ULB UNIVERSITÉ LIBRE DE BRUXELLES

Plan du cours

1. Introduction
2. Statistique descriptive – séries univariées
3. Calcul des probabilités
4. Variables aléatoires et lois de probabilité
5. Arbres de décision
6. Statistique descriptive – séries bivariées
7. Méthodes de prévision

ULB UNIVERSITÉ LIBRE DE BRUXELLES

Séries bivariées

- Objectif : étudier les liens entre deux variables x et y sur base d'une série d'observations.
- Plusieurs cas possibles :
  - 2 variables quantitatives,
  - 2 variables ordinales (ou quantitatives),
  - 2 variables nominales (ou ord. ou quant.).
- Généralisation à plus de 2 variables.

ULB UNIVERSITÉ LIBRE DE BRUXELLES

Cas 1 : variables quantitatives

- Tableau individus × caractères

	x	y
1	x_1	y_1
⋮	⋮	⋮
i	x_i	y_i
⋮	⋮	⋮
n	x_n	y_n

2 caractères

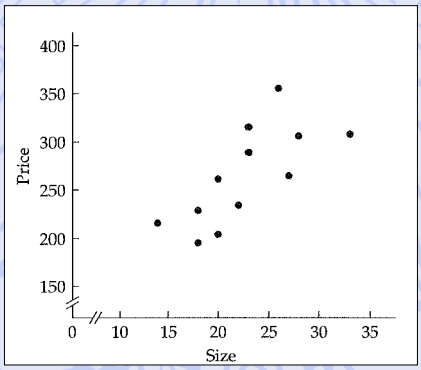
n individus

ULB UNIVERSITÉ LIBRE DE BRUXELLES

Exemple 1

- Lien entre le prix de vente d'une maison et sa surface habitable

Size	*Price* ($000s)
23	315
18	229
26	355
20	261
22	234
14	216
33	308
28	306
23	289
20	204
27	265
18	195



Exemple 2

- Série bivariée de 10 observations

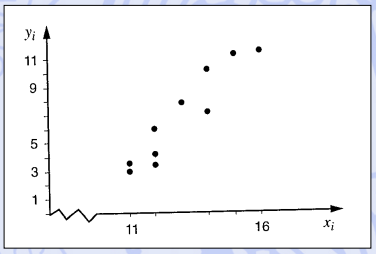
x_i	12	12	15	14	16	14	12	13	11	11
y_i	4.1	3.4	11.3	10.2	11.5	7.2	6.0	7.8	3.5	3.0

Représentation graphique

- Série bivariée :

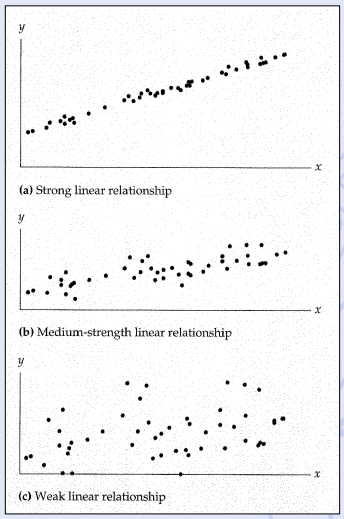
$$\{(x_i, y_i); i = 1, 2, \ldots, n\}$$

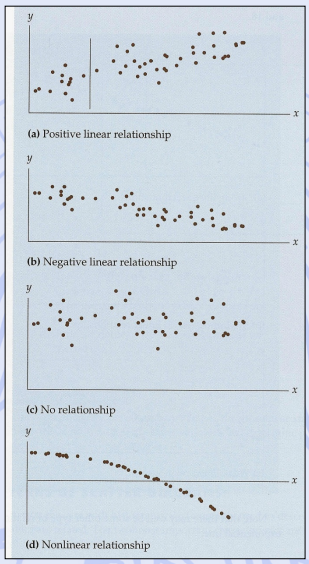
- Représentation graphique – scatterplot :
  - Nuage de points dans le plan (x,y).
  - Exemple 2 :



ULB UNIVERSITÉ LIBRE DE BRUXELLES

Représentation graphique





ULB UNIVERSITÉ LIBRE DE BRUXELLES

Séries marginales

- Série marginale en x :

$$\{x_i; i = 1, \ldots, n\}$$

$$\Rightarrow \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad \cdots$$

- Série marginale en y :

$$\{y_i; i = 1, \ldots, n\}$$

$$\Rightarrow \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad s_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 \quad \cdots$$

