CHAPITRE 7

AJUSTEMENT LINÉAIRE, CORRÉLATION

- *Nuages de points*
- *Ajustement linéaire*
- *Méthode des moindres carrés*
- *Droites de régression*
- *Coefficient de corrélation*
- *Application aux modèles exponentiels*

REPÈRES

Les chapitres précédents ont montré comment étudier une série statistique, mais deux séries ou deux caractères enregistrés à l'intérieur d'une même série peuvent être reliés entre eux. On examinera les plus simples de ces relations : les relations linéaires.

1. Données, nuages de points

On note : (x_1,y_1), (x_2,y_2), ... , (x_N,y_N) la série des observations relevant deux caractères quantitatifs x et y pour les N individus d'une population. Exemples : la taille et le poids d'un groupe d'étudiants, le PNB par habitant et le taux d'alphabétisation de pays en voie de développement, le revenu et la consommation de différents ménages, etc.

Des unités étant convenablement choisies sur chacun des axes, on peut représenter l'individu i de la population précédente par le point : (x_i,y_i) du plan $x \times y$. Figurant ainsi les N individus, on obtient le *nuage de points* associé à la série statistique.

Les nuages de points associés à des séries statistiques à deux caractères peuvent présenter différentes formes :

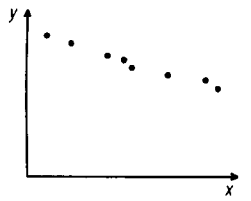


Figure 1

Figure 2

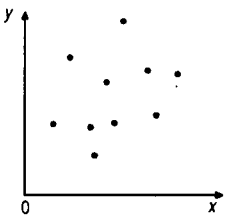


Figure 3

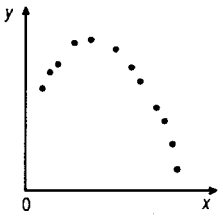


Figure 4

Les points du nuage 1 sont presque alignés, tandis que le nuage 2 laisse simplement apparaître une direction d'allongement privilégiée. Dans ces deux cas, on dit que le nuage présente un caractère linéaire. Le nuage 3 ne manifeste pas de structure particulière ; le nuage 4, enfin, semble se placer approximativement selon une courbe régulière.

L'ajustement linéaire est la recherche de la droite résumant le mieux la structure du nuage. Une telle recherche n'a donc d'intérêt que pour des nuages de l'un des deux premiers types.

2. Ajustement linéaire

a) Ajustement graphique

Lorsque le nuage présente un caractère linéaire, on peut tenter de tracer « à main levée » la droite qui résume le mieux la structure du nuage. La subjectivité du procédé est évidente.

b) Méthode des moindres carrés
Droite de régression de Y en X

On considère le plus souvent que l'un des caractères, ou l'une des variables, dépend de l'autre (par exemple, la consommation dépend du revenu) ; soit Y le premier caractère, ou *variable à expliquer*, et X le second, ou *variable explicative*. On cherche une expression de Y en fonction de X, de la forme $y = a \cdot x + b$, qui approche « le mieux » les données.

D'un point de vue géométrique, cela revient à chercher la droite d'équation : $y = a \cdot x + b$, qui traduit le mieux l'aspect linéaire du nuage de points.

● Critère des moindres carrés :

Soit : $y = a \cdot x + b$, la droite retenue ; on donne les définitions (fig. 5) :

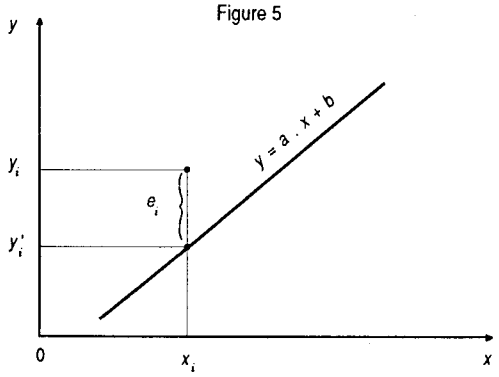
• x_i est la *valeur observée* de la variable explicative X pour l'individu i ;

• y_i est la *valeur observée* de la variable à expliquer Y pour l'individu i ;

• $y'_i = a \cdot x_i + b$ est la *valeur théorique* ou *ajustée* de la variable y, associée à la valeur observée x_i de la variable X ;

• $e_i = y_i - y'_i$ est l'*erreur d'ajustement*, c'est-à-dire l'écart entre la valeur observée et la valeur théorique calculée de Y.

Figure 5



La « meilleure » droite retenue est en fait celle qui rend minimale la somme des carrés des erreurs d'ajustement. On l'appelle *droite des moindres carrés* de y en x, ou *droite de régression* de y en x.

● Calculs :

Les coefficients a et b de la droite des moindres carrés de Y en X s'expriment en fonction des observations : (x_i, y_i).

On montre que a et b sont donnés par :

$$a = \frac{\sum (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum (x_i - \bar{x})^2},$$

et $\bar{y} = a \cdot \bar{x} + b$.

Cette dernière relation signifie que la droite passe par le point moyen : $(\bar{x}, \bar{y})$, ce qui permet de calculer b après a.

Pratiquement, ces calculs s'apparentent au calcul d'une variance. Si on doit utiliser une calculatrice, on peut calculer successivement les moyennes $\bar{x}$ et $\bar{y}$ de x et y, les deux séries d'écarts à la moyenne, la série des produits : $(x_i - \bar{x}) \cdot (y_i - \bar{y})$; celle des carrés : $(x_i - \bar{x})^2$; et, enfin, la somme de ces séries et leur rapport.

