

# Cours de Statistique Inférentielle

Jean Christophe meunier

## Module 4 Covariance, corrélation et régression linéaire

2<sup>ème</sup> Bac, Commerce Extérieur  
Année académique 2015-2016

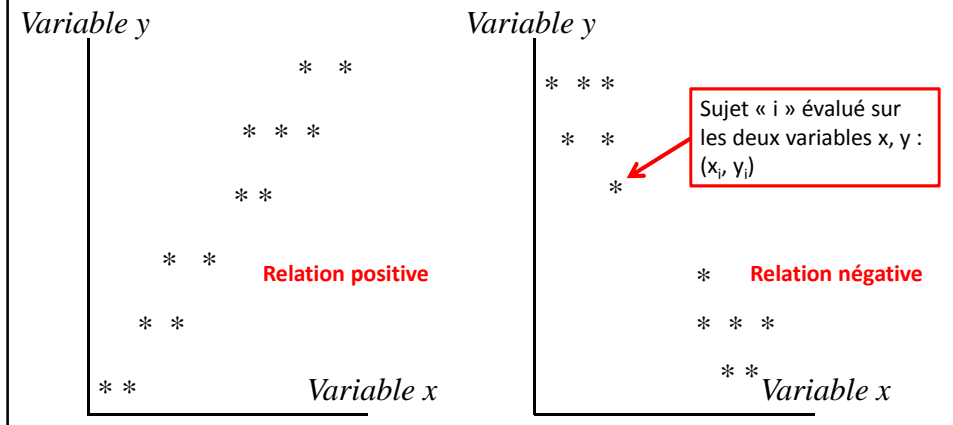


### Mesure d'association entre 2 variables

- Jusqu'à présent, travail sur une seule variable
  - Statistique descriptive, lois de probabilité, IC, comparaisons de moyenne...
- Dans ce module, intérêt sur l'association entre deux variables :
  - Dans quelle mesure un changement sur une variable (x) est-il associé à un changement sur une seconde variable (y) ?
  - Ex :
    - Temps d'étude (x) - réussite (y)
    - Opinion sur un produit (x) – achat de ce produit (y)
    - ...

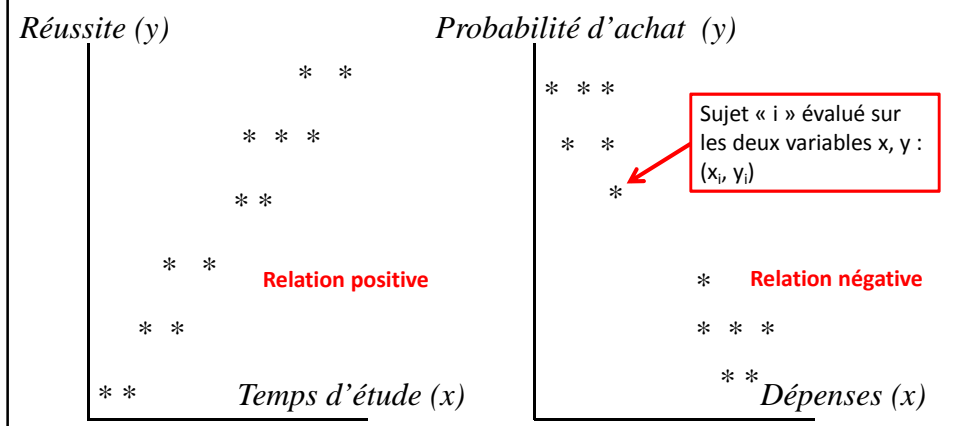
## Mesure d'association entre 2 variables

- Visuellement :
  - ‘Nuage de points’ ou ‘Diagramme de dispersion’



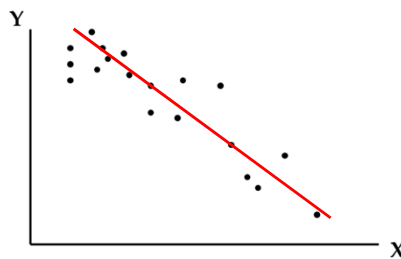
## Mesure d'association entre 2 variables

- Exemples :



## Mesure d'association entre 2 variables

- On parle d'association linéaire :
  - la relation entre les variables  $x$  et  $y$  peut être décrite/estimée par une ligne droite
  - Pour variables d'échelle intervalle :
    - Les variables d'échelle nominale ou ordinale n'ont pas de gradation à intervalle régulier (association avec d'autres variables impossible ou difficile à établir)



## Mesure d'association entre 2 variables

- Principe :
  - Trouver une mesure numérique pour représenter la relation entre deux variables
- Trois mesures les plus courantes :
  - Covariance : mesure d'association
  - Corrélation : mesure 'standardisée' d'association
  - Régression linéaire : mesure d'association causale (ex :  $x$  prédit  $y$ )

