III/ DESCRIPTION BIDIMENSIONNELLE ET MESURES DE LIAISON ENTRE VARIABLES

Après les descriptions unidimensionnelles on étudie généralement les liaisons entre les variables observées : c'est ce qu'on appelle communément l'étude des corrélations. Les méthodes et les indices de dépendance varient selon la nature des variables étudiées (qualitative, ordinale, numérique).

III.1/ DEUX VARIABLES QUALITATIVES

III.1.1/ TABLEAU DE CONTINGENCE

Une façon habituelle pour représenter ces données consiste à générer un *tableau* représentant la distribution conjointe de deux variables qualitatives observées chez les mêmes individus.

Soit I le nombre de modalités distinctes de la variable X

On note dans ce cas:

x_i : la $i^{ième}$ modalité de la variable X

avec :

$1 \leq i \leq I$

ATTENTION :

Dans ce cas, x_i ne représente plus la valeur de la variable X pour le ième individu mais la $i^{ième}$ modalité de la variable X !!!!

Soit J le nombre de modalités distinctes de la variable Y

On note :

y_j : la $j^{ième}$ modalité de la variable Y

avec :

$1 \leq j \leq J$

Le tableau représente l'ensemble des I*J couples de valeurs possibles d'une part, les effectifs n_{ij} (ou fréquences relatives f_{ij}) correspondant(es) d'autre part. On appelle ce tableau un :

TABLEAU DE CONTINGENCE

fréquences absolues ou effectifs

Y / X	y_1 y_2 ..... y_j .....y_J	Totaux
x_1	n_{11} n_{12} ..... $.n_{1j}$ .....n_{1J}	$n_{1.}$
x_2	n_{21} n_{22} .....n_{2j} .....n_{2J}	$n_{2.}$
.		
x_i	n_{i1} n_{i2} ..... n_{ij} .....n_{iJ}	$n_{i.}$
.		
x_I	n_{I1} n_{I2} .....n_{Ij} .....n_{IJ}	$n_{I.}$
Totaux	$n_{.1}$ $n_{.2}$ .....$n_{.j}$ .....$n_{.J}$	n

fréquences relatives

Y / X	y_1 y_2 .....y_j .....y_J	Totaux
x_1	f_{11} f_{12} .....f_{1j} .....f_{1J}	$f_{1.}$
x_2	f_{21} f_{22} .....f_{2j} .....f_{2J}	$f_{2.}$
.		
x_i	f_{i1} f_{i2} ..... . f_{ij} .....f_{iJ}	$f_{i.}$
.		
x_I	f_{I1} f_{I2} ..... $.f_{Ij}$ .....f_{IJ}	$f_{I.}$
Totaux	$f_{.1}$ $f_{.2}$ ..... $.f_{.j}$ .....$f_{.J}$	1

Rappel : $f_{ij} = n_{ij}/n$

$$n_{i.} = \sum_{j=1}^{J} n_{ij} \quad i = 1,\ldots..,\ I \qquad n_{.j} = \sum_{i=1}^{I} n_{ij} \quad j = 1,\ldots..,\ J$$

$$\sum_{i=1}^{I}\sum_{j=1}^{J} nij = \sum_{i=1}^{I} ni. = \sum_{j=1}^{J} n.j = n$$

Exemple de tableau de contingence :

Mobilité professionnelle entre deux générations

La population est constituée de 720 familles

Profession du père	Profession du fils					
	Cadre	Ouvrier	Agriculteur	Vente	Autre	Totaux
Cadre	52	7	2	23	8	92
Ouvrier	12	123	11	74	21	241
Agriculteur	31	61	25	29	17	163
Vente	46	14	8	79	11	158
Autre	13	19	22	9	3	66
Totaux	154	224	68	214	60	720

III.1.2/ DISTRIBUTIONS MARGINALES

Les $n_{i.}$ et les $n_{.j}$ s'appellent respectivement distribution marginale en ligne et distribution marginale en colonne.

Dans l'exemple précédent :

Distribution marginale en ligne :

Père cadre	92
Père ouvrier	241
Père agriculteur	163
Père vente	158
Père autre	66

Il y a 163 familles dans la population dont le père est agriculteur

Distribution marginale en colonne:

Fils cadre	154
Fils ouvrier	224
Fils agriculteur	68
Fils vente	214
Fils autre	60

Il y a 224 familles dont le fils est ouvrier

III.1.3/ DISTRIBUTIONS CONDITIONNELLES

1/ Distribution conditionnelle de Y par rapport à la modalité x_i de X :

Soit:

Y X	y_1 y_2 ..... y_j .....y_J	Totaux
x_1	n_{11} n_{12} .....n_{1j} .....n_{1J}	$n_{1.}$
x_2	n_{21} n_{22} .....n_{2j} .....n_{2J}	$n_{2.}$
.		
x_i	**n_{i1} n_{i2} ..... n_{ij} .....n_{iJ}**	$n_{i.}$
.		
x_I	n_{I1} n_{I2} ..... .n_{Ij} .....n_{IJ}	$n_{I.}$
Totaux	$n_{.1}$ $n_{.2}$ ..... .$n_{.j}$ .....$n_{.J}$	n

Les effectifs de la $i^{ième}$ ligne (en gras) définissent la distribution selon la variable Y des $n_{i.}$ individus.

