

Série statistique à deux variables.

Problèmes d'ajustement

Le problème qui se pose dans les séries statistiques à deux variables est principalement celui du lien qui existe ou non entre chacune des variables. En effet, l'étude séparée de chacune des variables revient à l'étude de séries à une variable, ce qui a été l'objet du chapitre précédent. Nous nous intéresserons donc exclusivement au problème de l'ajustement, qui est celui de savoir si un lien existe entre deux séries de mesure, ou si les événements dont elles sont issues sont sans rapport. Il convient d'être prudent sur la nature du travail que nous allons mener : en effet, établir un rapport entre deux séries de nombres ne signifie nullement qu'il y ait un quelconque rapport de cause à effet entre les deux, assimilation qui est malheureusement fort courante dans les médias : des événements "fortement corrélés" sont très souvent présentés au public, avec l'idée d'une causalité forte entre eux.

1 Position du problème. Vocabulaire

Dans tout ce chapitre, nous ne parlerons que de séries statistiques discrètes. Nous débutons par un exemple, destiné à éclairer les enjeux et méthodes de ce chapitre.

Exemple 1 *La consommation moyenne aux 100 kilomètres d'une automobile a été relevée à diverses vitesses. On a obtenu les résultats suivants :*

Vitesse	10	20	30	40	50	60	70	80	90	100	110	120	130
Consommation moyenne aux 100 Km	17	11.2	9.1	8	7.3	7.1	6.9	7.5	8.6	10.1	10.9	11.7	13

Un acheteur se demande quelle est la nature précise du rapport qui existe - ou pas - entre la consommation et la vitesse du véhicule.

1.1 Notion de nuage de points

La première réponse qui peut être formulée tient à la réalisation d'un graphique qui traduise les deux séries statistiques ci-dessus. En plaçant par exemple la vitesse en abscisses, et la consommation en ordonnées, on peut représenter par un point chaque état relevé expérimentalement. On obtient ainsi une succession de points, dont les coordonnées sont les suivantes : (10; 17), (20; 11.2), (30; 9.1), (40; 8), ... (130; 13), qui forment un nuage de points. Ce nuage traduit graphiquement une partie de l'information contenue dans le tableau, de manière beaucoup plus immédiatement perçue : le fait que pour des vitesses extrêmes, la consommation augmente. Le nuage de points associé à une série statistique à deux variables donne donc immédiatement des informations de nature qualitatives. Pour en tirer des informations plus quantitatives, il nous faut poser le problème de l'ajustement.

1.2 Le problème de l'ajustement

Dans notre exemple, la disposition du nuage de points ne permet pas de tracer une droite qui rendrait convenablement compte de la structure sous-jacente du nuage de points. En revanche, une portion de parabole peut sembler plus appropriée.

Ces tracés mettent en évidence la possibilité de "reconnaître" graphiquement la possibilité d'une relation fonctionnelle entre les deux grandeurs observées (vitesse et consommation). Le problème de l'établissement d'une relation fonctionnelle entre les deux séries observées est le problème de l'ajustement. Ce problème est en général complexe, et nécessite des outils mathématiques hors de portée de ce cours. Nous nous contenterons d'explorer le modèle de l'ajustement affine, qui consiste à chercher la relation fonctionnelle affine reliant les grandeurs considérées (ce qui revient graphiquement à rechercher la droite qui traduise le mieux la structure du nuage). De ce cas particulier, on peut élargir notre propos à d'autres types d'ajustements, qui seront présentés en exercices.

1.3 Point moyen

Définition 1 Soit Z une série statistique à deux variables, X et Y , dont les valeurs sont des couples $(x_i; y_i)$ (dans ces conditions, chacune des deux séries statistiques X et Y composant Z sont à une variable). On appelle point moyen de la série Z , le point de coordonnées $(\bar{X}; \bar{Y})$.

1.4 Covariance de deux séries statistiques

Définition 2 Soient X et Y deux séries statistiques à une variable, dont les n valeurs sont respectivement des réels $(x_i)_{1 \leq i \leq n}$ et $(y_i)_{1 \leq i \leq n}$. On appelle covariance de X et Y le réel, noté $\text{Cov}(X; Y)$, défini par la relation

$$\text{Cov}(X; Y) = \frac{(x_1 - \bar{X})(y_1 - \bar{Y}) + (x_2 - \bar{X})(y_2 - \bar{Y}) + \dots + (x_n - \bar{X})(y_n - \bar{Y})}{n} = \frac{1}{n} \sum_{i=1}^n (X - \bar{X})(Y - \bar{Y}),$$

où $\bar{X}$ et $\bar{Y}$ sont respectivement les moyennes des séries X et Y considérées.