Comme dans le cas de la variance, le numérateur et le dénominateur du coefficient a s'expriment sous une autre forme qui peut simplifier le calcul à la main :

$$a = \frac{\sum x_i \cdot y_i - N \cdot \bar{x} \cdot \bar{y}}{\sum x_i^2 - N \cdot \bar{x}^2}.$$

L'emploi de cette expression dispense du calcul des écarts aux moyennes.

Quelle que soit l'expression de a utilisée, ces calculs sont aussi fastidieux que peu instructifs. Ils se font en revanche très simplement sur un tableur, quand ils ne sont pas déjà programmés, comme sur les calculatrices statistiques ou mathématiques ; dans ce cas, il suffit d'entrer les séries d'observations de x et y puis d'utiliser la commande appropriée pour lancer le calcul.

c) Droite de régression de X en Y

On a noté que la régression de y en x donne des rôles différents aux deux variables. On peut renverser ces rôles et régresser la variable X en Y. On cherche une expression : $x = c \cdot y + d$, de X en fonction de Y, optimale au sens des moindres carrés.

La simple transposition des formules précédentes, donnant les valeurs de a et b, donne c et d :

$$c = \frac{\sum (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum (y_i - \bar{y})^2},$$

et $\bar{x} = c \cdot \bar{y} + d$.

Les deux droites de régression (de Y en X et de X en Y) représentées dans le même repère $x \times y$ ne sont confondues que si les points du nuage sont exactement alignés.

Elles diffèrent d'autant plus que le nuage s'éloigne de l'alignement.

3. Corrélation

a) Coefficient R^2

On montre que la série des valeurs ajustées y'_i, dans la régression de la variable y sur x, a la même moyenne $\bar{y}$ que la série initiale des y_i.

On définit le coefficient noté R^2 :

$$R^2 = \frac{\sum (y'_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}.$$

Ce rapport entre la « variation expliquée » par la régression et la « variation totale » mesure la qualité de la régression : si celle-ci est bonne, les y'_i approchent les y_i et R^2 est proche de 1.

b) Coefficient de corrélation linéaire R

En fait, on calcule plus fréquemment le *coefficient de corrélation linéaire*, noté R (on montre en effet qu'il est égal, au signe près, à la racine carrée de la quantité R^2 précédente) :

$$R = \frac{\sum (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \cdot \sqrt{\sum (y_i - \bar{y})^2}} ;$$

on voit sur cette expression que les variables X et Y jouent des rôles symétriques.

Le coefficient R mesure non seulement la qualité de l'une ou l'autre régression, mais, plus généralement, le caractère linéaire du nuage de points, ou encore l'intensité de la liaison linéaire entre les deux variables (en particulier lorsqu'aucune des deux ne paraît devoir « expliquer » l'autre, ainsi du taux d'équipement en machines à laver et du taux d'équipement en réfrigérateurs).

Son emploi est précisé par les propriétés suivantes :

- R est toujours compris entre – 1 et + 1 ;
- R vaut + 1 (respectivement – 1) lorsque les points sont alignés sur une droite ascendante, traduisant une variation dans le même sens des deux caractères (respectivement descendante, pour une variation de sens contraire) ;
- R est proche de + 1 (respectivement – 1) lorsque les caractères montrent une liaison linéaire marquée et croissante (respectivement décroissante). En ce cas, la régression est *a priori* intéressante, et les deux droites de régression ne seront guère éloignées ;
- R est proche de 0 en l'absence de liaison linéaire apparente, la régression linéaire est alors peu justifiée.

R peut être calculé en premier lieu (c'est-à-dire avant les droites de régression) et — par exemple dans le dernier cas — ne pas donner suite à une régression.

R est souvent appelé « coefficient de corrélation », en omettant le qualificatif « linéaire ».

c) Autres formules

Comme la variance, R peut être calculé sans passer par les écarts aux moyennes, à l'aide de la formule :

$$R = \frac{\sum x_i \cdot y_i - N \cdot \bar{x} \cdot \bar{y}}{\sqrt{\sum x_i^2 - N \cdot \bar{x}^2} \cdot \sqrt{\sum y_i^2 - N \cdot \bar{y}^2}}.$$

On a également la relation liant R aux pentes des deux droites de régression (de Y sur X et de X sur Y) :

$$R^2 = a \cdot c.$$

4. Élargissements

a) Modèles exponentiels

Les relations de nature linéaire, pour être simples à étudier, ne sont pas les seules qui puissent exister entre deux grandeurs. D'autres peuvent cependant s'y ramener.

Si Y et X sont liés par une relation de la forme $Y = A \cdot B^X$, on dit que Y dépend exponentiellement de X. Cette relation n'est pas linéaire, mais, en prenant les logarithmes, on a :

$$\ln(Y) = \ln(B) \cdot X + \ln(A),$$

c'est-à-dire une relation linéaire entre la variable transformée, ln(Y), et X, qui s'écrit encore : $\ln(Y) = \alpha \cdot X + \beta$, en notant $\alpha = \ln(B)$ et $\beta = \ln(A)$.

L'étude de cette dernière relation s'effectue par régression linéaire ; on revient ensuite à la forme exponentielle de départ.

b) Régression multiple

Certains modèles, plus élaborés, veulent expliquer une grandeur en fonction de plusieurs autres (par exemple, la consommation d'un ménage pourra dépendre de son revenu, mais également de son effectif).

Bien que les idées de départ soient les mêmes, la représentation graphique n'est plus possible ; les calculs ne se font plus à la main et leur interprétation est plus délicate.