L’ajustement linéaire et corrélation

Daamouch

Données quantitatives : nuages de points

CA et spots publicitaires pour le magasin Truc

Semaines	Nombres de spots publicitaires	CA en centaines de dollars
1	2	50
2	5	57
3	1	41
4	6	54
5	5	54
6	1	38
7	6	63
8	3	48
9	4	59
10	7	65

Source : adapté de Anderson et *alii* (2001)

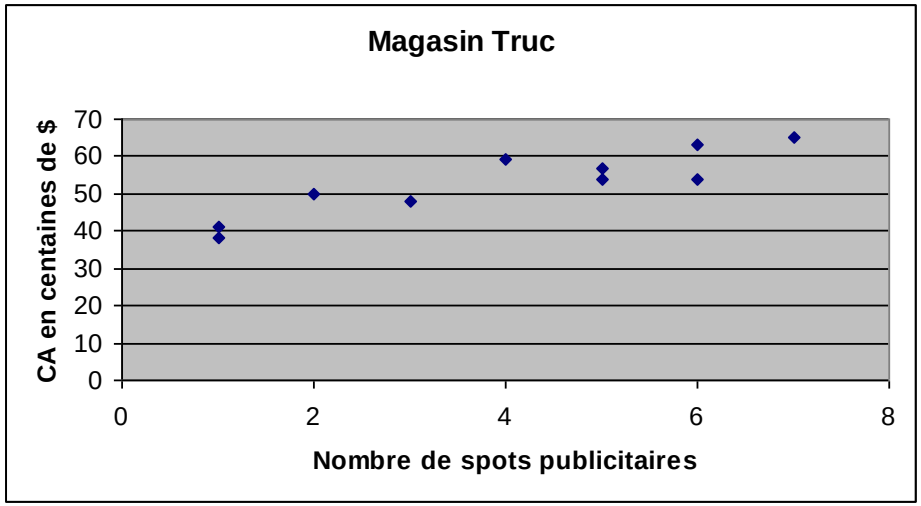
Question : existe-t-il une liaison statistique entre le nombre de spots et le CA ?

Le CA et le nombre de spots évoluent-ils de manière concomitante ?

Données quantitatives : nuages de points

Un représentation graphique du nuage de points (ou diagramme de corrélation) permet :

- D'apprécier l'existence ou non d'une éventuelle liaison
- De déterminer la forme de la liaison