Remarque 1 On a en particulier dans ce cadre, $\text{Cov}(X; X) = \mathbb{V}(X)$.

Proposition 1 Si les séries statistiques X et Y sont indépendantes, alors $\text{Cov}(X; Y) = 0$. La réciproque est fausse.

Remarque 2 Cette proposition est équivalente à la formulation ci-dessous :

Si la covariance $\text{Cov}(X; Y)$ de X et Y est non-nulle, alors X et Y ne sont pas indépendants

(et dans ce cas, il existe une forme de lien entre eux.

Attention aux charlatans de tous poils. Sur la question de l'inocuité des ondes GSM des téléphones portables pour la santé, la communauté scientifique affirme que la covariance des effets sur la santé et des émissions est voisine de 0. Ceci signifie qu'on n'est pas en mesure de conclure à un lien entre les deux. Rien ne prouve cependant qu'il y ait un lien ou pas ... contrairement à ce qu'affirment ceux qui disent que les événements sont non reliés !

Proposition 2 Si X et Y sont deux séries statistiques, on peut aussi calculer leur covariance au moyen de la relation

$$\text{Cov}(X; Y) = \frac{x_1 y_1 + x_2 y_2 + \dots + x_n y_n}{n} - \bar{X} \bar{Y} = \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{X} \bar{Y} = \bar{X} \cdot \bar{Y} - \bar{X} \bar{Y}$$

Preuve

◀ La démonstration ci-dessus consiste à remarquer que pour tout $i \in \{1; 2; \dots; n\}$, on a

$$(x_i - \bar{X})(y_i - \bar{Y}) = x_i y_i - x_i \bar{Y} - y_i \bar{X} + \bar{X} \bar{Y}$$

et à conclure par une réécriture des sommes utilisant cette relation. ▶

2 Méthodes d'ajustement affine

2.1 Méthode artisanale (à la règle)

Cette méthode est purement graphique : ayant tracé le nuage de point associé à une série statistique à deux variables, on tente, au jugé, de placer la règle de manière à tracer une droite équilibrant de part et d'autre les points du nuage, et qui soit de même direction que celle qui semble poindre sous le nuage de point. Ayant tracé cette droite, on détermine graphiquement une équation de celle-ci par les méthodes usuelles de Seconde, ou du Collège.

Cette méthode très approximative peut être un peu améliorée en imposant à la droite tracée de passer par le point moyen du nuage, que dont on aura au préalable calculé les coordonnées.

On rappelle néanmoins comment obtenir les coefficients de l'équation d'une droite dont on connaît, dans un repère $O; \vec{i}; \vec{j}$ deux points :

Théorème 3 Soit $(O; \vec{i}; \vec{j})$ un repère du plan. Toute droite non parallèle à l'axe des ordonnées admet une équation réduite de la forme $y = ax + b$ pour des réels uniques a et b .

En outre, on a

$$a = \frac{y_B - y_A}{x_B - x_A} \text{ et } b = \frac{x_A y_B - x_B y_A}{x_B - x_A}.$$

Définition 3 Avec les notations précédentes, a est appelé le coefficient directeur de la droite, et b est appelé l'ordonnée à l'origine.

Preuve

◀ Soit $\mathcal{D}$ une droite non parallèle à l'axe des ordonnées, et $A(x_A; y_A)$ et $B(x_B; y_B)$ deux points distincts de $\mathcal{D}$. Comme $\mathcal{D}$ n'est pas parallèle à l'axe des ordonnées, on a $x_A \neq x_B$, ou encore $x_B - x_A \neq 0$.

Soit $M(x; y)$ un point du plan. On a $\overrightarrow{AM} \begin{pmatrix} x - x_A \\ y - y_A \end{pmatrix}$, $\overrightarrow{AB} \begin{pmatrix} x_B - x_A \\ y_B - y_A \end{pmatrix}$ et

$$\begin{aligned} M \in (AB) &\Leftrightarrow \overrightarrow{AM} \text{ col. } \overrightarrow{AB} \Leftrightarrow \begin{vmatrix} x - x_A & x_B - x_A \\ y - y_A & y_B - y_A \end{vmatrix} = 0 \\ &\Leftrightarrow (y_B - y_A)(x - x_A) - (y - y_A)(x_B - x_A) = 0 \\ &\Leftrightarrow y_B x - y_B x_A - y_A x + y_A x_A - x_B y + x_A y + y_A x_B - y_A x_A = 0 \\ &\Leftrightarrow (y_B - y_A)x + (x_A - x_B)y + (x_B y_A - x_A y_B) = 0 \\ &\Leftrightarrow y = \frac{y_B - y_A}{x_B - x_A} x + \frac{x_A y_B - x_B y_A}{x_B - x_A} \text{ car } x_B - x_A \neq 0. \end{aligned}$$

Remarquons que la relation $y_B - y_A \neq 0$ traduit simplement le fait que les points A et B sont placés de sorte que la droite (AB) soit parallèle à l'axe des ordonnées.

