croit avec x (la pente de la droite des moindres carrés est positive), tandis que si r est négatif, y tend à décroitre avec x (la pente de la droite des moindres carrés est négative). Ce signe est automatiquement pris en compte si nous utilisons les expressions (II .29), (II.31),(II.32)(II. 33) ou (II.34). Par contre, si nous utilisons (II.30) pour calculer r, nous devons y ajouter le signe convenable.

II.1.10 coefficient de corrélation généralisé

La définition (II.29) (ou toute autre définition équivalente de (II.31) à (II.34)) du coefficient de corrélation ne comporte que des valeurs d'échantillonnage x et y. elle conduit, par conséquent, au même nombre pour toutes les formes de courbes de régression et ne peut être utilisée comme mesure de l'ajustement, sauf dans le cas de la régression linéaire ou elle coïncide avec (II.33).la définition, cependant, soit

$$r^2 = \frac{ecart\ expliqué}{ecart\ total} = \frac{\sum (y_{est} - \bar{y})^2}{\sum (y - \bar{y})^2} \qquad \text{II.35}$$

Réfléchit la forme de la courbe de régression (via y_{est}) et peut convenir pour définir un coefficient de corrélation généralisé r. nous utilisons (II.35) pour obtenir des coefficients de corrélation non linéaire (qui mesurent la qualité d'ajustement par une courbe de régression non linéaire) ou, au moyen d'une généralisation appropriée, les coefficients de corrélation multiple. La relation (II.25) entre le coefficient de corrélation et l'erreur-type d'estimation est également valable pour une corrélation non linéaire.

II.1.11 corrélation d'ordre

Au lieu d'utiliser des valeurs précises d'échantillonnage, ou dans le cas où la précision ne peut-être approchée, les données peuvent être ordonnées par ordre de grandeur, d'importance, etc…, en

utilisant les nombres 1,2,…,n.si deux ensembles de valeurs de x et y correspondantes sont rangées de cette façon, le coefficient de corrélation d'ordre, noté r_{rang} ou, tout simplement r, s'exprime

$$r_{rang} = 1 - \frac{6\sum d^2}{n(n^2-1)} \qquad \text{II.36}$$

Ou d=différence entre les ranges des x et y correspondants

n=nombres de paires de valeurs (x,y) dans les données.

L'expression (II.36), établie la formule de corrélation d'ordre.