## A. Covariance

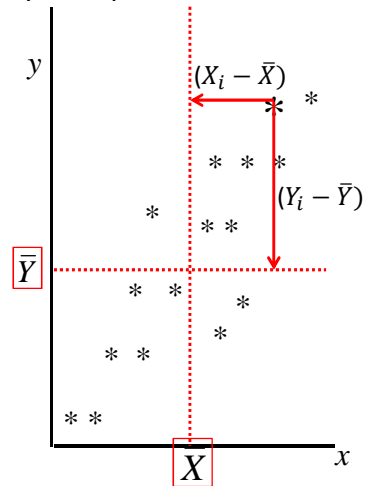
### I. Covariance

- Notation :  $\text{Cov}(x,y)$  ou  $S_{xy}$
- Mesure numérique qui donne la direction de la relation entre deux variables  $x$  et  $y$ 
  - Ne donne pas l'intensité de la relation
  - N'établit pas de causalité entre les variables ( $x$  prédit  $y$  ou  $y$  est prédit par  $x$ )
- L'une des mesures les plus fondamentales d'association

## I. Covariance

- Pour chaque sujet
  - Multiplier écarts à la moyenne pour les deux variables

$$(X_i - \bar{X}) * (Y_i - \bar{Y})$$



## I. Covariance

- Exemple pour un ensemble de sujets (échantillon)

Sujets	Variable X	Variable Y
1	$X_1$	$Y_1$
2	$X_2$	$Y_2$
:	:	:
n	$X_n$	$Y_n$

## I. Covariance

Sujets	X	Y	$X_i - \bar{X}$	$Y_i - \bar{Y}$	$(X_i - \bar{X}) * (Y_i - \bar{Y})$
1	$X_1$	$Y_1$	$(X_1 - \bar{X})$	$(Y_1 - \bar{Y})$	$(X_1 - \bar{X}) * (Y_1 - \bar{Y})$
2	$X_2$	$Y_2$	$(X_2 - \bar{X})$	$(Y_2 - \bar{Y})$	$(X_2 - \bar{X}) * (Y_2 - \bar{Y})$
n	$X_n$	$Y_n$	$(X_n - \bar{X})$	$(Y_n - \bar{Y})$	$(X_n - \bar{X}) * (Y_n - \bar{Y})$
Moyenne	$\bar{X}$	$\bar{Y}$			

Somme de ces valeurs  
pour chacun des sujets  
puis divisé par (n-1)

## I. Covariances

- Cov (X,Y) ou  $s_{xy}$

$$Cov(x, y) = s_{xy} = \frac{(x_1 - \bar{X})(y_1 - \bar{Y}) + (x_2 - \bar{X})(y_2 - \bar{Y}) + \dots + (x_n - \bar{X})(y_n - \bar{Y})}{n-1}$$

$$OU$$

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

Où

$x_i$  et  $y_i$  sont les valeurs observées,

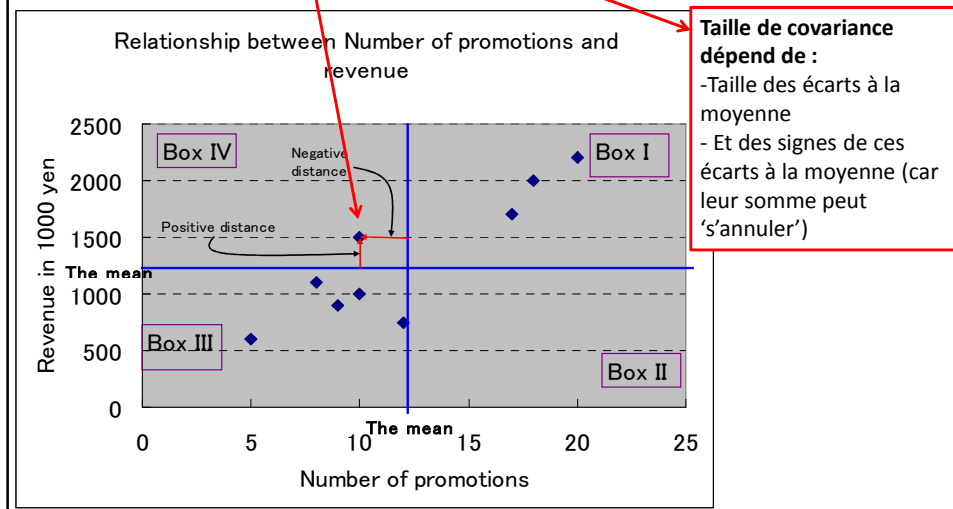
$\bar{X}$  et  $\bar{Y}$  les moyennes des deux variables

n la taille de l'échantillon.