ULB UNIVERSITÉ LIBRE DE BRUXELLES

Distribution observée à 2 dimensions

- Lorsque les mêmes valeurs de x et de y se présentent plusieurs fois :

$$\{(x_j, y_k, n_{jk}); j = 1,\ldots,J; k = 1,\ldots,K\}$$

- Effectifs :

$$n_{jk} = \text{effectif associé à } (x_j, y_k)$$

- Propriété :

$$\sum_{j=1}^{J}\sum_{k=1}^{K} n_{jk} = n$$

ULB UNIVERSITÉ LIBRE DE BRUXELLES

Tableau de contingence

x \ y	y_1	$\cdots$	y_k	$\cdots$	y_K
x_1	n_{11}	$\cdots$	n_{1k}	$\cdots$	n_{1K}
$\vdots$	$\vdots$		$\vdots$		$\vdots$
x_j	n_{j1}	$\cdots$	n_{jk}	$\cdots$	n_{jK}
$\vdots$	$\vdots$		$\vdots$		$\vdots$
x_J	n_{J1}	$\cdots$	n_{Jk}	$\cdots$	n_{JK}

Exemple 3

- 80 femmes ayant eu au moins un enfant,
  - x : nombre d'enfants,
  - y : nombre de frères et sœurs.

x_j	y_k 0	1	2	3	4
1	4	4	2	0	0
2	9	16	4	0	0
3	4	12	9	2	0
4	1	6	1	1	2
5	0	1	0	1	1

Distributions marginales

- Distribution marginale en x :

$$\{(x_j, n_{j.}); j = 1,\ldots,J\} \qquad n_{j.} = \sum_{k=1}^{K} n_{jk}$$

- Distribution marginale en y :

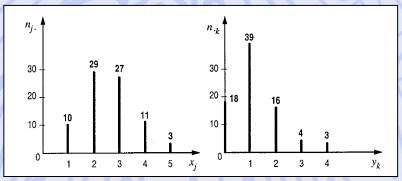
$$\{(y_k, n_{.k}); k = 1,\ldots,K\} \qquad n_{.k} = \sum_{j=1}^{J} n_{jk}$$

- Propriété :

$$\sum_{j=1}^{J} n_{j.} = \sum_{k=1}^{K} n_{.k} = \sum_{j=1}^{J}\sum_{k=1}^{K} n_{jk} = n$$

Exemple 3

	y_k					
x_j	0	1	2	3	4	$n_{j.}$
1	4	4	2	0	0	10
2	9	16	4	0	0	29
3	4	12	9	2	0	27
4	1	6	1	1	2	11
5	0	1	0	1	1	3
$n_{.k}$	18	39	16	4	3	



Distributions marginales

- Fréquences marginales :

$$f_{j.} = \frac{n_{j.}}{n} \qquad f_{.k} = \frac{n_{.k}}{n}$$

- Paramètres :

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{J} n_{j.} x_j \qquad s_x^2 = \frac{1}{n}\sum_{j=1}^{J} n_{j.} (x_j - \bar{x})^2$$

$$\bar{y} = \frac{1}{n}\sum_{k=1}^{K} n_{.k} y_k \qquad s_y^2 = \frac{1}{n}\sum_{k=1}^{K} n_{.k} (y_k - \bar{y})^2$$

Distributions conditionnelles

- Distribution d'une des variables lorsque la valeur de l'autre est fixée.
- Distribution conditionnelle de y en x :
  - On fixe la valeur de x : $x = x_j$

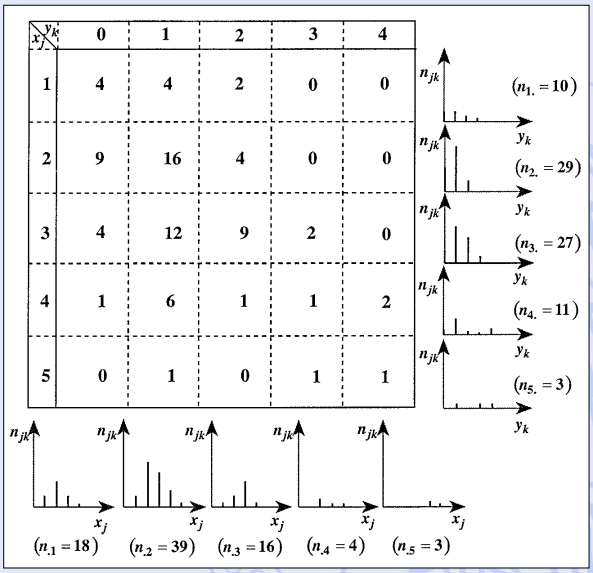
$$\{(y_k, n_{jk}); k = 1, \ldots, K\}$$

  - Effectif total : $n_{j.}$
- Distribution conditionnelle de x en y :
  - On fixe la valeur de y : $y = y_k$

