Chapitre 8

Statistiques à deux variables

8.1 Introduction

Dans un chapitre précédent, nous traitions des séries statistiques à 1 variables : il s'agissait d'une suite de nombres $x_1, \ldots, x_p$ (notes d'élèves, nombre de buts marqués, nombres de cigarettes par jour, etc).

Dans ce chapitre nous allons étudier des séries statistiques à deux variables. Cela signifie que l'ensemble des valeurs obtenues sont des points $(x_1; y_1), \ldots, (x_p; y_p)$ du plan.

Exemple 8.1.1. Une grande surface s'intéresse au lien entre ses dépenses publicitaires et son chiffre d'affaires : elle recueille les données suivantes, exprimées en millions d'euros, portant sur cinq périodes ; les dépenses sont notées $x_1, ..., x_5$ et les chiffres d'affaires $y_1, ..., y_5$.

Dépenses publicitaires : x_i	0,5	2,0	2,9	4,5	5,6
Chiffres d'affaires : y_i	35	37	75	92	90

Il est bien entendu possible de placer les points $M(x_i; y_i)$ (pour $i = 1, \ldots, 5$) dans un repère orthonormé. Le nuage de points est « allongé » de manière oblique (cf. figure suivante) et laisse penser qu'il existe une relation entre x_i et y_i

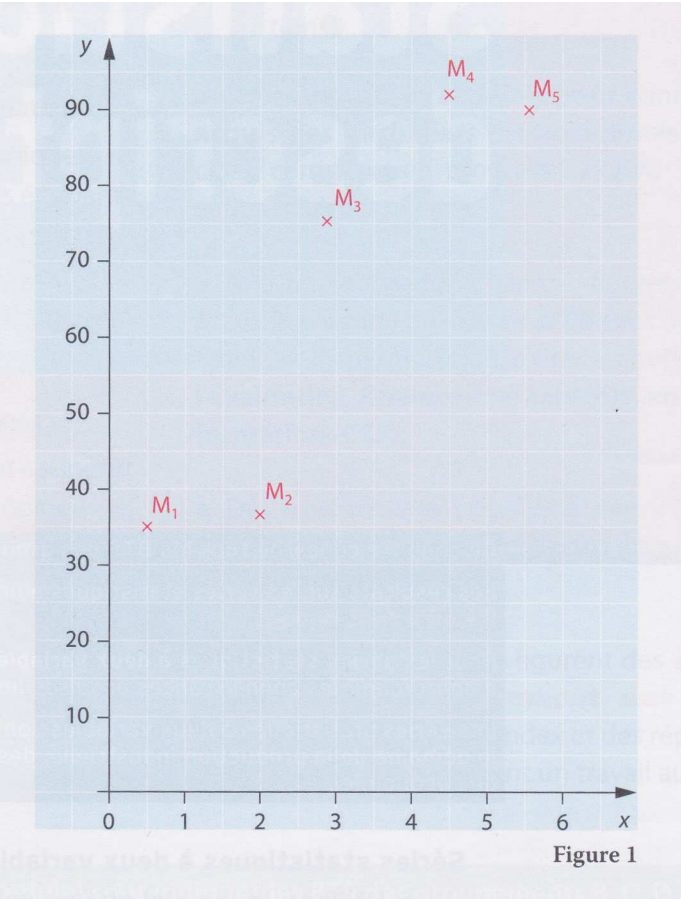


Figure 1

Définition 8.1.1. *Le point moyen d'une série statistique est* $(\bar{x}; \bar{y})$*. Autrement dit, le point dont l'abscisse (resp. l'ordonnée) vaut la moyenne des abscisses (resp. des ordonnées).*

Exemple 8.1.2. Dans l'exemple précédent, le point moyen a pour coordonnées $(3,1; 65,8)$.

