

Régression et corrélation

BUT : expliquer une *variable dépendante* ("dependent variable") y en fonction de *variables explicatives* ("explanatory variables") x_1, x_2, \dots :

Erreur aléatoire notée $e \Rightarrow y = f(x_0, x_1, x_2, \dots) + e$

Régression linéaire: $y = b_0x_0 + b_1x_1 + b_2x_2 + \dots + e$

Remarque 1. = *équation* du modèle

Remarque 2. *Coefficients de régression* b_0, b_1, b_2, \dots

Remarque 3. *Constante ("intercept")* $b_0: x_0 = 1$

Droite de régression: 1 variable (+ constante)

Régression linéaire simple: $y = b_0 + b_1x + e.$

Exemple: x = dosage, y = réponse

Autres variables explicatives potentielles: poids, durée, etc.

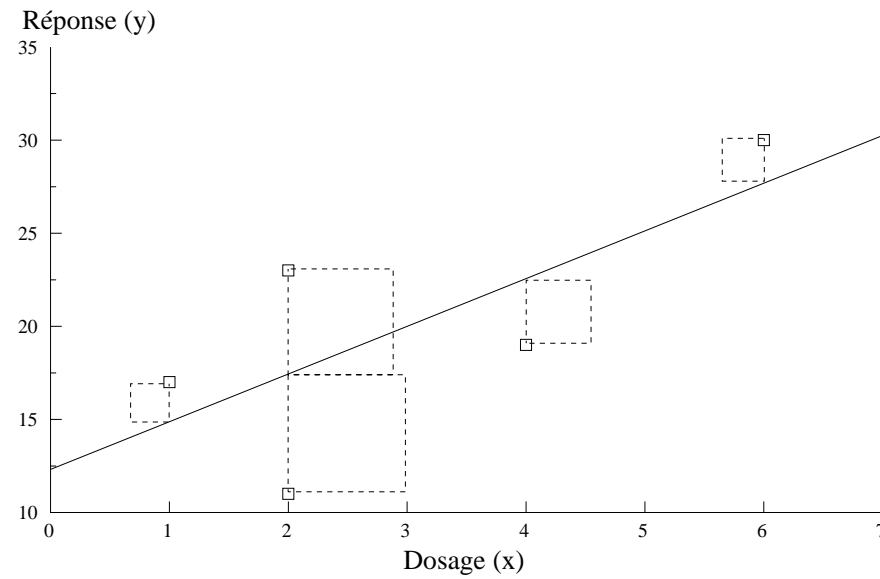
Méthode des moindres carrés ("least square method")

x (dosage)	y (réponse)
1	17
2	11
2	23
4	19
6	30

Meilleure prévision de y sans employer x : $\bar{y} = 20$

**au sens que la somme des carrés des écarts entre les y_i et la prévision est minimum
pour cette prévision = \bar{y}**

Minimiser la somme des carrés (S.C.) des écarts par rapport à la droite, mesurés sur l'axe des y



Droite d'équation $y = 13 + 2x \Rightarrow$ résidus e_i

x_i	y_i	y_i^*	e_i	e_i^2
1	17	15,0000	2,0000	4,0000
2	11	17,0000	-6,0000	36,0000
2	23	17,0000	6,0000	36,0000
4	19	21,0000	-2,0000	4,0000
6	30	25,0000	5,0000	25,0000
Somme			5,0000	105,0000

Pour diminuer la S.C. résiduelle, réduire la somme de 5 à 0 \Rightarrow ajouter 1 à la constante $\Rightarrow b_0 = 13 + 1 = 14$

Moyennes arithmétiques de x_i et y_i : $\bar{x} = 3, \bar{y} = 20$

Puisque $\bar{y} = 14 + 2 \bar{x}$ ($= 20$),

par soustraction de l'équation $y = 14 + 2x$:

$y - \bar{y} = 2(x - \bar{x})$ ou $y = 20 + 2(x - 3)$

Droite d'équation $y = 14 + 2x$

x_i	y_i	y_i^*	e_i	e_i^2
1	17	16,0000	1,0000	1,0000
2	11	18,0000	-7,0000	49,0000
2	23	18,0000	5,0000	25,0000
4	19	22,0000	-3,0000	9,0000
6	30	26,0000	4,0000	16,0000
Somme			0,0000	100,0000