Cette distribution est la **distribution conditionnelle de Y si $X=x_{i.}$**.

Il y a I distributions conditionnelles de Y pour $X= x_i$ i=1,.....,I.

2/ Distributions conditionnelles de X par rapport aux modalités de Y :

Par symétrie, on définit ces distributions de la même manière.

Les effectifs de la $j^{ième}$ colonne définissent la distribution selon la variable X des $n_{.j}$ individus.

Cette distribution est la **distribution conditionnelle de X si $Y=y_j$**

Il y a J distributions conditionnelles de X pour $Y= y_j$ j=1,.....,J.

III.1.4/ PROFIL LIGNE –PROFIL COLONNE

Deux lectures différentes d'un même tableau de contingence sont possibles selon que l'on privilégie l'une ou l'autre des deux variables :

Lecture en ligne ou lecture en colonne.

On appelle **Tableau des profils-lignes** le tableau des fréquences (ici implicitement relative) conditionnelles $\frac{n_{ij}}{n_{i.}}$ (la somme de chaque ligne est ramenée à 1)

Tableau des profils-lignes :

Profession du père	Profession du fils					
	Cadre	Ouvrier	Agriculteur	Vente	Autre	Totaux
Cadre	52/92 = 0.56	7/92 = 0.08	2/92 = 0.02	23/92 = 0.25	8/92 = 0.09	92/92
Ouvrier	12/241	123/241	11/241	74/241	21/241	241/241
Agriculteur	31/163	61/163	25/163	29/163	17/163	163/163
Vente	46/158	14/158	8/158	79/158	11/158	158/158
Autre	13/66	19/66	22/66	9/66	3/66	66/66

Parmi les familles dont le père est cadre, 56% de ces familles ont le fils cadre, 8% ont le fils ouvrier, 2% ont le fils agriculteur, 25% ont le fils dans la vente et 9% ont le fils dans une autre profession.

On appelle **Tableau des profils-colonnes** le tableau des fréquences (ici implicitement relative) conditionnelles $\frac{n_{ij}}{n_{.j}}$ (la somme de chaque colonne est ramenée à 1)

Tableau des profils-colonnes :

Profession du père	Profession du fils				
	Cadre	Ouvrier	Agriculteur	Vente	Autre
Cadre	52/154 = 0.34	7/224	2/68	23/214	8/60
Ouvrier	12/154 = 0.08	123/224	11/68	74/214	21/60
Agriculteur	31/154 = 0.20	61/224	25/68	29/214	17/60
Vente	46/154 = 0.30	14/224	8/68	79/214	11/60
Autre	13/154 = 0.08	19/224	22/68	9/214	3/60
Totaux	154/154	224/224	68/68	214/214	60/60

Parmi les familles dont le fils est cadre, 34% des ces familles ont le père cadre, 8% ont le père ouvrier, 20% ont el père ouvrier, 30% ont le père dans la vente et 8% ont le père dans une autre profession.

III.1.5/ L'ECART A L'INDEPENDANCE

Lorsque tous les profils-lignes sont identiques on peut parler d'indépendance entre les X et Y puisque la connaissance de X ne change pas les distributions conditionnelles de Y. Il s'ensuit d'ailleurs que tous les profils-colonnes sont également identiques.

On doit donc avoir :

$$\frac{n1j}{n1.} = \frac{n2j}{n2.} = \ldots = \frac{nij}{ni.} = \ldots . \frac{nIj}{nI.} \quad \forall \textit{ la colonne } j$$

Ce qui entraîne, par sommation des numérateurs et dénominateurs:

$$\frac{nij}{ni.} = \frac{n.j}{n} \quad \forall i, j$$

L'indépendance empirique se traduit donc par :

$$nij = \frac{ni.\, n.j}{n} \quad \forall i, j$$

Démonstration :

III.1.6/ LE X^2 OU d^2 D' ECART A L'INDEPENDANCE ET LES AUTRES MESURES ASSOCIEES

On adopte généralement la mesure suivante de liaison X^2 (notée aussi d^2)

$$X^2 = d^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{\left(n_{ij} - \frac{n_{i.} n_{.j}}{n} \right)^2}{\frac{n_{i.} n_{.j}}{n}}$$

On voit que X^2 est nul dans le cas d'indépendance.

Quelle est sa borne supérieure ?

On montre que :

$$\frac{d^2}{n} \leq \inf(I-1, J-1)$$

Démonstration :

AUTRES MESURES ASSOCIEES :

Divers coefficients liés au d^2 ont été proposés pour obtenir une mesure comprise entre 0 (indépendance) et 1 (liaison fonctionnelle). Citons :

- Le coefficient de contingence de K. Pearson :

$$C = \left(\frac{d^2}{n + d^2} \right)^{1/2}$$

- Le coefficient de Tschuprow :

$$T = \left(\frac{d^2}{n\sqrt{(I-1)(J-1)}} \right)^{1/2}$$

- Le coefficient de Cramer :

$$V = \left(\frac{d^2}{n \inf\{(I-1);(J-1)\}} \right)^{1/2}$$

- $\frac{d^2}{n}$ est usuellement noté φ^2

III.1.7/ CONTRIBUTION AU d^2

La construction du tableau $\frac{n_{i.} n_{.j}}{n}$ (tableau d'indépendance) et sa comparaison avec le tableau des n_{ij} est en général très instructive : en particulier le calcul pour chaque case du terme :

$$\frac{\left(n_{ij} - \frac{n_{i.} n_{.j}}{n} \right)^2}{\frac{n_{i.} n_{.j}}{n}} * \frac{1}{d^2}$$

Appelé **CONTRIBUTION AU d^2** permet de mettre en évidence les associations significatives entre catégories des deux variables.