8.2 Ajustement affine par la méthode des moindres carrés

8.2.1 Régression de y en x

Au vu du nuage de point, nous allons chercher une relation affine entre les variables x_i et les variables y_i. Nous aimerions donc trouver une droite $\mathcal{D}$ vérifiant les équations

$$y_i = ax_i + b \quad \text{pour tout} \quad i = 1, \ldots, p.$$

Pour trouver des coefficients a et b convenables, nous allons chercher à minimiser la distance (impliquant donc des carrés) entre la droite est les points $M(x_i; y_i)$. C'est pour cette raison que l'on utilise l'appellation « méthode des moindres carrés ». Graphiquement, on a l'impression d'avoir

autant de points au dessus et en dessous de la droite $\mathcal{D}$ (cf. figure ci-dessous).

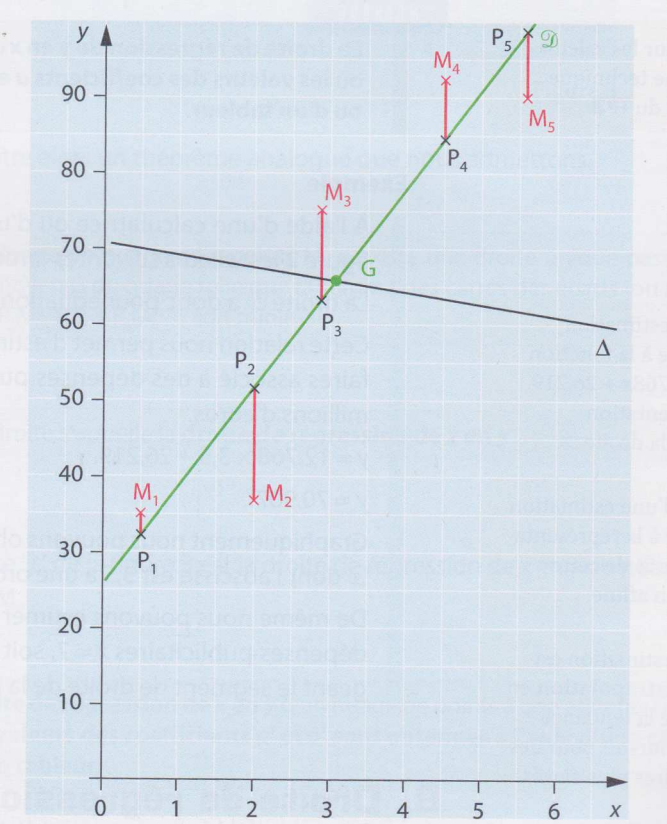


Figure 2

Cette droite s'appelle la droite de régression de y en x car nous avons voulu expliquer la variable y en fonction de la variable x.

En pratique, les coefficients a et b sont déterminés à l'aide de la calculatrice :

1. Compléter deux listes, l'une avec les variables x_i, l'autre avec les variables y_i.
2. Utiliser la fonction « Stats 2 var » de la calculatrice.

Exemple 8.2.1. Dans l'exemple précédent, la calculatrice fournit les valeurs suivantes :

$$a = 12,768 \quad \text{et} \quad b = 26,219$$

La droite $\mathcal{D}$ a donc pour équation $y = 12,768x + 26,219$.

8.2.2 Régression de x en y

Bien entendu, il est possible de faire le contraire et chercher à expliquer la variable x en fonction de la variable y. En reprenant ce qui précède, nous obtenons alors une nouvelle droite $\mathcal{D}'$ d'équation

$$x = a'y + b'$$

avec des coefficients a' et b' obtenues à l'aide d'une calculatrice (en échangeant les colonnes x_i et y_i). Cette nouvelle droite est visible sur la figure suivante :