Ensuite, remplacer 2 par b et chercher b qui minimise la S. C. de

$$e_i = y_i - [20 + b(x_i - 3)]$$

x_i	$x_i - \bar{x}$	y_i	$\bar{y} + b(x_i - \bar{x})$	e_i	e_i^2
1	-2	17	$20 + (-2)b$	$-3 + (2)b$	$9 + (-12)b + 4b^2$
2	-1	11	$20 + (-1)b$	$-9 + (1)b$	$81 + (-18)b + 1b^2$
2	-1	23	$20 + (-1)b$	$3 + (1)b$	$9 + (6)b + 1b^2$
4	1	19	$20 + (1)b$	$-1 + (-1)b$	$1 + (2)b + 1b^2$
6	3	30	$20 + (3)b$	$10 + (-3)b$	$100 + (-60)b + 9b^2$
$\bar{x} = 3$		$\bar{y} = 20$	$0 + (0)b$	$200 + (-82)b + 16b^2$	à minimiser

Minimum de $b = 82/(2 \times 16) = 2,5625 \Rightarrow b_1 = 2,5625$

Variance de x_i : $s_x^2 = \frac{1}{n} \sum_{i=1}^n [(x_i - \bar{x})^2] = 16/5 = 3,2$

Covariance entre x_i et y_i : $s_{xy} = \frac{1}{n} \sum_{i=1}^n [(x_i - \bar{x})(y_i - \bar{y})] = 41/5 = 8,2$

$$\Rightarrow b_1 = \frac{\text{cov}(x, y)}{\text{var}(x)} = \frac{s_{xy}}{s_x^2} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} = 8,2/3,2 = 2,5625$$

(= rapport entre la covariance entre x_i et y_i et la variance de x_i)

Constante ("intercept"): $\bar{y} = b_0 + b_1\bar{x}$

Signification: la droite contient le point (\bar{x}, \bar{y})

ou la moyenne \bar{e} des résidus est égale à zéro \Rightarrow

$$b_0 = \bar{y} - b_1\bar{x} = 20 - 2,5625 \times 3 = 12,3125$$

Remarques

1) Plus généralement: soit la somme des carrés

$$Q(b_0, b_1) = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n [y_i - (b_0 + b_1 x_i)]^2.$$

Pour la minimiser en tant que fonction de b_0 et b_1 , exprimer que les dérivées (partielles) par rapport à b_0 et b_1 sont égales à zéro \Rightarrow 2 équations linéaires à deux inconnues \Rightarrow FACILE A CALCULER

quand on résout, donne b_0 et b_1 comme ci-dessus: SOLUTION UNIQUE (sauf dans un cas: $\text{var}(x) = 0$)

2) Présentation des calculs

x_i	$x_i - \bar{x}$	y_i	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1	-2	17	-3	4	9	6
2	-1	11	-9	1	81	9
2	-1	23	3	1	9	-3
4	1	19	-1	1	1	-1
6	3	30	10	9	100	30
\bar{x} : 3		\bar{y} : 20		16	200	41

Qualité de l'ajustement ("goodness of fit"): variance des résidus

$$s_{y \cdot x}^2 = \text{MSE} = \frac{\sum_{i=1}^n e_i^2}{n}$$

x_i	y_i	\hat{y}_i	e_i	e_i^2
1	17	14,8750	2,1250	4,5156
2	11	17,4375	-6,4375	41,4414
2	23	17,4375	5,5625	30,9414
4	19	22,5625	-3,5625	12,6914
6	30	27,6875	2,3125	5,3477
Somme			0,0000	94,9375
Variance résiduelle = $\frac{94,9375}{5} - \left(\frac{0}{5} \right)^2 = 18,9875$				

La meilleure prévision de y

a) sans employer x : $\bar{y} = 20$

b) en employant x et la régression linéaire: pour $\hat{x} = 5$, par exemple, $\hat{y} = 12,31 + 2,56 \times 5 = 25,1$