$$\{(x_j, n_{jk}); j = 1, \ldots, J\}$$

  - Effectif total : $n_{.k}$

Exemple 3

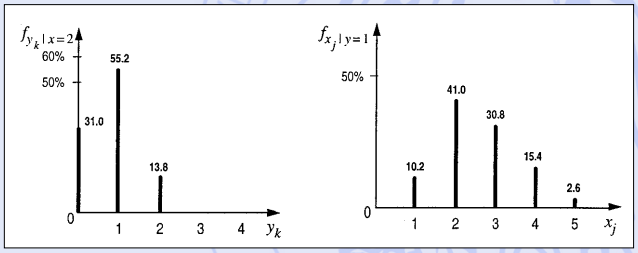


Distributions conditionnelles

- Fréquences conditionnelles :

$$f_{y_k|x_j} = f_{k|j} = \frac{n_{jk}}{n_{j.}} \qquad f_{x_j|y_k} = f_{j|k} = \frac{n_{jk}}{n_{.k}}$$

- Exemple 3 :



Moyennes et variances conditionnelles

$$\bar{y}_{|x_j} = \frac{1}{n_{j.}} \sum_{k=1}^{K} n_{jk} y_k$$

$$s^2_{y|x_j} = \frac{1}{n_{j.}} \sum_{k=1}^{K} n_{jk} \left(y_k - \bar{y}_{|x_j} \right)^2$$

ULB UNIVERSITÉ LIBRE DE BRUXELLES

Moments

- Moments centrés :

$$m_{rs} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^r (y_i - \bar{y})^s \qquad r,s \in N$$

- Cas particuliers : $m_{20} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = s_x^2$

$$m_{02} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = s_y^2$$

Covariance ⟶ $m_{11} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = s_{xy}$

ULB UNIVERSITÉ LIBRE DE BRUXELLES

Moments

- Moments par rapport à l'origine :

$$m'_{rs} = \frac{1}{n}\sum_{i=1}^{n} x_i^r y_i^s$$

- Cas particuliers :

$$m'_{10} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x} \qquad m'_{01} = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}$$

Propriétés des moments

- Changements d'origine et d'unité :

$$u_i = \frac{x_i - x_0}{d_x} \qquad v_i = \frac{y_i - y_0}{d_y} \qquad x_0, y_0 \in R; d_x, d_y \in R_0^+$$

$$\tilde{m}_{rs} = \frac{1}{n}\sum_{i=1}^{n}(u_i - \bar{u})^r (v_i - \bar{v})^s$$

$$\Rightarrow \quad \tilde{m}_{rs} = \frac{m_{rs}}{d_x^r d_y^s}$$

Propriétés des moments

- Cas particulier :

$$s_{uv} = \frac{s_{xy}}{d_x d_y} \quad \Rightarrow \quad s_{xy} = s_{uv} d_x d_y$$

- Calcul de la covariance (cf. variance) :

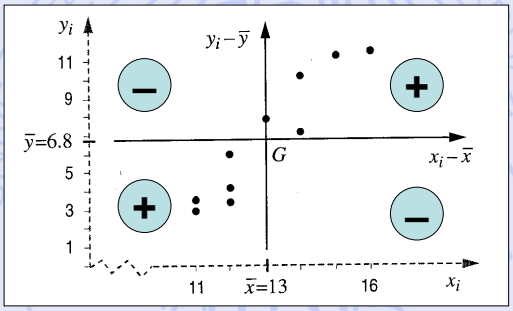
$$m_{11} = s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

$$= \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\,\bar{y} = m'_{11} - m'_{10} m'_{01}$$

ULB UNIVERSITÉ LIBRE DE BRUXELLES

Covariance

- Permet de détecter une relation linéaire entre x et y :
- Positive si relation croissante,
- Négative si relation décroissante.