On pose alors $a = \frac{y_B - y_A}{x_B - x_A}$ et $b = \frac{x_A y_B - x_B y_A}{x_B - x_A}$ et le résultat est acquis. ▶

Remarque 3 Enfin, on peut affiner un peu le tracé initial de la droite cherchée, en imposant que celle-ci passe par le point moyen G du nuage. On procède alors comme indiqué précédemment pour obtenir une équation de la droite.

2.2 Ajustement affine selon la méthode de Meyer

Il s'agit d'un perfectionnement relativement élémentaire, qui peut s'avérer utile lorsque l'on ne souhaite pas appliquer la méthode des moindres carrés, qui nécessite une calculatrice. Le principe est le suivant, et s'appuie sur la dernière amélioration présentée dans le cas d'une droité tracée "au jugé" : on partage la série statistique Z en deux sous séries équilibrées et approximativement de même taille, Z_1 et Z_2 , et on calcule les coordonnées des points moyens G_1 et G_2 , associés au deux séries extraites. La droite choisie pour réaliser l'ajustement graphique est la droite $(G_1 G_2)$, dont on détermine une équation par le calcul.

Notons que cette méthode n'est pas adaptée lorsque l'on se trouve face à une série statistique comportant peu de points, ou lorsque les points sont visiblement mal alignés.

Proposition 4 Le moyen moyen G de la série Z est situé sur toutes les droites $(G_1 G_2)$, où G_1 et G_2 sont les points moyens de n'importe quelles sous-séries Z_1 et Z_2 de Z .

Preuve

◀ Cette proposition est une conséquence des propriétés d'associativité des moyennes. ▶
Ceci conduit à ce que l'on ait $G \in (G_1 G_2)$.

2.3 Ajustement affine selon les moindres carrés

Cette méthode, bien que plus perfectionnée que la méthode artisanale, repose néanmoins sur la même idée. Il s'agit d'obtenir une droite équidistante des points situés de part et d'autre d'elle-même. Pour réaliser ceci, on considère les distances de tous les points à la droite, et on cherche à minimiser la somme de ces distances. Pour des raisons qui sortent de notre propos, on utilise la distance quadratique.

2.3.1 Distance quadratique

Définition 4 Si deux points $P(x_P; y_P)$ et $Q(x_Q; y_Q)$ sont deux points du plan muni d'un repère orthonormal, la distance quadratique de P à Q est le nombre $PQ^2 = (x_Q - x_P)^2 + (y_Q - y_P)^2 = \|\overrightarrow{PQ}\|^2$.

2.3.2 Droite de régression

Dans le plan muni d'un repère orthonormal, on considère un nuage de n points de coordonnées $(x_i; y_i)_{1 \leq i \leq n}$, représentant deux séries statistiques X et Y à une variable, que l'on cherche à ajuster de manière affine. Soit $\mathcal{D}$ une droite d'équation $y = ax + b$. Cette droite est appelée droite de régression de y en x ssi la quantité $\sum_{i=1}^n (ax_i + b - y_i)^2$ est minimale.

Remarque 4 Il serait tout aussi judicieux de s'intéresser à la droite $\mathcal{D}'$ d'équation $x = ay + b$, qui minimise la quantité $\sum_{i=1}^n (ay_i + b - x_i)^2$ minimale. Cette droite est appelée droite de régression de x en y et elle est en générale différente de la précédente.

Théorème 5 Sous les hypothèses précédentes, il existe une droite $\mathcal{D}$ qui rend la quantité $\sum_{i=1}^n ((ax_i + b) - y_i)^2$ minimale. Cette droite $\mathcal{D}$ est d'équation $y = ax + b$, avec les valeurs a et b suivantes :

$$\begin{cases} a = \frac{\text{Cov}(X; Y)}{\mathbb{V}(X)} = \frac{\text{Cov}(X; Y)}{\sigma_X^2} \\ \text{et} \\ b = \bar{Y} - a\bar{X} \end{cases}$$