Qualité de l'ajustement (en utilisant l'échantillon):

a) erreur = $y_i - \bar{y}$ et MSE = $s_y^2 = 40$

b) erreur = résidu des moindres carrés et MSE = variance résiduelle $s_{y \cdot x}^2 = 94,9375/5 = 18,9875$

Coefficient de corrélation entre x et y

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

x_i	y_i	x_i^2	$x_i y_i$	y_i^2
1	17	1	17	289
2	11	4	22	121
2	23	4	46	529
4	19	16	76	361
6	30	36	180	900
$\bar{x}: 3$	$\bar{y}: 20$	61	341	2200

$s_x^2 = \frac{61}{5} - (3)^2 = 3,2$
$s_y^2 = \frac{2200}{5} - (20)^2 = 40$
$s_{xy} = \frac{341}{5} - 60 = 8,2$
$r = \frac{8,2}{(3,2 \cdot 40)^{1/2}} = 0,7248$
$r^2 = R^2 = 0,5253$

Coefficient de détermination

= carré du coefficient de corrélation

$$R^2 = r_{xy}^2 = \left(\frac{s_{xy}}{s_x s_y} \right)^2$$

Le rapport $s_{y \cdot x}^2 / s_y^2 = 18,988 / 40 = 0,4747 = 1 - R^2$

Plus généralement $MSE = s_{y \cdot x}^2 = s_y^2(1 - R^2)$

\Rightarrow pour le calculer, pas besoin de calculer les résidus

$$MSE = 40 (1 - 0,5253) = 40 \times 0,4747 = 18,988$$

Conséquence: décomposition de la variance :

$$s_y^2 = s_y^2(1 - R^2) + s_y^2 R^2.$$

$1 - R^2 =$ *proportion de var(y) qui n'est pas expliquée par la régression*

$R^2 =$ *proportion de var(y) expliquée par la régression*

Exemple:

proportion de la variance (de la réponse) non expliquée (par le dosage) = $18,988/40 = 0,4747$ (< 50%)

proportion de la variance (de la réponse) expliquée (par le dosage) = $0,5253$ (> 50%)

Ecart-type résiduel ("residual standard deviation") = $RMSE = \sqrt{MSE}$

Lien entre corrélation et régression

1) Droite de régression horizontale: $y = \bar{y}$.

Résidus = $y_i - \bar{y} \Rightarrow$ **moyenne** = 0; **variance** = s_y^2

$\Rightarrow R^2 = 0 \Rightarrow r = 0 \Rightarrow s_{xy} = 0 \Rightarrow b_1 = 0$

x et y non liés (linéairement du moins)

2) Pente positive: $b_1 > 0 \Rightarrow r > 0 \Rightarrow R^2 > 0 \Rightarrow s_{y \cdot x}^2 < s_y^2$

\Rightarrow points autour d'une droite croissante

x et y liés dans la même direction

3) Pente négative: $b_1 < 0 \Rightarrow r < 0 \Rightarrow R^2 > 0 \Rightarrow s_{y \cdot x}^2 < s_y^2$

\Rightarrow points autour d'une droite décroissante

x et y liés en sens opposés

Si les points se rapprochent de la droite

alors

$$s_{y \cdot x}^2 \downarrow 0,$$
$$R^2 \uparrow 1,$$
$$r \downarrow -1 \text{ ou } \uparrow 1$$