## Remarques sur covariance

- Signe de covariance détermine direction de l'association :
  - Cov + → association positive
  - Cov - → association négative
- La taille de la valeur de la covariance
  - ne détermine pas la force d'association entre les deux variables x et y
  - Dépend notamment de la taille des écarts à la moyenne et de leur signe (cf. dia suivante)
- **En bref**, Covariance donne
  - La direction de l'association (+ ou -)
  - Pas la grandeur de l'association ni la causalité

$$Cov(x, y) = s_{xy} = \frac{(x_1 - \bar{X})(y_1 - \bar{Y}) + (x_2 - \bar{X})(y_2 - \bar{Y}) + \dots + (x_n - \bar{X})(y_n - \bar{Y})}{n-1}$$



## Ex: calcul de covariance

- Lien entre nombre de cigarettes et capacité pulmonaire

Cigs ( $X$ )	Lung Cap ( $Y$ )
0	45
5	42
10	33
15	31
20	29
10	36

## Ex: calcul de covariance

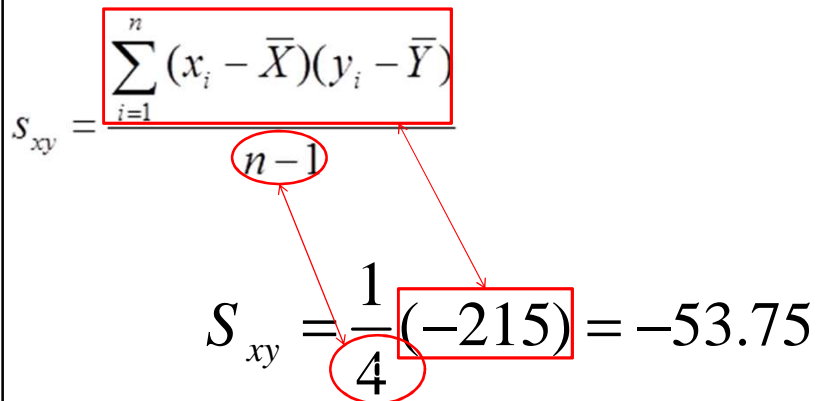
- Lien entre nombre de cigarettes et capacité pulmonaire

Cigs ( $X$ )	$(X - \bar{X})$	$(X - \bar{X})(Y - \bar{Y})$	$(Y - \bar{Y})$	Cap ( $Y$ )
0	-10	-90	9	45
5	-5	-30	6	42
10	0	0	-3	33
15	5	-25	-5	31
20	10	-70	-7	29
		$\Sigma = -215$		



## Ex: calcul de covariance

- Lien entre nombre de cigarettes et capacité pulmonaire

$$S_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{X})(y_i - \bar{Y})}{n-1}$$
$$S_{xy} = \frac{1}{4}(-215) = -53.75$$


## B. Corrélation

## II. Corrélation

- Notation :  $r_{xy}$  ou  $r$
- Mesure numérique qui donne la direction et l'intensité de la relation entre deux variables  $x$  et  $y$ 
  - N'établit pas de causalité entre les variables ( $x$  prédit  $y$  ou  $y$  prédit  $x$ )

## II. Corrélation

- Formule

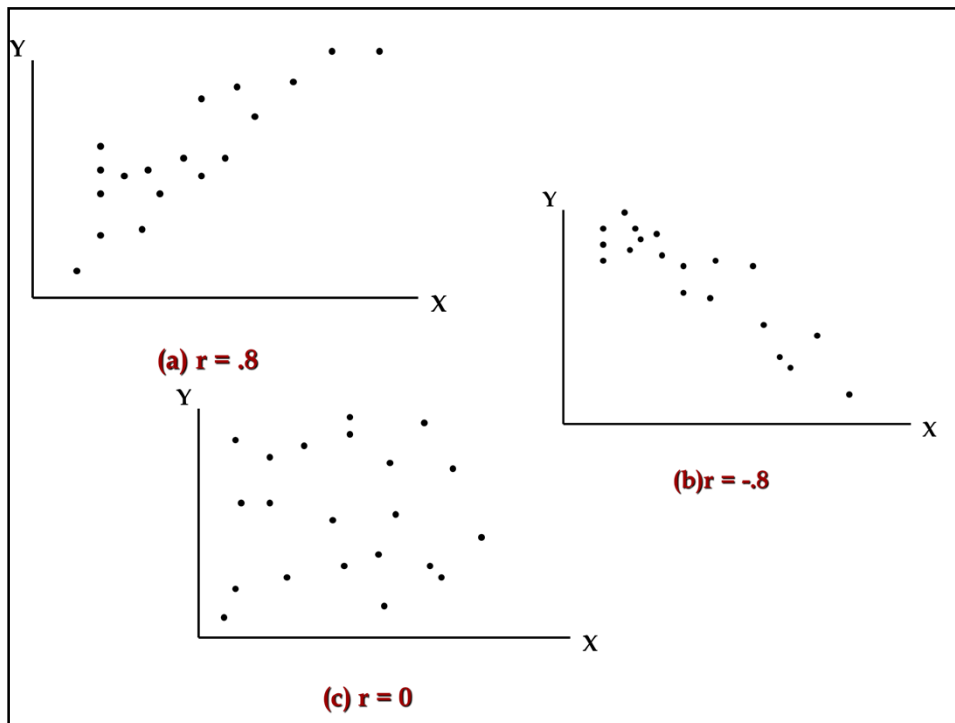
$$r_{xy} = \frac{S_{xy}}{S_x * S_y}$$

où,

$$S_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{X})(y_i - \bar{Y})}{n-1} \quad \text{Covariance entre x et y}$$

$$S_x = \frac{\sum_{i=1}^n (x_i - \bar{X})^2}{n-1} \quad S_y = \frac{\sum_{i=1}^n (y_i - \bar{Y})^2}{n-1} \quad \text{Ecart-type de x et y}$$