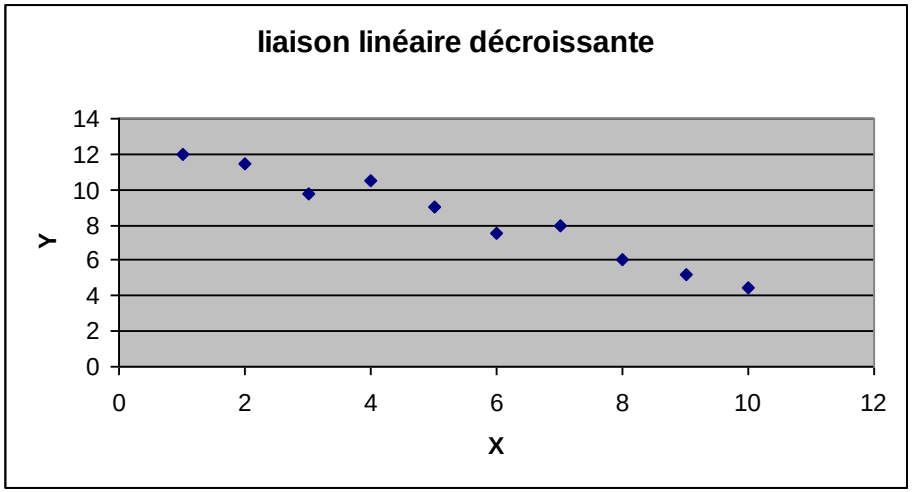


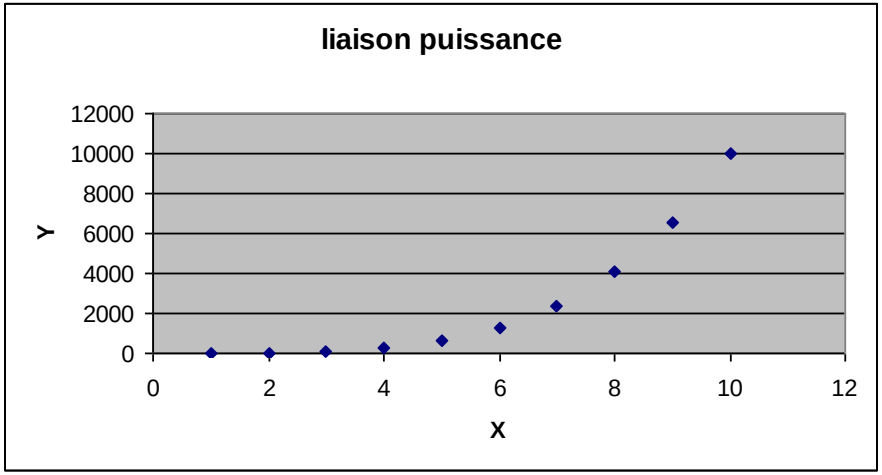
Données quantitatives : nuages de points

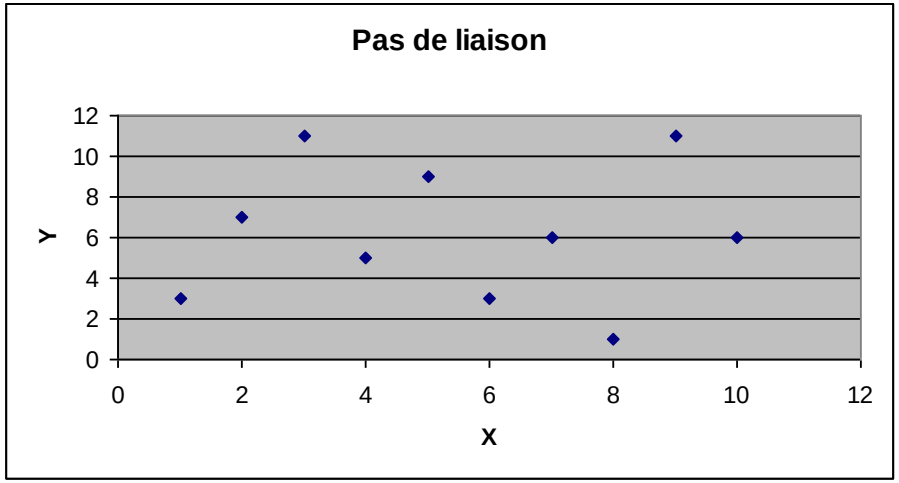
La forme du nuage de point suggère les interprétations suivantes :

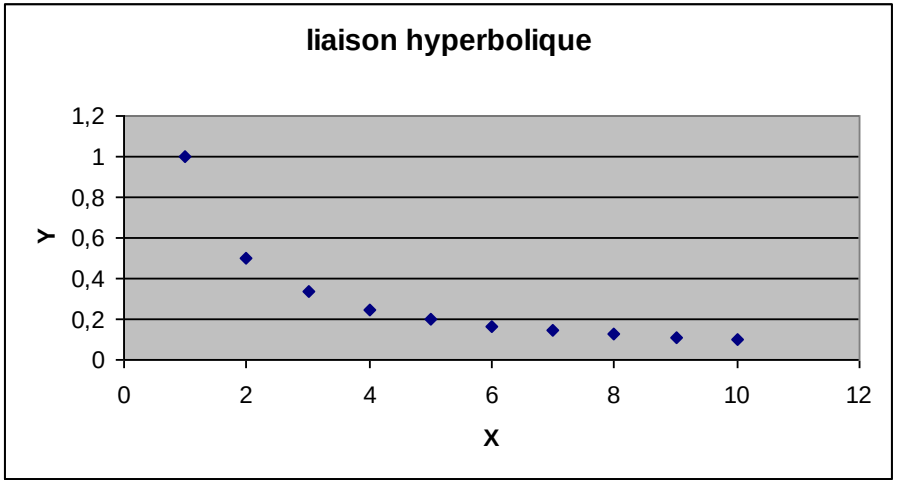
- Il existe une liaison entre les 2 variables : si le nombre de spots varient alors le CA a tendance à varier aussi
- Cette liaison est linéaire : les points sont à peu près alignés sur une droite
- Cette liaison est positive : plus le nombre de spots s'accroît, plus le CA augmente.