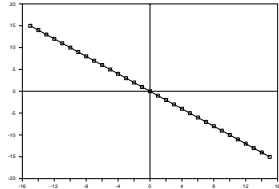
Propriétés de r

(1) r situé entre -1 et 1

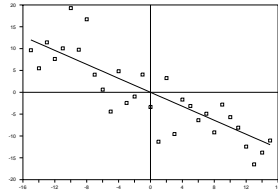
(2) r , b_1 et s_{xy} ont le même signe

(3) le signe et l'ordre de grandeur de r peuvent être interprétés

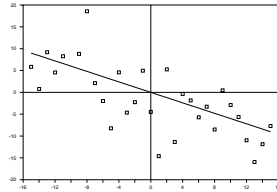
$$r = -1$$



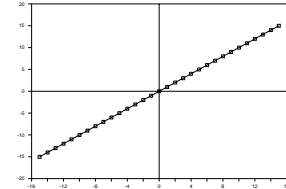
$$r = -0,8$$



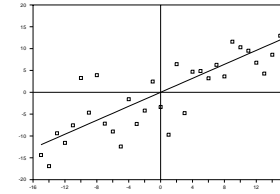
$$r = -0,6$$



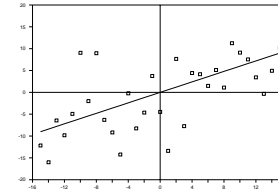
$$r = 1$$



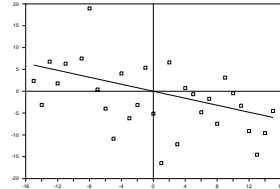
$$r = 0,8$$



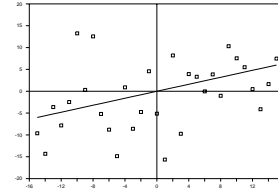
$$r = 0,6$$



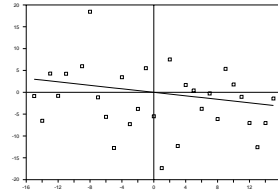
$$r = -0,4$$



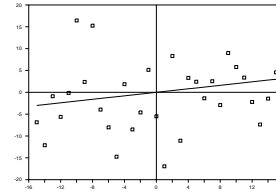
$$r = 0,4$$



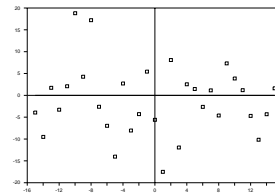
$$r = -0,2$$



$$r = 0,2$$



$$r = 0$$



Résumé

Relation linéaire

négative

inexistante

positive

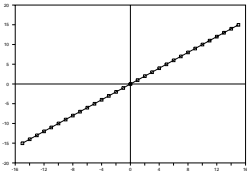
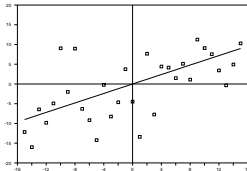
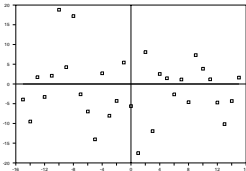
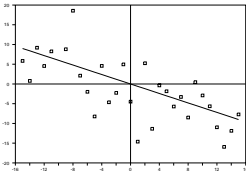
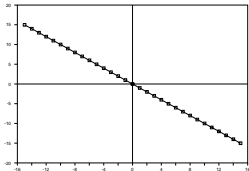
$r = -1$

$r = 0$

$r = 1$

I
M
P
O
S
S
I
B
L
E

I
M
P
O
S
S
I
B
L
E



Moindres carrés = une méthode exceptionnellement fiable?

- **solution unique**
- **aisée à implanter**
- **ne nécessite qu'un passage sur les données**
- **mesure de la qualité d'ajustement**
- **pas besoin de calculer les résidus**

Mais toute l'information est-elle condensée dans un petit nombre de paramètres b_1 , R^2 , $s_{y \cdot x}$?

Mais toute l'information est-elle condensée dans un petit nombre de paramètres b_1 , R^2 , $s_{y \cdot x}$?

En fait non !

L'information peut aussi se cacher dans les résidus

⇒ conclusions hâtives

Comment le savoir?

Regarder les résidus

Si on peut voir une structure, les M.C. sont mal adaptés

le modèle doit être amélioré

(pas une régression linéaire simple

ou pas les moindres carrés!)

Autrement (pas de structure dans les résidus)

alors c'est O.K.

Exemple des jeux de données d'Anscombe [1973]

Jeu A		Jeu B		Jeu C		Jeu D	
x	y	x	y	x	y	x	y
10	8,04	10	9,14	10	7,46	8	6,58
14	9,96	14	8,10	14	8,84	8	5,76
5	5,68	5	4,74	5	5,73	8	7,71
8	6,95	8	8,14	8	6,77	8	8,84
9	8,81	9	8,77	9	7,11	8	8,47
12	10,84	12	9,13	12	8,15	8	7,04
4	4,26	4	3,10	4	5,39	8	5,25
7	4,82	7	7,26	7	6,42	19	12,50
11	8,33	11	9,26	11	7,81	8	5,56
13	7,58	13	8,74	13	12,74	8	7,91
6	7,24	6	6,13	6	6,08	8	6,89