Le signe de la différence $n_{ij} - \frac{n_{i.} n_{.j}}{n}$ indique alors s'il y a association positive ou négative entre les catégories i de X et j de Y.

Un tel calcul devrait être systématiquement associé à chaque calcul de d^2.

III.1.8/ AUTRES MESURES DE DEPENDANCE

Les indices dérivés du d^2 sont loin d'être les seules mesures de dépendance utilisables, elles ont d'ailleurs été souvent critiquées. La littérature statistique abonde en la matière et le problème est d'ailleurs celui du trop grand nombre d'indices proposés.

On se reportera utilement aux ouvrages de Goodman et Kruskal et de Marcotorcino.

III.1.9/ MISE EN PRATIQUE

Reprendre l'exemple précédent et calculer d^2 et la contribution de chaque cas au d^2.

Calculer également les mesures associées. Conclure.

III.2/ DEUX VARIABLES QUANTITATIVES

Deux variables quantitatives X et Y sont observées chez n sujets d'une population.
Par exemple, X le poids et Y la taille.
Les deux séries de données peuvent être représentées conjointement sous la forme d'une série statistique double $\{(x_1,y_1), (x_2,y_2),.... (x_i,y_i),..... (x_n,y_n)\}$.
La notation (x_i,y_i) désigne ici le couple de valeurs obtenues en observant le poids x_i et la taille y_i chez le ième individu.
x_i : valeur de X pour le ième individu
y_i : valeur de Y pour le ième individu

III.2.1/ DIAGRAMME DE DISPERSION OU NUAGE DE POINTS

Pour étudier la force de liaison entre deux variables quantitatives X et Y observées sur n individus, la méthode graphique de référence est le ***diagramme de dispersion*** (ou nuage de points).

Il fait correspondre à chaque couple (x_i , y_i) un point d'abscisse x_i et d'ordonnée y_i dans un repère cartésien.

Il permet de visualiser si les points sont groupés autour d'une droite ou d'une courbe, si ils sont groupés en plusieurs ensembles isolés les uns des autres, ou si ils sont complètement dispersés.

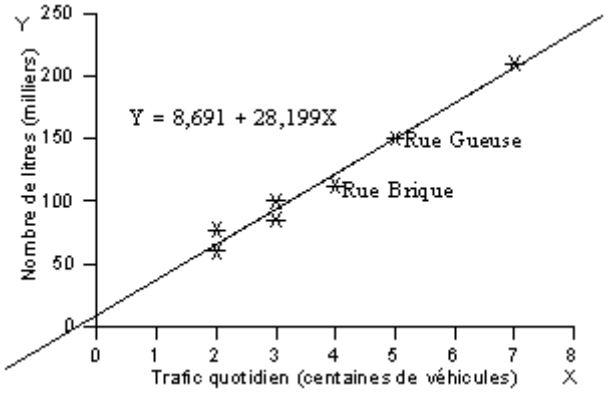
Quelle variable sur quel axe ?

Lorsqu'il n'existe aucun argument *a priori* pour penser qu'une variable dépend de l'autre (on parle alors d'analyse de corrélation ou d'association), les deux variables peuvent se situer sur l'un ou l'autre des 2 axes, de manière indifférente.

Lorsqu'il existe en revanche des arguments pour considérer qu'une des variables dépend de l'autre, alors la variable expliquée (par convention Y) est portée sur l'axe des ordonnées, et l'autre variable dite " explicative " (par convention X) est portée sur l'axe des abscisses.

Exemple (Pompaluile):
On cherche à déterminer si, dans la ville Pompaluile, il existe une relation (une liaison) entre le nombre de véhicules qui passent devant une station d'essence et le nombre de litres d'essence vendus (moyennes par jour, sur un an). Voici les résultats:

Emplacement	**X = Nombre de véhicules (centaines)**	**Y = Nombre de litres (milliers)**
Rue Barbe	3	100
Rue Brique	4	112
Rue Gueuse	5	150
Avenue Anse	7	210
Rue Elle	2	60
Chemin Sire	3	85
Chemin Soeur	2	77