- Donne une mesure standardisé de la covariance
  - Covariance pondérée par l'écart-type des deux variables
    - corrélation ne dépend donc plus de la taille des écarts à la moyenne
  - force de l'association comprise entre -1 et 1
    - 1 association parfaite (positive)
    - 0 association nulle
    - -1 association parfaite (négative)



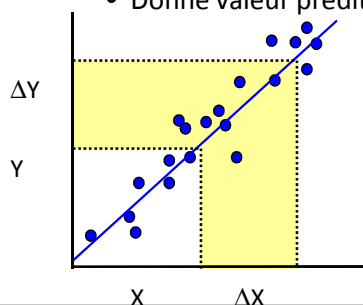
## C. Régression linéaire

### III. Régression linéaire

- Notation :  $\beta$
- Mesure numérique qui donne la direction et l'intensité de la relation et la causalité entre deux variables x et y
  - X : variable indépendante (prédicteur)
  - y : variable dépendante (dépend de/prédit par x)

### III. Régression linéaire

- ‘Nuage de points’ ou ‘Diagramme de dispersion’
  - Par convention
    - X (prédicteur, var. indépendante) sur l'axe horizontal
    - Y (var. dépendante) sur l'axe vertical
  - La ‘pente’ (donnée par  $\beta$ , coefficient de régression)
    - Ajustement d'une ligne droite à travers le nuage de points
    - Donne valeur prédite de Y ( $\hat{Y}$ ) pour n'importe quelle valeur de X



$$\beta = \Delta Y / \Delta X$$

$$\Delta X = X_i - X_j$$

$$\Delta Y = Y_i - Y_j$$

Coefficient  $\beta$  :

- Quantifie la prédiction de X sur Y
- Dans quelle mesure un changement de X est associé à un changement de Y

# Modèle de régression linéaire

$$\hat{Y} = \alpha + \beta X + \varepsilon$$

Où,

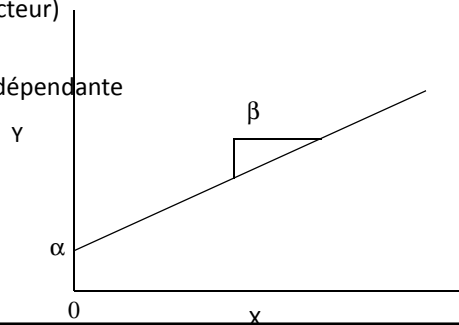
$\alpha$  = intercepte (constante) = valeur prédite de  $y$  à l'origine de  $X$  (quand  $X_i=0$ ) ou à sa valeur moyenne (quand  $X_i = \bar{X}$ )

$\beta$  = pente (coefficient de régression)

$X$  = variable indépendante (prédicteur)

$\varepsilon$  = erreur ou résidu

$\hat{Y}$  = valeur prédite de la variable dépendante



# Modèle de régression linéaire

- Intercepte ( $\alpha$ ) et pente ( $\beta$ )

$$\hat{Y} = \alpha + \beta X + \varepsilon$$

**Intercepte :**

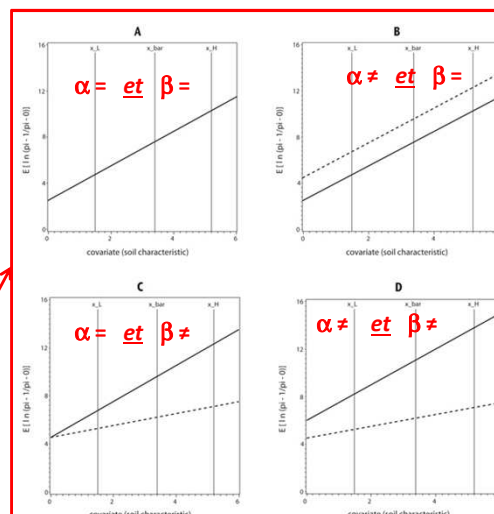
- Constante qui donne la valeur prédite de  $Y$  à l'origine de  $X$

**Pente :**

- Donne degré d'association entre  $x$  et  $y$   
- Toujours associé à  $X$  ( $\beta X$ ) car donne valeur prédite de  $Y$  ( $\hat{Y}$ ) pour toutes valeurs de  $X$

Imaginons 2 modèles de régression

- Ex :  $X_i$  prédit  $Y_i$  et  $X_j$  prédit  $Y_j$



## Modèle de régression linéaire

- Calcul intercepte ( $\alpha$ ) et pente ( $\beta$ )

$$\hat{Y} = \alpha + \beta X (+\varepsilon) \rightarrow \text{On ne s'occupe pas de l'erreur } (\varepsilon) \text{ pour l'instant}$$

- Calcul de  $\varepsilon$  possible une fois que  $\alpha$  et  $\beta$  ont été calculés (cf. diapos suivantes)