la même droite de régression

$$y = 3 + 0,5x$$

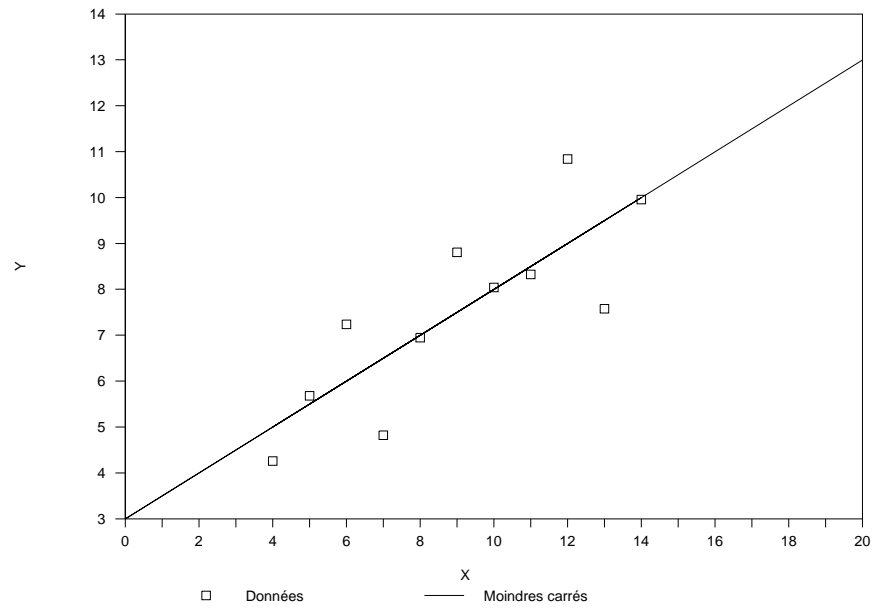
le même écart-type résiduel

1,236

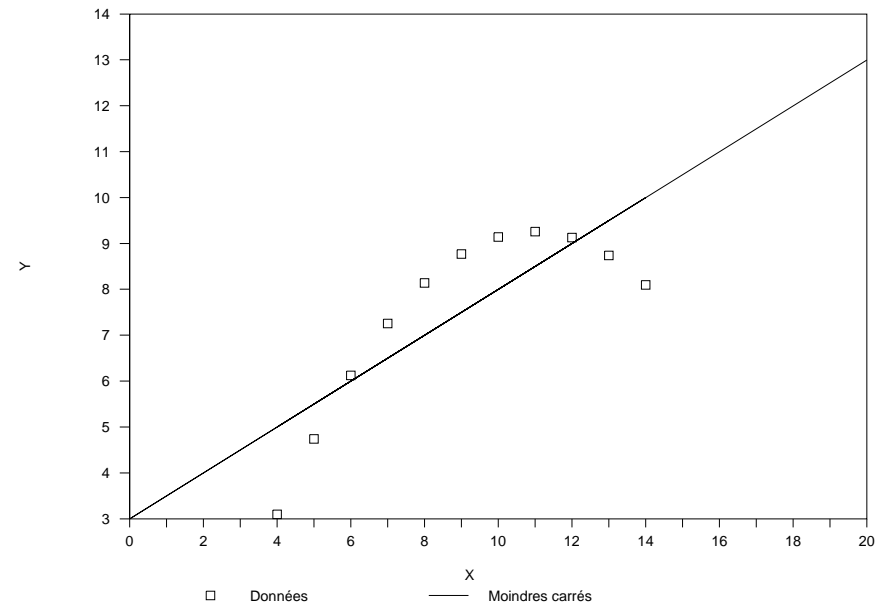
le même coefficient de détermination

0,667

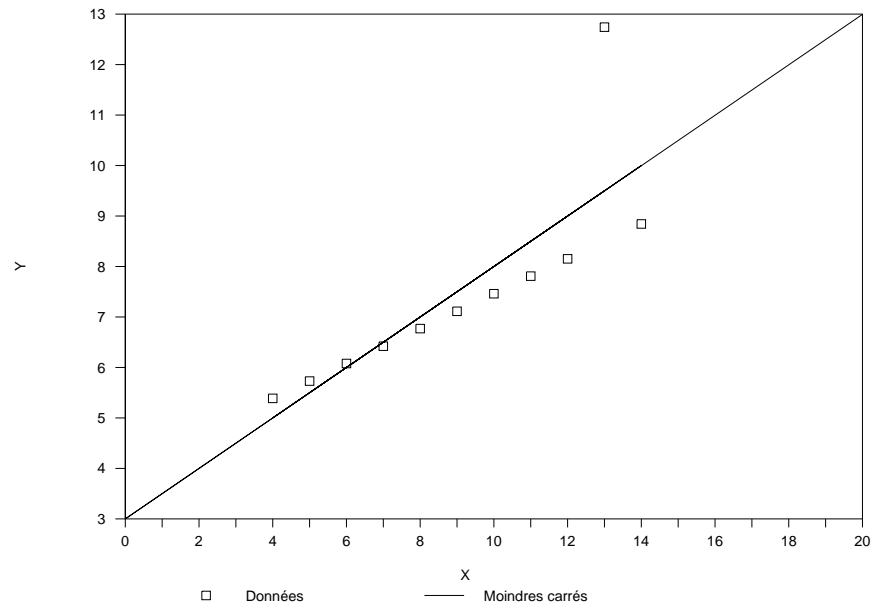
Jeu A



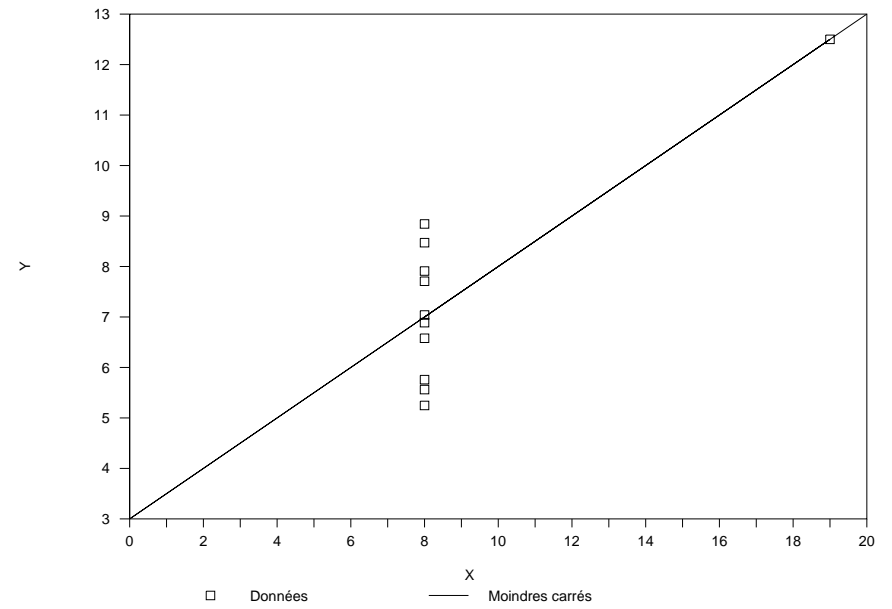
Jeu B



Jeu C



Jeu D



Exemple

Corrélation entre les dosages d'enzymes et l'âge:

FNU	-0,32	dépendance en sens inverse
ADA	-0,35	dépendance en sens inverse
PNP	0,37	dépendance en sens direct

Corrélation entre les dosages d'enzymes et les pourcentages

	FNU	ADA	PNP
PCLYMPH	-0,08	-0,18	0,50
PCOKT4	0,05	0,28	0,07
PCOKT8	0,13	-0,65	0,14

Ces corrélations ont été calculées sur 8 observations

Remarque:

- 1) la dépendance entre les dosages d'enzymes et les traitements peuvent mieux s'exprimer en une comparaison de moyennes (moyenne des dosages avec ou sans traitement)**
- 2) la dépendance entre les dosages d'enzymes et la sévérité de l'affection ne peut pas s'exprimer par un coefficient de corrélation car les différences entre sévérités ne sont pas quantitatives mais qualitatives; on pourrait employer les *rangs* des données**