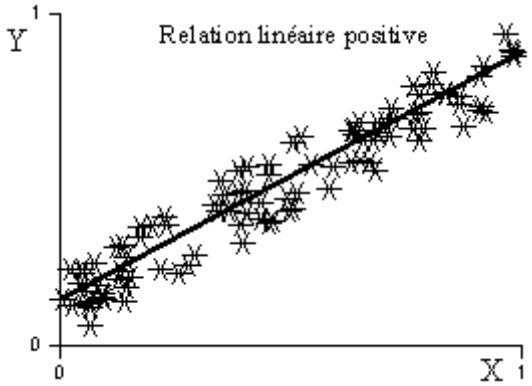


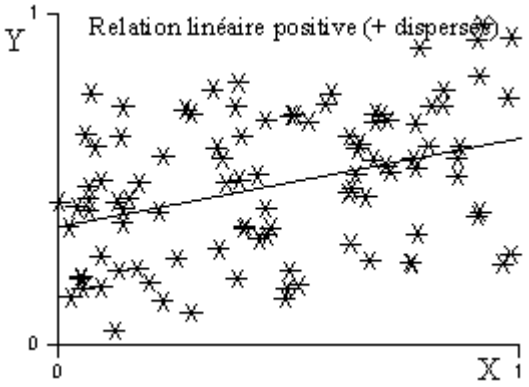
le diagramme de dispersion correspondant

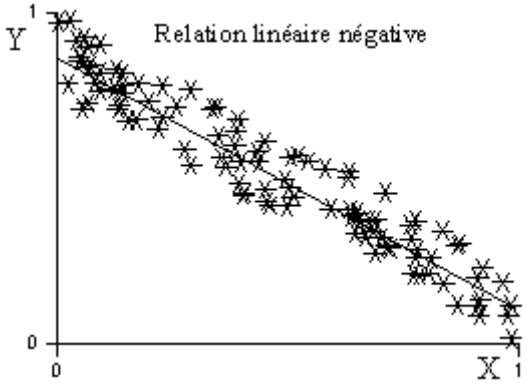
Dans ce cas, on essaie bien d' »expliquer » le nombre de litres vendus (Y : variable expliquée) en fonction du trafic quotidien (X : variable explicative) !!

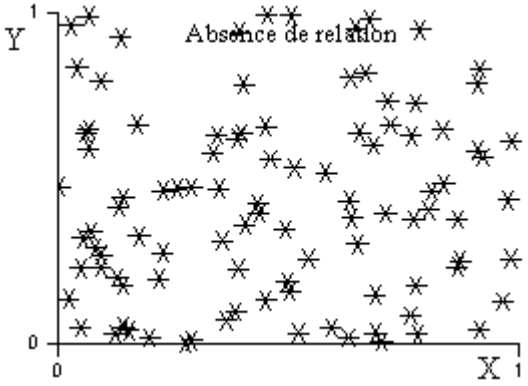
III.2.2/ DIFFERENTS TYPES DE RELATION

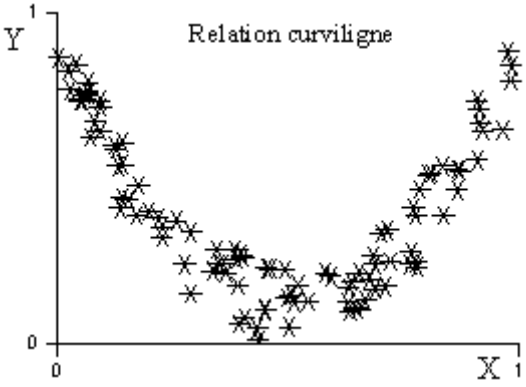
La relation entre deux variables quelconques peut être de différents types
C'est ce qu'illustrent les diagrammes suivants:

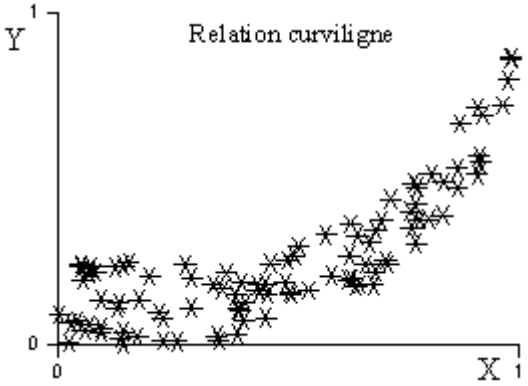

Relation linéaire positive
Y
1
0
0
X
1


Relation linéaire positive (+ dispersé)
Y
1
0
0
X
1


Relation linéaire négative
Y
1
0
0
X
1


Absence de relation
Y
1
0
0
X
1


Relation curviligne
Y
1
0
0
X
1


Relation curviligne
Y
1
0
0
X
1

III.2.3/ LA DROITE DES MOINDRES CARRES

Dans les cas où le diagramme de dispersion montre l'existence d'une relation linéaire, on désire déterminer la droite qui décrira le «mieux» cette relation Cependant, le choix de cette droite dépend d'un critère qu'il faudra fixer.

Le critère mathématique habituel est celui des *moindres carrés.*

Soit I le point de coordonnées (x_i,y_i). Il y aura en effet une différence ou erreur entre la valeur de y_i et la valeur qui lui « correspond » sur la droite.

Cette erreur est indiquée par e_i sur la figure suivante :

Graphique :

Les valeurs que peuvent prendre les e_i (pour l'ensemble des n points : $1 \leq i \leq n$) peuvent être positives, négatives ou simplement, nulles.