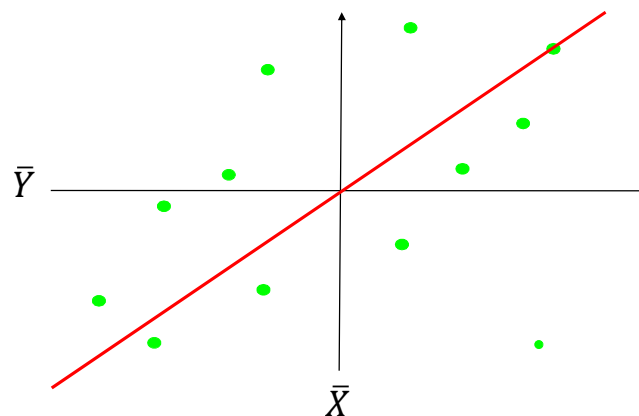
$$\beta = \frac{S_{YX}}{S_X^2} = \frac{\frac{\sum_{i=1}^n (x_i - \bar{X})(y_i - \bar{Y})}{n-1}}{\frac{\sum_{i=1}^n (x_i - \bar{X})^2}{n-1}} = \frac{\sum_{i=1}^n (x_i - \bar{X})(y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{X})^2} \rightarrow \beta = \frac{\sum xy}{\sum x^2}$$

Notation simplifiée !

$$\alpha = \bar{Y} - \beta \bar{X} \rightarrow \text{Valeur moyenne de X et de Y comme référence car la droite de régression passe par le centre de gravité du nuage de points } (\bar{X}, \bar{Y})$$

## Modèle de régression linéaire

- Calcul intercepte ( $\alpha$ ) et pente ( $\beta$ )



$$\alpha = \bar{Y} - \beta \bar{X} \rightarrow \text{Valeur moyenne de X et de Y comme référence car la droite de régression passe par le centre de gravité du nuage de points } (\bar{X}, \bar{Y})$$

## Rappel :

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

**Covariance**

$$r_{xy} = \frac{S_{xy}}{S_x * S_y}$$

**Corrélation**

$$\beta = \frac{S_{xy}}{S_x^2}$$

**Régression linéaire**

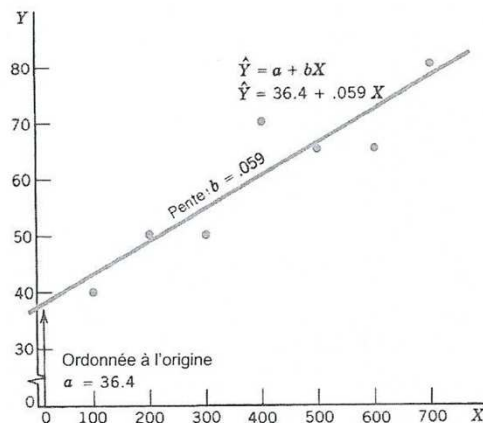
## Modèle de régression linéaire

- Calcul intercepte ( $\alpha$ ) et pente ( $\beta$ )

### Exemple

Etude de l'influence d'un engrais ( $X = \text{kg/ha}$ ) sur le rendement ( $Y = \text{volume de production par hectare} - \text{Q/ha}$ ) des cultures.

X	Y
Engrais (kg/ha)	Rendement (Q/ha)
100	40
200	50
300	50
400	70
500	65
600	65
700	80



Dans Excel, l'option ajouter une courbe de tendance permet de tracer la droite des moindres carrés et d'afficher les coefficients a et b

## Modèle de régression linéaire

- Calcul intercepte ( $\alpha$ ) et pente ( $\beta$ )

Données		Ecart		Produits	
$X$	$Y$	$x = X - \bar{X}$ $= X - 400$	$y = Y - \bar{Y}$ $= Y - 60$	$xy$	$x^2$
100	40	-300	-20	6 000	90 000
200	50	-200	-10	2 000	40 000
300	50	-100	-10	1 000	10 000
400	70	0	10	0	0
500	65	100	5	500	10 000
600	65	200	5	1 000	40 000
700	80	300	20	6 000	90 000
$\bar{X} = 400$ $\bar{Y} = 60$		$\sum x = 0$	$\sum y = 0$	$\sum xy = 16 500$	$\sum x^2 = 280 000$

D'où on tire la valeur de  $\beta$

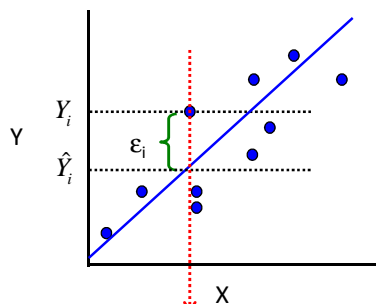
$$\beta = \frac{\sum xy}{\sum x^2} = \frac{16500}{280000} = 0,059$$

Par ailleurs, la droite de régression passe par le centre de gravité  $(\bar{X}, \bar{Y})$  du nuage de points. On a donc:

$$\bar{Y} = \alpha + \beta \bar{X} \rightarrow \alpha = \bar{Y} - \beta \bar{X} = 60 - 0,059 \times 400 = 36,4$$