Preuve

◀ On considère la fonction $f : (a; b) \mapsto \sum_{i=1}^n ((ax_i + b) - y_i)^2$. Cette fonction est polynômiale, donc différentiable sur $\mathbb{R}^2$. Si $(a; b)$ réalise un minimum de cette fonction, alors en ce point on a

$$\frac{\partial f}{\partial a}(a; b) = 0 \text{ et } \frac{\partial f}{\partial b}(a; b) = 0.$$

$$\text{Or } \frac{\partial f}{\partial a}(a; b) = 2 \sum_{i=1}^n x_i(ax_i + b - y_i) \text{ et } \frac{\partial f}{\partial b}(a; b) = 2 \sum_{i=1}^n (ax_i + b - y_i).$$

De cette seconde relation, on déduit que $\sum_{i=1}^n ax_i - y_i = nb$, soit donc $b = a \frac{1}{n} \sum_{i=1}^n x_i - \frac{1}{n} \sum_{i=1}^n y_i = a\bar{X} - \bar{Y}$.

De plus, la première condition conduit à $0 = \sum_{i=1}^n x_i(ax_i + b - y_i) = a \sum_{i=1}^n x_i^2 + b \sum_{i=1}^n x_i - \sum_{i=1}^n x_i y_i$. En divisant par n cette relation, on obtient l'égalité

$$a \sum_{i=1}^n \frac{1}{n} x_i^2 + b \sum_{i=1}^n \frac{1}{n} x_i - \sum_{i=1}^n \frac{1}{n} x_i y_i = 0,$$

soit encore $a\mathbb{E}(X^2) + b\mathbb{E}(X) - \mathbb{E}(XY) = 0$. En tenant compte de la valeur de b obtenue plus haut, il vient finalement $a = \frac{\text{Cov}(X; Y)}{\mathbb{V}(X)}$, puisque $\mathbb{V}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$.

Enfin, puisque a et b sont uniques, f présente au plus un extremum. Il est par ailleurs clair que f est minorée, par exemple par la fonction $(a; b) \mapsto (ax_1 + b - y_1)^2$, donc que sa limite lorsque a et b tendent vers l'infini est infinie. Un argument de compacité et de continuité de f permet de conclure à l'existence d'un minimum de f sur $\mathbb{R}^2$, qui ne peut être atteint qu'en ces valeurs de a et b . ▶

Remarque 5 Les réels a et b sont fournies par la calculatrice, en mode statistique (régression linéaire). Voir pour ceci les TP.

Proposition 6 Le point moyen du nuage appartient à la droite de régression de y en x .

Preuve

◀ C'est une conséquence immédiate de la valeur de b obtenue : on avait en effet $b = a\bar{X} - \bar{Y}$, qui s'écrit aussi $\bar{Y} = a\bar{X} + b$, et comme les coordonnées du point moyen sont $G(\bar{X}; \bar{Y})$, G appartient à la droite de régression de y en x . ▶

Remarque 6 Par symétrie du problème, l'équation de la droite de régression de x en y , obtenue pour des distances quadratiques calculées horizontalement, est $x = ay + b$ avec $a = \frac{\text{Cov}(Y; X)}{\mathbb{V}(Y)}$, et $b = \bar{X} - a\bar{Y}$.

2.3.3 Coefficient de corrélation

Il s'agit d'un nombre qui est également fourni par les calculatrices dans les résultats des études de régression linéaire.

Définition 5 Le coefficient de corrélation calculé lors de l'ajustement linéaire de deux séries statistiques X et Y est noté ρ (ou r sur les calculatrices, et qui vaut par définition

$$\rho = \frac{\text{Cov}(X; Y)}{\sqrt{\mathbb{V}(X)\mathbb{V}(Y)}} = \frac{\text{Cov}(X; Y)}{\sigma_X \sigma_Y}.$$

Ce coefficient mesure la qualité de l'ajustement affine réalisé par le calcul des paramètres de la droite de régression. De manière précise, on a le résultat suivant :

Proposition 7 Plus le coefficient de régression est proche de 1 en valeur absolue, meilleur est l'ajustement linéaire.

Lorsque $\rho = \pm 1$, la droite de régression passe par tous les points du nuage, qui sont donc alignés.