On peut mesurer l'efficacité d'un ajustement en effectuant la somme des carrés des erreurs de l'ensemble des données.

Si cette somme est petite, l'ajustement linéaire est considéré comme étant bon. À l'inverse, si elle est grande, l'ajustement linéaire est mauvais.

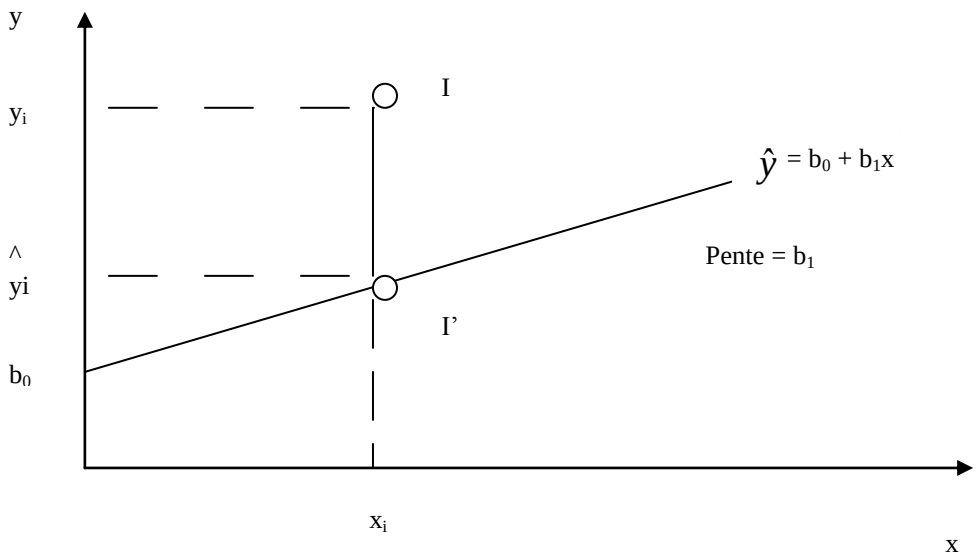
On dit donc que parmi toutes les droites possibles qui s'approchent d'un ensemble de données, celle révélant le meilleur ajustement est celle qui correspond à la propriété suivante:

$$\text{minimum} = e_1^2 + e_2^2 + .. e_i^2 + ... + e_n^2$$

On appelle droite des moindres carrés la droite qui vérifie cette propriété

En <u>formule</u>, la droite des moindres carrés (comme toute droite) sera donnée par:

$$\hat{y} = \mathbf{b_0 + b_1 x}$$



où

I le point de coordonnées (x_i, y_i)

$\hat{y}_i$: valeur de y_i « estimée » par la droite

Ainsi le point $I'(x_i, \hat{y}_i)$ appartient à la droite des moindres carrés

b_0 = l'ordonnée à l'origine, i.e. la valeur de y lorsque x = 0

b_1 = la pente, i.e. la variation de y pour une variation d'une unité de x

Remarque sur la pente d'une droite :

Considérons deux points quelconques de coordonnées: (x_1, y_1) et (x_2, y_2) qui se retrouvent sur une droite d'équation : $y = b_0 + b_1x$

Donc:

$y_1 = b_0 + b_1x_1$

$y_2 = b_0 + b_1x_2$

À partir de ces équations, il devient possible de « découvrir » la pente b_1.

Si on soustrait les deux équations, on obtient comme valeur de b_1 (la pente) :

$(y_2 - y_1) / (x_2 - x_1)$

Ainsi :
b_1 = la pente, i.e. la variation de y pour une variation d'une unité de x

cqfd !!

1/ Propriétés de la droite de régression

Par définition, la somme

$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$ est minimale

$e_i = y_i - \hat{y}_i$

De plus:

$\sum_{i=1}^{n} (y_i - \hat{y}_i) = \sum_{i=1}^{n} e_i = 0$

i.e. que les écarts e_i «positifs» sont compensés par des écarts e_i «négatifs» équivalents

2/ Calcul des coefficients b_0 et b_1

La *méthode des moindres carrés* consiste à chercher les valeurs des paramètres b_0 et b_1 qui rendent minimale la *somme des carrés des écarts résiduelle* (*SS_r : sum of squared residuals*)

entre les valeurs observées y_i et les valeurs estimées $\hat{y}_i$:

i.e:

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2 = SS_r$$

où n est le nombre de points et :

$$\hat{y}_i = b_0 + b_1 x_i$$

d'où :

$$SSr = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 = \Phi(b_0 , b_1)$$ → Recherche d'extrema d'une fonction à 2 var.