## Modèle de régression linéaire

- Variable dépendante :  $Y$ 
  - Valeur observée ( $Y_i$ ) vs. Valeur prédite ( $\hat{Y}_i$ )



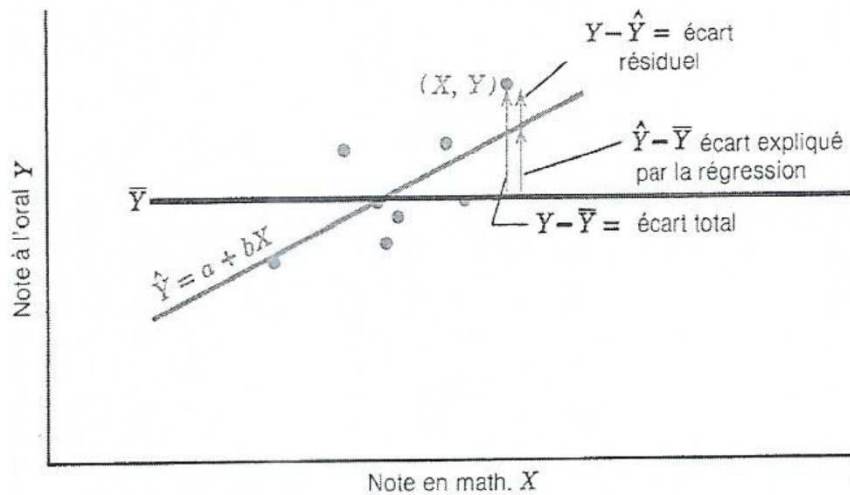
Pour une même valeur  $X_i$

- $Y_i$  : valeur réellement observée
- $\hat{Y}_i$  : valeur prédite par la droite de régression



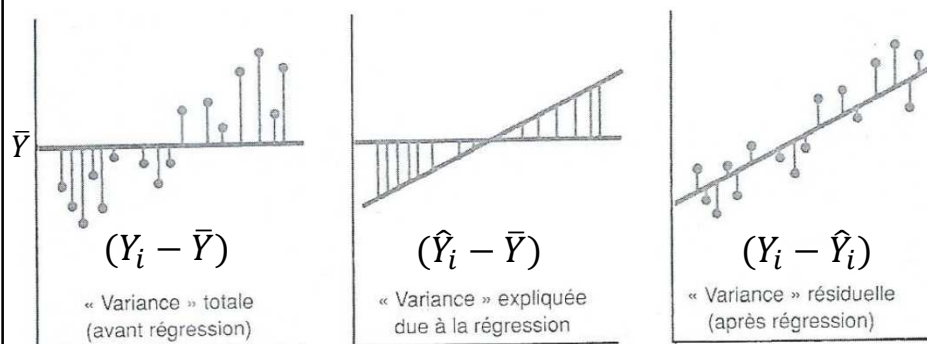
## Modèle de régression linéaire

- Variance expliquée et variance résiduelle



## Modèle de régression linéaire

- Variance expliquée et variance résiduelle

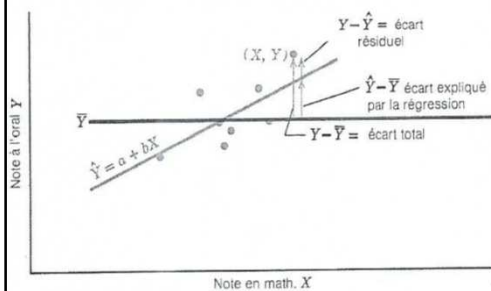


La régression réduit la « variance »

# Modèle de régression linéaire

- Variance expliquée et variance résiduelle

## 2. Variance expliquée et variance résiduelle



### Ecart total

$$Y - \bar{Y} = (\hat{Y} - \bar{Y}) + (Y - \hat{Y})$$

« variance » (en fait, somme des carrés)

$$\sum (Y - \bar{Y})^2 = \sum (\hat{Y} - \bar{Y})^2 + \sum (Y - \hat{Y})^2$$

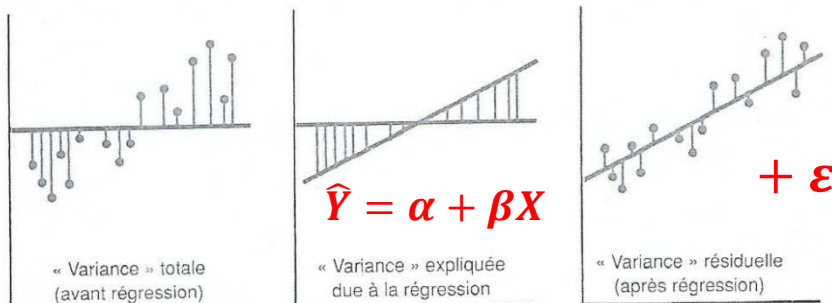
expliqué par X      non expliquée  
résiduelle

# Modèle de régression linéaire

- Variance expliquée et variance résiduelle

$$\sum (Y - \bar{Y})^2 = b^2 \sum x^2 + \sum (Y - \hat{Y})^2$$

Analyse de la « variance » = décomposition



La régression réduit la « variance »

## Modèle de régression linéaire

- Coefficient de détermination :  $r^2$

### Coefficient de détermination $r^2$ et d'indétermination $(1 - r^2)$

On peut montrer que:

$$r^2 = \frac{\text{"variance" expliquée de } Y}{\text{"variance" totale de } Y}$$

$r^2$  est appelé coefficient de détermination. Il fournit la part de la variance totale de  $Y$  expliquée par l'ajustement réalisé grâce à la régression linéaire.