Nuages de points : formes de liaison









Covariance

Pour le magasin, le nuage de points montre que les variables ont tendance à covarier (varier ensemble)

$\Rightarrow$ Construction d'un indicateur qui mesure la variabilité conjointe des 2 variables.

- Mesure descriptive de la relation entre les 2 variables
- Mesure les fluctuations simultanées de chaque variable par rapport à sa moyenne

Covariance : calculs

$$COV(X,Y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{N}$$

$$COV(X,Y) = \frac{1}{N}\sum_i x_i y_i - \bar{x}\bar{y}$$

COV (X,Y) = moyenne du produit XY – produit des moyennes de X et de Y

Calcul de la covariance pour le magasin Truc

Semaines	Nombres de spots publicitaires (X)	CA en centaines de dollars (Y)	XY
1	2	50	100
2	5	57	285
3	1	41	41
4	6	54	324
5	5	54	270
6	1	38	38
7	6	63	378
8	3	48	144
9	4	59	236
10	7	65	455
Moyenne	**4**	**52,9**	**227,1**

Covariance = 227,1 - 4*52,9 = 15,5

Covariance : interprétation

Covariance > 0 $\Rightarrow$ les variables ont tendance à varier dans le même sens

Covariance < 0 $\Rightarrow$ les variables ont tendance à varier en sens opposée

$\Rightarrow$ Plus la valeur (>0 ou <0) de la covariance est élevée plus la relation entre les variables est forte

$\Rightarrow$ S'il n'y a pas de tendance à la croissance ou à la décroissance entre les variables covariance nulle

☞ La covariance est un indicateur de <u>relation linéaire</u> entre les variables

$\Rightarrow$ Covariance = 0 peut signifier une relation non linéaire.

Coefficient de corrélation linéaire

Covariance dépend des unités des variables $\Rightarrow$ coefficient de corrélation linéaire.

Coefficient de corrélation linéaire

$$r = \frac{COV(X,Y)}{\sigma_x \sigma_y} \qquad r = \frac{15{,}5}{2{,}049 * 8{,}37} = 0{,}903$$

- -1 < r < 1
- Si r = 1 ou r = -1 alors points parfaitement alignés

Régression linéaire

Il s'agit de caractériser quantitativement le lien entre les deux variables.

Seule situation envisagée : le nuage de points suggère une liaison linéaire :

$$y = ax + b \Rightarrow$$

En connaissant l'équation de la droite qui résume la relation, il est possible de faire des prévisions

Remarque : attention à la véracité statistique de ces prévisions lorsqu'on sort de l'intervalle de l'échantillon

Régression linéaire

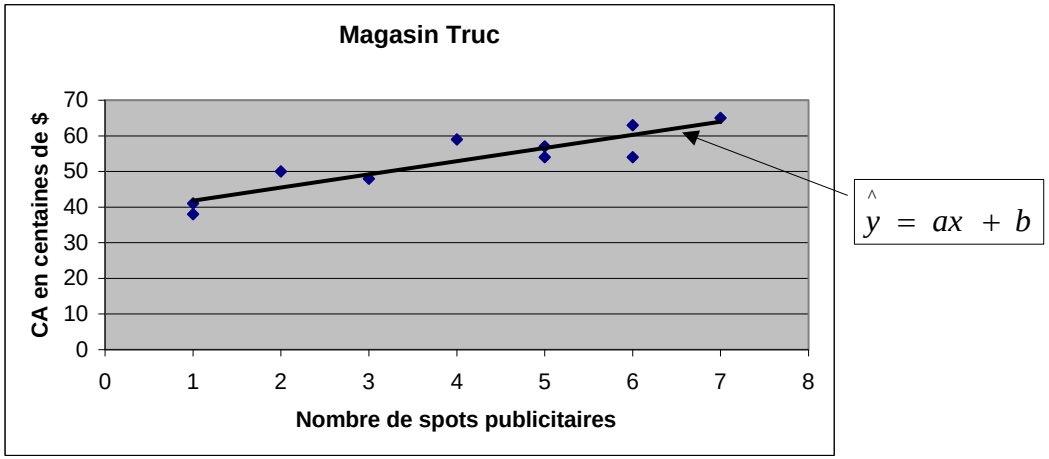
On cherche donc à estimer la droite qui s'ajuste le mieux au nuage de point