<u>Rappel</u> *: Optimisation d'une fonction à 2 variables f(x,y)*

Conditions nécessaires d'optimalité:

(1) $\partial\Phi(b_0 , b_1)/ \partial b_0 = -2 \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) = 0$

(2) $\partial\Phi(b_0 , b_1)/ \partial b_1 = -2 \sum_{i=1}^{n} x_i(y_i - b_0 - b_1 x_i) = 0$

Soit:

(1) $b_0 n + b_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$

(2) $b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$

Ce système d'équations admet pour solutions :

$$b_1 = \frac{n\Sigma x_i y_i - \Sigma x_i \ \Sigma y_i}{n\Sigma x_i^2 - (\Sigma x_i)^2} \quad \rightarrow \quad [(1)^* \ \Sigma xi - (2)^* n]$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

On démontre facilement (à faire en exercice) que :

CE SONT LES FORMULES A RETENIR !!

$$b_1 = \frac{s_{xy}}{s^2_x} \quad \text{et} \quad b_0 = \bar{y} - b_1 \bar{x}$$

où s^2_x est la variance empirique de la variable X (déjà évoquée) et s_{xy} la covariance empirique des variables X et Y (nouveau !!)

Formules:

$$s^2_x = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

et

$$s_{xy} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

Conditions suffisantes d'optimalité (à démontrer en TD):

…….

Remarque:

La droite des moindres carrés passe par le point $(\bar{x}, \bar{y})$

Exemple précédent (Pompaluile):

Emplacement	**X**	**Y**	**XY**	$\mathbf{X^2}$	$\mathbf{Y^2}$
Rue Barbe	3	100	300	9	10 000
Rue Brique	4	112	448	16	12 544
Rue Gueuse	5	150	750	25	22 500
Avenue Anse	7	210	1 470	49	44 100
Rue Elle	2	60	120	4	3 600
Chemin Sire	3	85	255	9	7 225
Chemin Soeur	2	77	154	4	5 929
	26	**794**	**3 497**	**116**	**105 898**

On trouve les valeurs suivantes :

$b_1 = 28.20$ et $b_0 = 8.69$

Vérifiez ce résultat « à la main » puis retrouvez directement ce résultat avec votre calculatrice !!

III.2.4/ COEFFICIENT DE DETERMINATION EMPIRIQUE

Objectif:
Evaluer le degré d'association entre les deux variables i.e. juger de la qualité de l'ajustement des points par la droite de régression

Soit :

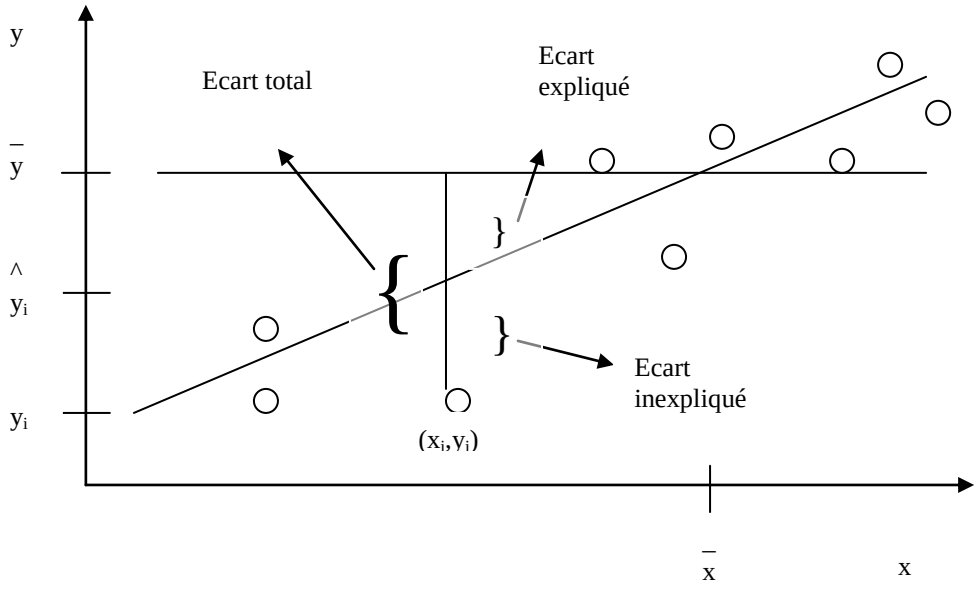
$\bar{y}$ = la moyenne des valeurs de la variable à expliquer Y
y_i = la valeur de Y pour un individu i
$\hat{y}_i$ = la valeur de y_i estimée par la droite des moindres carrés

alors :

$y_i - \bar{y}$ = l'écart total
$\hat{y}_i - \bar{y}$ = l'écart expliqué « par la variable explicative X »
$e_i = y_i - \hat{y}_i$ = l'écart inexpliqué « par la variable explicative X »

Pour tout y_i, on a la relation suivante :

Ecart total = Ecart expliqué + Ecart inexpliqué
$y_i - \bar{y}$ = $\hat{y}_i - \bar{y}$ + $y_i - \hat{y}_i$



En additionnant tous les écarts sur l'ensemble des points (somme sur les i), on a
(la démonstration sera faite géométriquement ultérieurement)

Formule de décomposition de la variance :

$$\sum_i (y_i - \bar{y})^2 = \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i (y_i - \hat{y}_i)^2$$

Variation Totale (SST)	=	Variation expliquée (SSE)	+	Variation inexpliquée (SSR)

Définition :
Le coefficient de détermination empirique, notée r^2 est une mesure de la proportion de la variation de la variable Y qui s' »explique » les variations de la variable X.