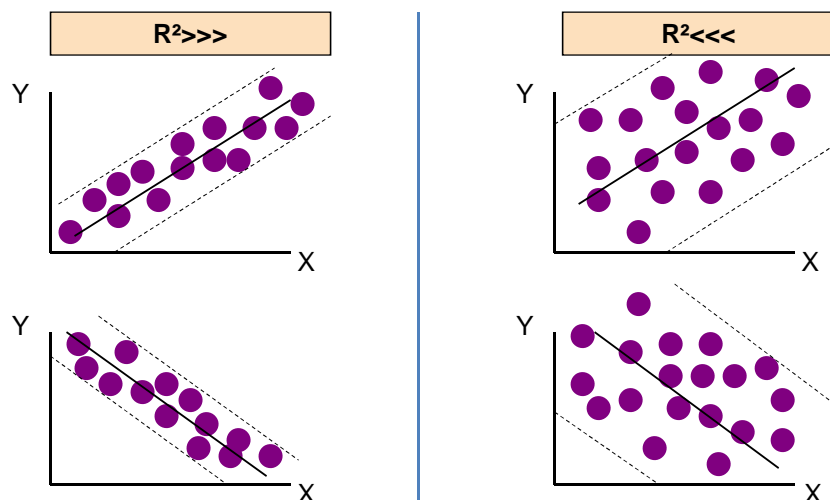
Enfin, on peut aussi montrer que:

$$s^2 = (1 - r^2)s_y^2$$

$(1 - r^2)$  est appelé coefficient d'indétermination.

## Modèle de régression linéaire

- Coefficient de détermination :  $r^2$



# Modèle de régression linéaire

## Analyse de la « variance » sous forme de tableau ANOVA

Source de variance	« Variance » (somme des carrés, SS)	d.d.l.	Vraie variance (moyenne des carrés, MS)	Rapport F
Expliquée (par la régression)	$\sum(\hat{Y} - \bar{Y})^2$ ou $b^2 \sum x^2$	1	$\frac{b^2 \sum x^2}{1}$	$\frac{b^2 \sum x^2}{s^2}$
Non expliquée (résiduelle)	$\sum(Y - \hat{Y})^2$	$n - 2$	$s^2 = \frac{\sum(Y - \hat{Y})^2}{n - 2}$	
Total	$\sum(Y - \bar{Y})^2$	$n - 1$		

Tableau 11.2

Hypothèse nulle:  $\beta = 0$  (absence de relation linéaire entre X et Y)

Test classique de l'ANOVA:

F suffisamment grand pour  $H_0$  ?

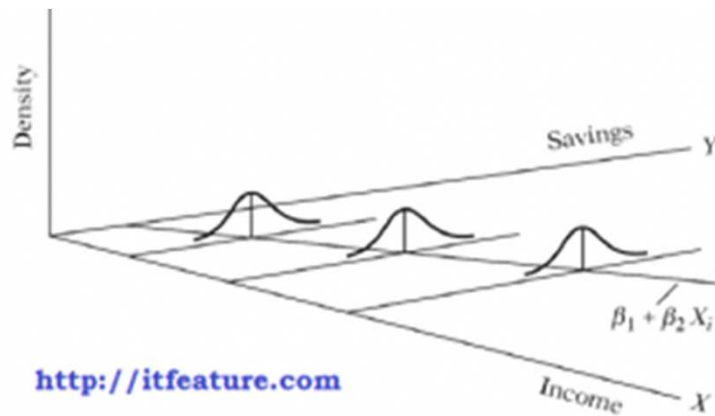
$$F = \frac{\text{variance expliquée par la régression}}{\text{variance inexpliquée}} = \frac{b^2 \sum x^2}{s^2}$$

## Conditions d'application de la régression (I)

- Les observations sont indépendantes entre elles
  - Les scores d'un sujet  $i$  n'influencent pas/ne sont pas influencés par les scores d'un sujet  $j$

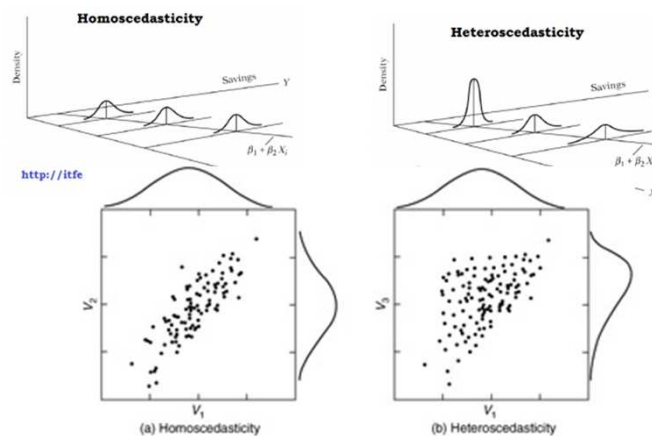
## Conditions d'application de la régression (II)