Preuve

► Le minimum de la distance quadratique, dans le cas d'une droite de régression de y en x , est donnée par l'image de $\left(\frac{\text{Cov}(X; Y)}{\mathbb{V}(X)}; \mathbb{E}(Y) - \mathbb{E}(X)\frac{\text{Cov}(X; Y)}{\mathbb{V}(X)}\right)$ par la fonction $f : (a; b) \mapsto \sum_{i=1}^n (ax_i + b - y_i)^2$. Ce minimum vaut donc : $M = \sum_{i=1}^n \left(\frac{\text{Cov}(X; Y)}{\mathbb{V}(X)}x_i + \mathbb{E}(Y) - \frac{\mathbb{E}(X)\text{Cov}(X; Y)}{\mathbb{V}(X)} - y_i\right)^2$. Commençons par estimer l'intérieur de la parenthèse :

$$\frac{\text{Cov}(X; Y)}{\mathbb{V}(X)}x_i + \mathbb{E}(Y) - \frac{\mathbb{E}(X)\text{Cov}(X; Y)}{\mathbb{V}(X)} - y_i = \frac{\text{Cov}(X; Y)}{\mathbb{V}(X)}(x_i - \mathbb{E}(X)) - (y_i - \mathbb{E}(Y)).$$

Ainsi,

$$\begin{aligned} M &= \sum_{i=1}^n \left(\frac{\text{Cov}(X; Y)}{\mathbb{V}(X)}(x_i - \mathbb{E}(X)) - (y_i - \mathbb{E}(Y)) \right)^2 \\ &= \sum_{i=1}^n \left(\frac{\text{Cov}(X; Y)^2}{\mathbb{V}(X)^2}(x_i - \mathbb{E}(X))^2 + (y_i - \mathbb{E}(Y))^2 - 2\frac{\text{Cov}(X; Y)}{\mathbb{V}(X)}(x_i - \mathbb{E}(X))(y_i - \mathbb{E}(Y)) \right) \\ &= \frac{\text{Cov}(X; Y)^2}{\mathbb{V}(X)^2} \sum_{i=1}^n (x_i - \mathbb{E}(X))^2 + \sum_{i=1}^n (y_i - \mathbb{E}(Y))^2 - 2\frac{\text{Cov}(X; Y)}{\mathbb{V}(X)} \sum_{i=1}^n (x_i - \mathbb{E}(X))(y_i - \mathbb{E}(Y)) \\ &= n\frac{\text{Cov}(X; Y)^2}{\mathbb{V}(X)^2}V(X) + n\mathbb{V}(Y) - 2n\frac{\text{Cov}(X; Y)}{\mathbb{V}(X)}\text{Cov}(X; Y) \\ &= n\frac{\text{Cov}(X; Y)^2}{\mathbb{V}(X)^2}V(X) + n\mathbb{V}(Y) - 2n\frac{\text{Cov}(X; Y)^2}{\mathbb{V}(X)} \\ &= n\mathbb{V}(Y) - n\frac{\text{Cov}(X; Y)^2}{\mathbb{V}(X)} = n\mathbb{V}(Y)\left(1 - \frac{\text{Cov}(X; Y)^2}{\mathbb{V}(X)\mathbb{V}(Y)}\right) \end{aligned}$$

En posant $\rho = \frac{\text{Cov}(X; Y)}{\sqrt{\mathbb{V}(X)\mathbb{V}(Y)}}$, on a donc

$$M = n\mathbb{V}(Y)(1 - \rho^2).$$

Cette quantité est évidemment minimum (et nulle ...) lorsque $\rho = \pm 1$.

La deuxième affirmation de la proposition découle bien sûr du fait que si $\rho = \pm 1$, alors la somme des distances quadratiques est nulle, donc toutes ces distances sont nulles aussi. ►

Remarque 7 Dans le cas où ρ serait éloigné de ± 1 , il est clair qu'un ajustement linéaire est absolument dénué de pertinence.

Par ailleurs, chaque domaine d'application détermine les valeurs critiques du coefficient de corrélation qui rendent compte d'un ajustement affine acceptable dans le domaine considéré. Enfin, en pratique, avant de calculer une équation de la droite de régression, il serait plus logique de commencer par calculer le coefficient de corrélation pour apprécier la pertinence de la recherche d'une droite de régression (ceci étant dit, comme en pratique la plupart des calculatrices fournissent simultanément une équation et le coefficient en question, cette chronologie n'est pas nécessairement respectée). Un regard critique est néanmoins apprécié ...

3 Autres exemples d'ajustement

Les autres types d'ajustement se déduisent de l'ajustement affine par des changements de variables opportuns. Dans ces domaines, le choix de la nouvelle variable étudiée est déterminant, et l'intuition du statisticien est déterminante pour choisir le "bon" changement de variable.

3.1 Exemple d'ajustement logarithmique

On reprend l'exemple d'introduction concernant la consommation d'une voiture.

3.2 Exemple d'ajustement exponentiel

Décomposition du Carbone 14 et datation

Références