Notation

y = vraies valeurs de la valeur de variable y c'est la variable expliquée

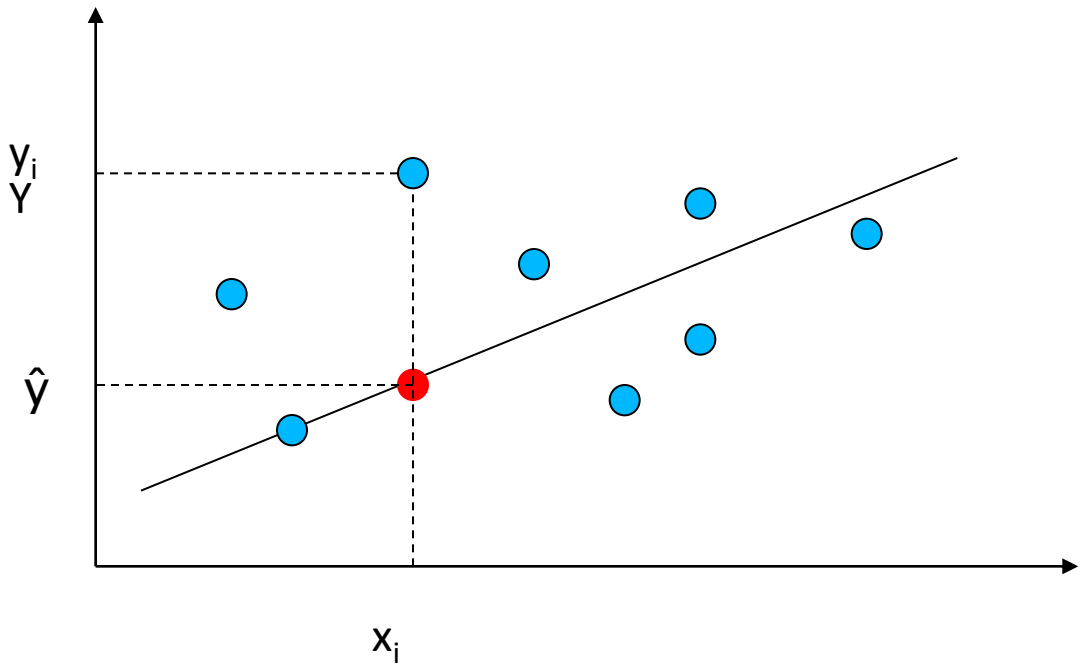
$\hat{y}$ = valeurs de la variables y obtenues à l'aide du modèle

x = variable dépendante ou variable explicative



Régression linéaire

Méthodologie : minimisation de la somme des carrés des écarts entre la véritable valeurs de y_i et son estimation



Régression linéaire

La droite de régression

a pour équation

$$a = \frac{COV(X,Y)}{Var(X)}$$

$$b = \bar{y} - a\,\bar{x}$$

Calcul de la de la droite de régression pour le magasin Truc

Semaines	Nombres de spots publicitaires (X)	CA en centaines de dollars (Y)	XY	X^2
1	2	50	100	4
2	5	57	285	25
3	1	41	41	1
4	6	54	324	36
5	5	54	270	25
6	1	38	38	1
7	6	63	378	36
8	3	48	144	9
9	4	59	236	16
10	7	65	455	49
Total	**40**	**529**	**2271**	**202**
Moyenne	***4***	***52,9***	***227,1***	