- Résidus normalement distribués



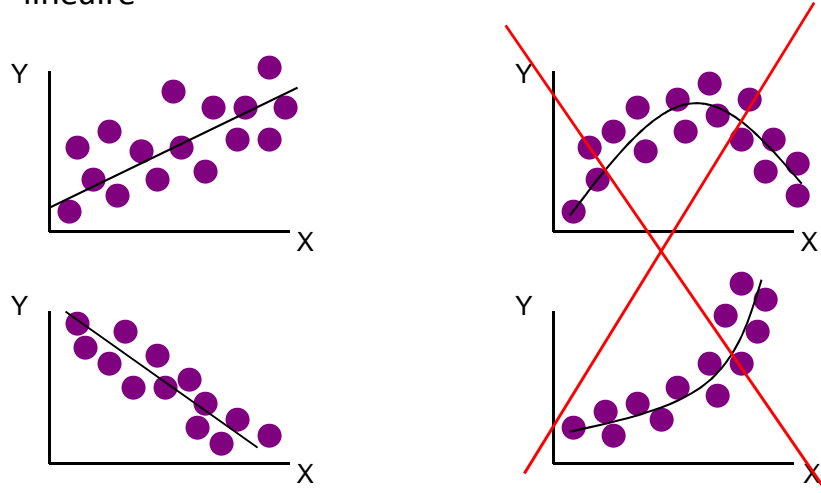
## Conditions d'application de la régression (II)

- La variance des résidus est égale pour toutes les valeurs de X
  - homoscedasticité



## Conditions d'application de la régression (IV)

- Et évidemment, la relation entre les deux variables est linéaire



## Exercice

### Estimation de la droite de régression de Y en X

Parmi tous les points de Y pour  $X = X_1$ , lequel donne la meilleure prédiction ?

Le point situé au milieu du segment ( $P_1$ ).

Si la distribution de Y en X est normale bivariée, alors, on peut montrer que pour toute autre note X, la valeur de Y située au milieu du segment correspondant est la meilleure prédiction et que les points ainsi déterminés sont alignés sur une droite d'équation:

$$Y = \alpha + \beta X$$

Cette droite peut être estimée à partir d'un échantillon sur lequel on fait une régression

$$\hat{Y} = a + bX$$

où, comme démontré précédemment,

$$b = \frac{\sum xy}{\sum x^2} = r \frac{s_Y}{s_X}$$

$$a = \bar{Y} - b\bar{X}$$

# Exercice

## Exercice

A l'aide de l'échantillon des notes de math et à l'oral du tableau 11.1 et des statistiques qui en découlent:

- Déterminer la régression de Y en X et la régression de X en Y. Représenter les 2 droites.
- Pour un étudiant dont la note de math est de X = 90, quelle est la meilleure prédiction de la note d'oral Y ?
- Pour un étudiant dont la note d'oral est de Y = 10, quelle est la meilleure prédiction de la note de math X ?

Solution

- Les calculs dont on a besoin,  $\Sigma xy$ ,  $\Sigma x^2$ , etc... on déjà été effectués dans le tableau 11.1

Il suffit de les introduire dans les formules appropriées.

Pour la droite de régression de Y en X

$$b = \frac{\sum xy}{\sum x^2} = \frac{654}{1304}$$

$$\text{donc: } \hat{Y} = a + bX = 20 + 0,50X$$

$$a = \bar{Y} - b\bar{X} = 50 - 0,50(60) = 20$$

# Exercice

Nous avons, pour 8 étudiants, la note en math (X) et la note à l'oral (Y):

Données		Ecart à la moyenne		Produits		
X	Y	$x = X - \bar{X}$	$y = Y - \bar{Y}$	xy	$x^2$	$y^2$
80	65	20	15	300	400	225
50	60	-10	10	-100	100	100
36	35	-24	-15	360	576	225
58	39	-2	-11	22	4	121
72	48	12	-2	-24	144	4
60	44	0	-6	0	0	36
56	48	-4	-2	8	16	4
68	61	8	11	88	64	121
$\bar{X} = 60$	$\bar{Y} = 50$	0√	0√	$\Sigma xy = 654$	$\Sigma x^2 = 1\ 304$	$\Sigma y^2 = 836$

Tableau 11.1

Le coefficient de régression b est:  $b = \frac{\sum xy}{\sum x^2} = \frac{654}{1304} = 0,50$  pour prédire Y à partir de X

Le coefficient de corrélation r est:  $r = \frac{\sum xy}{\sqrt{\sum x^2} \sqrt{\sum y^2}} = \frac{654}{\sqrt{1304} \sqrt{836}} = 0,63$   
intensité du lien linéaire