En formule,

$$r^2 = \frac{\text{Variation expliquée}}{\text{Variation totale}} = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} = \frac{SSE}{SST}$$

*Exemple (Pompaluile)***:**

x_i	y_i	$\hat{y}_i$	$(\hat{y}_i - \bar{y})$	$(\hat{y}_i - \bar{y})^2$	$(y_i - \bar{y})$	$(y_i - \bar{y})^2$
3	100	93,29	-20,14	405,62	-13,43	180,36
4	112	121,49	8,06	64,96	-1,43	2,04
5	150	149,69	36,26	1 314,79	36,57	1 337,36
7	210	206,09	92,66	8 585,88	96,57	9 325,76
2	60	65,09	-48,34	2 336,76	-53,43	2 854,76
3	85	93,29	-20,14	405,62	-28,43	808,26
2	77	65,09	-48,34	2 336,76	-36,43	1 327,14
26	**794**	**794,03**	**0**	**15 450,39**	**0**	**15 835,68**

D'où

$$r^2 = \frac{\text{Variation expliquée}}{\text{Variation totale}} = \frac{\Sigma(\hat{y}_i - \bar{y})^2}{\Sigma(y_i - \bar{y})^2} = \frac{15450.39}{15835.68} = 0.976 \ (97.6\ \%)$$

III.2.5/ COEFFICIENT DE CORRÉLATION LINEAIRE EMPIRIQUE

Définition: le coefficient de corrélation linéaire empirique (de Pearson), noté r, est tout simplement la racine carré du coefficient de détermination; son signe (±) donne le sens de la relation
En formule,
On peut calculer:

$$r = \pm\sqrt{r^2}$$

en décidant du signe selon le signe de la pente (b_1)

On note aussi que:

$$r = \frac{S_{XY}}{S_X * S_Y}$$

On a:

$-1 \le r \le 1 \quad et \quad |r| = 1$ est équivalent à l'existence d'une relation linéaire exacte entre les deux variables.

Exemple (Pompaluile):
Avec les données de cet exemple, on trouve
r = 0.988

III.2.6/ INTERPRETATION

Ne pas oublier que le coefficient de détermination donne un pourcentage de variation (ou variance) de la variable dépendante «expliquée» par la présence de la variable indépendante.
Cette notion de «pourcentage de variance expliquée» est fondamentale...
et reviendra dans nombre d'analyses subséquentes malheureusement, en pratique, beaucoup de gens l'oublient

Plus la valeur de *r* se rapproche de ±1 ,
plus la relation linéaire est forte !!!!!

et

Plus la valeur de *r* est voisine de 0,
plus la relation linéaire est faible !!!!!

Rappelons que la non corrélation linéaire n'est pas nécessairement l'indépendance.

III.2.7/ INTERPRETATION GEOMETRIQUE

III.3/ DEUX VARIABLES ORDINALES

Il arrive souvent de ne disposer que d'un ordre sur un ensemble d'individus et non de valeurs numériques d'une variable mesurable : soit parce qu'on ne dispose que de données du type classement (ordre de préférence, classement A, B, C, D, E), ou bien parce que les valeurs numériques d'une variable n'ont que peu de sens et n'importent que par leur ordre (notes d'une copie de français : avoir 12 ne signifie pas valoir deux fois plus que celui qui a 6).

A chaque individu de 1 à n on associe son rang selon une variable (un rang varie de 1 à n). Etudier la liaison entre deux variables revient donc à comparer les classements issus de ces deux variables :

Objet	1	2	3	….	….	….	n
Rang n°1	R1	R2	R3	….	….	….	Rn
Rang n°2	S1	S2	S3	….	….	….	Sn

Les Ri et Si sont des permutations différentes des n premiers entiers.

Le psychologue Spearman a proposé de calculer le coefficient de corrélation sur les rangs :

COEFFICIENT DE CORRELATION DES RANGS DE SPEARMAN

$$r_S = \frac{S_{RS}}{S_R * S_S}$$

Le fait que les rangs soient des permutations de [1,….n] simplifie les calculs et l'on a :

$$\bar{R} = \bar{S} = \frac{n+1}{2}$$

$$s_R^2 = s_S^2 = \frac{n^2-1}{12}$$

$$r_S = \frac{\frac{1}{n}\sum_{i=1}^{n} R_i S_i - \left(\frac{n+1}{2}\right)^2}{\frac{n^2-1}{12}}$$

Si on pose:

$d_i = R_i - S_i$ différence des rangs d'un même objet selon les deux classements, on trouve:

$$r_S = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n\left(n^2-1\right)} \quad : \textit{formule plus simple}$$

Démonstration:

La définition de r_s comme coefficient de corrélation linéaire sur des rangs nous indique que :

$r_s = 1$ → les deux classements sont identiques
$r_s = -1$ → les deux classements sont inverses l'un de l'autre
$r_s = 0$ → les deux classements sont indépendants
Exemple :

10 marques de jus d'orange ont été classés par ordre de préférence par deux gastronomes :

Ri	1	2	3	4	5	6	7	8	9	10
Si	3	1	4	2	6	5	9	8	10	7

Calculer le coefficient de corrélation de Spearman. Que peut-on en conclure ?