Cov(X,Y) = 227,1 - 4*52,9 = 15,5

Var (X) = 202/10 - 4 ^{2} = 4,2

a = 15,5/4,2 = 3,69

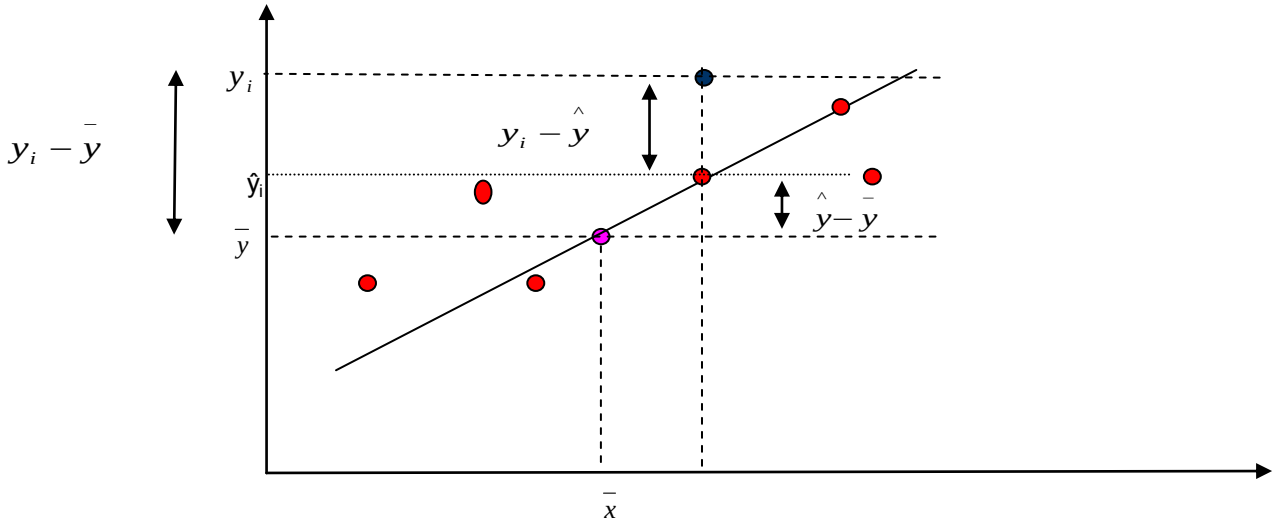
b = 52,9 - 3,69*4 = 38,14

$\hat{y}$ = 3,69x + 38,14

Régression linéaire : coefficient de détermination

Cette droite explique-t-elle de façon satisfaisante les variations de y (ou la variance de y)

La droite de régression passe par la covariance $\Rightarrow$ moy (ŷ) = $\bar{y}$



$$y_i - \bar{y} = y_i - \hat{y} + \hat{y} - \bar{y}$$

on montre que $\left(y_i - \bar{y}\right)^2 = \left(\hat{y} - \bar{y}\right)^2 + \left(y_i - \hat{y}\right)^2 \Rightarrow SCT = SCE + SCR$

Régression linéaire : coefficient de détermination

Calcul de la covariance pour le magasin Truc

Semaines	Nombres de spots publicitaires (X)	CA en centaines de dollars (Y)	Ŷ	(Y - Ŷ)	$(Y- m_y)^2$	$(\hat{Y} - m_{\hat{y}})^2$	$(Y - \hat{Y})^2$
1	2	50	45,52	4,48	8,41	54,48	20,08
2	5	57	56,59	0,41	16,81	13,62	0,17
3	1	41	41,83	-0,83	141,61	122,58	0,69
4	6	54	60,28	-6,28	1,21	54,48	39,45
5	5	54	56,59	-2,59	1,21	13,62	6,71
6	1	38	41,83	-3,83	222,01	122,58	14,66
7	6	63	60,28	2,72	102,01	54,48	7,39
8	3	48	49,21	-1,21	24,01	13,62	1,46
9	4	59	52,90	6,10	37,21	0,00	37,21
10	7	65	63,97	1,03	146,41	122,58	1,06
Total	**40**	**529**			**700,90**	**572,02**	**128,88**
Moyenne	**4**	**52,9**			***SCT***	***SCE***	***SCR***

a = 15,5/4,2 = 3,69
b = 52,9 - 3,69*4 = 38,14

ŷ = 3,69x + 38,14

SCT = 572,02 + 128,88 = 700,90

$$R2 = \frac{SCE}{SCT}$$

$$R2 = \frac{572,02}{700,9}$$

R2 = 81,61

Régression linéaire : coefficient de détermination

R^2 représente la part de la variabilité de Y « expliquée » par la droite de régression.

$R^2 \leq 1$

Si les observations sont parfaitement alignées, il n'y a pas de différence entre y et $\hat{y}$ $\Rightarrow$ pas de résidu $\Rightarrow$ SCT = SCE $\Rightarrow$ $R^2 = 1$

Donc R^2 exprime la qualité du modèle. Plus est proche de 1, meilleure est la qualité du modèle linéaire

Ici le nombre de spots publicitaires « explique » 81,61% de la dispersion des CA

<u>Remarque</u> : $R^2 = r^2$, uniquement pour un modèle linéaire