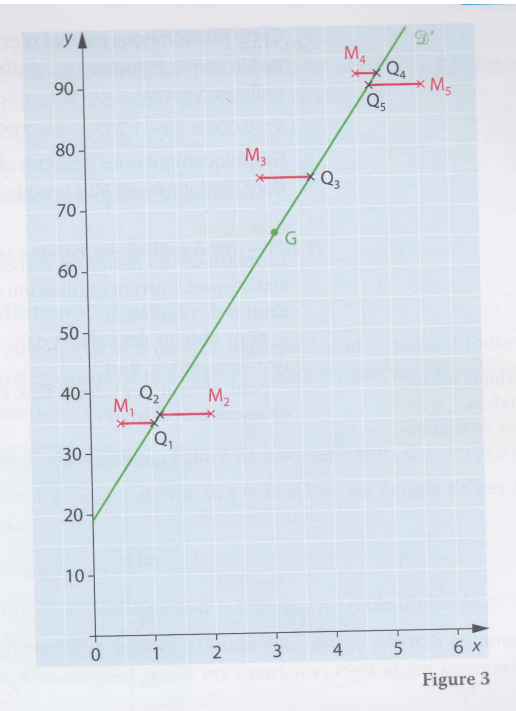


Figure 3

8.2.3 Pourquoi faire ceci ?

L'intérêt derrière ces calculs et d'obtenir une relation ($y = ax+b$ ou $x = a'y+b'$) permettant de prédire/estimer ce qui pourrait se produire à l'avenir. Ce point sera plus clair grâce aux exercices.

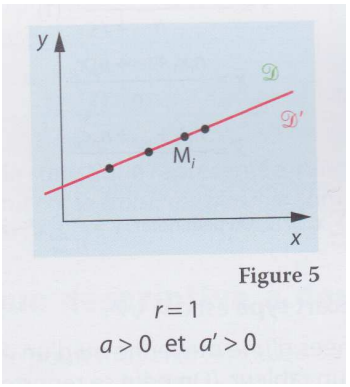
8.3 Qualité de l'ajustement

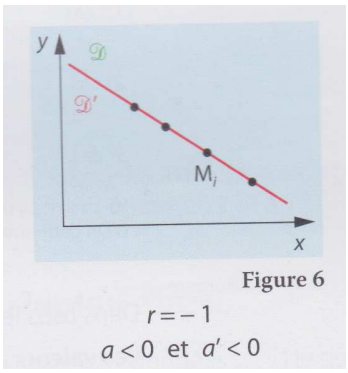
Lorsque nous avons des nuages de points de forme « allongés », il est naturel de chercher à obtenir une relation affine entre les variables. Il est ensuite naturel de s'interroger sur la qualité de l'ajustement qui a été fait (via les droites $\mathcal{D}$ ou $\mathcal{D}'$). Pour cela, on utilise un coefficient de corrélation, noté r.

Ce coefficient est obtenu à l'aide de la calculatrice et vérifie les propriétés suivantes :
- $r^2 = aa'$
- $r \in [-1;1]$

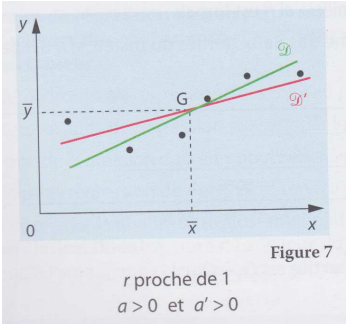
Pour mieux appréhender ce que signifie ce nombre voici quelques commentaires :

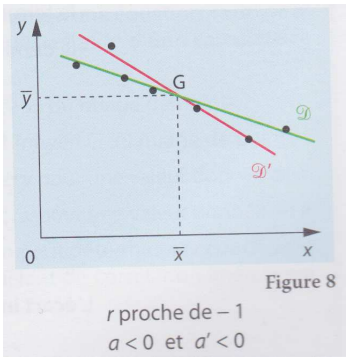
1. Lorsque $r = 1$ ou $r = -1$, cela signifie que les deux droites $\mathcal{D}$ ou $\mathcal{D}'$ possèdent le même coefficient directeur, l'alignement des points est parfait (cf. figure ci-dessous).





2. Plus r est proche de 1 ou de -1, plus l'ajustement affine est de qualité (cf. figure ci-dessous).





Remarque. Attention à ne pas confondre corrélation et liaison de cause à effets. Par exemple, les variables x_i et y_i peuvent mesurer deux effets d'une même cause.

Comme nous le verrons en exercice, la forme du nuage de points peut indiquer d'un ajustement affine n'est pas pertinent (notamment si la forme du nuage de points ressemble à une parabole). Il faudra alors trouver une autre manière de procéder (la plupart du temps, l'exercice nous proposera un changement de variables adéquat).