III.4/ UNE VARIABLE QUANTITATIVE Y ET UNE VARIABLE QUALITATIVE X

Exemple : On veut étudier la liaison entre le salaire Y et la catégorie socio-professionnelle X d'un ensemble d'individus.

Si X a k groupes (modalités) on notera n1, n2, …nk les effectifs observés et $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_k$ les moyennes de Y pour chaque groupe (il est indispensable qu'au moins un des ni soit supérieur à 1) et $\bar{y}$ la moyenne totale.

On note e^2, le rapport de corrélation empirique :

$$e^2 = \frac{\frac{1}{n}\sum_{c=1}^{k} n_c\left(\bar{y}_c - \bar{y}\right)^2}{s_Y^2} = \frac{\text{Variance inter-groupe}}{\text{Variance totale}}$$

$e^2 = 0$ si $\bar{y}_1 = \bar{y}_2 = \ldots = \bar{y}_k$ d'où absence de dépendance en moyenne.
$e^2 = 1$ si tous les individus d'un groupe de X ont même valeur de Y et ceci pour chaque groupe car :

Variance totale = Variance inter-groupe + Variance intra-groupe

$$S_Y^2 = \frac{1}{n}\sum_{c=1}^{k} n_c\left(\bar{y}_c - \bar{y}\right)^2 + \frac{1}{n}\sum_{c=1}^{k} n_c S_c^2$$

Où les S_c^2 sont les variances empiriques de Y à l'intérieur de chaque groupe c de X:

1/ $\frac{1}{n}\sum_{c=1}^{k} n_c\left(\bar{y}_c - \bar{y}\right)^2$ est appelée variance inter-groupe

2/ $\frac{1}{n}\sum_{c=1}^{k} n_c s_c^2 = \frac{1}{n}\sum_{c=1}^{k} n_c\left(\frac{1}{n_c}\sum_{i=1}^{nc}(y_{ic} - \bar{y}_c)^2\right) = \frac{1}{n}\sum_{c=1}^{k}\sum_{i=1}^{nc}(y_{ic} - \bar{y}_c)^2$

est appelée variance intra-groupe.

3/ $s_Y^2 = \frac{1}{n}\sum_{c=1}^{k}\sum_{i=1}^{nc}(y_{ic} - \bar{y})^2$ est appelée la variance totale

On remarquera que si on attribue à chaque groupe c de X une valeur numérique égale à

$\bar{y}_c$ ce qui revient à transformer X en une variable quantitative X' à k valeurs, e^2 est alors égal au coefficient de détermination empirique r^2 entre Y et X'.

Exemple :

Avec un litre d'essence super par voiture, trois voitures de marque différente (A, B et C) ont été conduites dans des conditions essentiellement identiques. Cet essai a été répété cinq fois (avec des voitures différentes à chaque essai) et le nombre de kilomètres parcourus a été retenu dans le tableau suivant:

	Marque A	Marque B	Marque
Essai 1	9.7	9.0	9.6
Essai 2	9.5	9.2	9.6
Essai 3	9.5	9.8	9.9
Essai 4	9.3	9.3	9.8
Essai 5	9.8	9.4	9.7

Définir les variables étudiées X et Y.
Calculez le rapport de corrélation empirique entre les variables X et Y.

III.5/ PLUSIEURS VARIABLES QUANTITATIVES

III.5.1/ DEFINITION DES MATRICES X, V et R

Lorsqu'on observe p variables quantitatives sur n individus on se trouve en présence d'un tableau (ou matrice) X à n lignes et p colonnes.

$$X = \begin{bmatrix} \ldots\ldots\ldots\ldots \\ \ldots\ldots\ x_i^j \ldots\ldots \\ \ldots\ldots\ldots\ldots \end{bmatrix}$$

x_i^j est la valeur prise par la variable n°j sur le i^{ième} individu.

On note V la matrice de variance covariance empirique des p variables

$$V = \begin{bmatrix} s_1^2 & s12 \ldots\ldots s1p \\ & s_2^2 \ldots\ldots s2p \\ & \\ & \\ \ldots\ldots\ldots\ldots & s_p^2 \end{bmatrix}$$

Définir les éléments de V

La matrice regroupant tous les coefficients de corrélation linéaire empirique entre les p variables prises 2 à 2 est notée R.

Ecrire R :

$$R = \begin{bmatrix} \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \\ \\ \\ \\ \\ \\ \\ \end{bmatrix}$$

Définir les éléments de la matrice R

III.5.2/ COEFFICIENT DE CORRELATION LINEAIRE MULTIPLE

Soit une variable quantitative Y et un ensemble de p variables quantitatives X^1, X^2, … X^p.
Le coefficient de corrélation linéaire multiple empirique R est alors la valeur maximale prise par le coefficient de corrélation linéaire entre Y et une combinaison linéaire des X^j.

$$R = \underbrace{\sup}_{a_1, a_2, \ldots a_p} r\left(Y, \sum_{j=1}^{p} a_j X^j\right)$$

On a donc toujours :

$0 \leq R \leq 1$

R = 1 s'il existe une combinaison linéaire des X^j telle que :

$$Y = a_0 + \sum_{j=1}^{p} a_j X^j$$

Interprétation géométrique de R :