ULB UNIVERSITÉ LIBRE DE BRUXELLES

Covariance

- Propriétés :
  - Indépendante de changements d'origine.
  - Lien avec les variances marginales :

$$|s_{xy}| \leq s_x s_y$$

ULB UNIVERSITÉ LIBRE DE BRUXELLES

Coefficient de corrélation

- Définition :

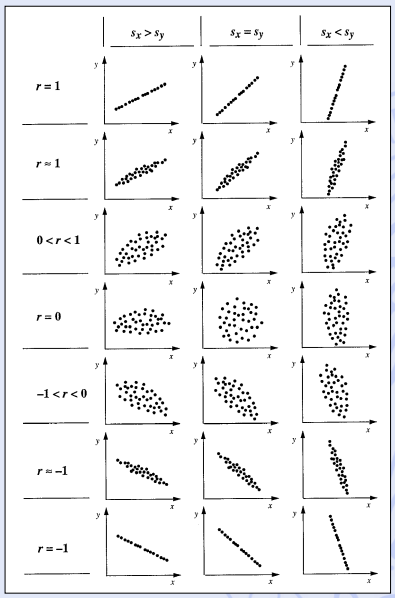
$$r = \frac{S_{xy}}{S_x S_y}$$

- Propriétés :
  - Indépendant des changements d'origine ET d'unité.
  - Même signe que la covariance.
  - Compris entre -1 et +1.
  - Mesure l'intensité de la relation linéaire entre x et y.

246

ULB UNIVERSITÉ LIBRE DE BRUXELLES

Coefficient de corrélation



247

ULB UNIVERSITÉ LIBRE DE BRUXELLES

Notation matricielle

- Vecteur moyenne (centre de gravité) :

$$\mathbf{g} = \begin{pmatrix} \bar{x} \\ \bar{y} \end{pmatrix}$$

- Matrice de variance-covariance :

$$\mathbf{V} = \begin{pmatrix} s_x^2 & s_{xy} \\ s_{xy} & s_y^2 \end{pmatrix}$$

  - Matrice symétrique

ULB UNIVERSITÉ LIBRE DE BRUXELLES

Notation matricielle

- Matrice définie par le tableau des données :

$$\mathbf{X} = \begin{pmatrix} x_1 & y_1 \\ x_2 & y_2 \\ \vdots & \vdots \\ x_n & y_n \end{pmatrix}$$

- Matrice des valeurs centrées :

$$\mathbf{X}_c = \begin{pmatrix} x_1 - \bar{x} & y_1 - \bar{y} \\ x_2 - \bar{x} & y_2 - \bar{y} \\ \vdots & \vdots \\ x_n - \bar{x} & y_n - \bar{y} \end{pmatrix}$$

Notation matricielle

- Propriété :

$$\mathbf{V} = \frac{1}{n}\mathbf{X}_c'\mathbf{X}_c$$

$$= \frac{1}{n}\begin{pmatrix} x_1-\bar{x} & x_2-\bar{x} & \cdots & x_n-\bar{x} \\ y_1-\bar{y} & y_2-\bar{y} & \cdots & y_n-\bar{y} \end{pmatrix}\begin{pmatrix} x_1-\bar{x} & y_1-\bar{y} \\ x_2-\bar{x} & y_2-\bar{y} \\ \vdots & \vdots \\ x_n-\bar{x} & y_n-\bar{y} \end{pmatrix}$$

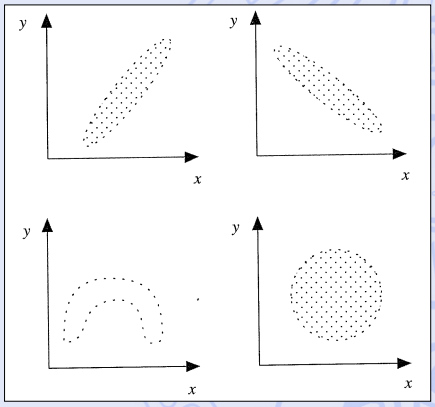
Régression linéaire

- Objectif : Déterminer une relation de dépendance linéaire entre une variable y (variable dépendante) et une ou plusieurs variables explicatives.
  - Régression linéaire simple : une seule variable explicative (x),
  - Régression linéaire multiple : plusieurs variables explicatives ($x_1, x_2, \ldots, x_p$).

ULB UNIVERSITÉ LIBRE DE BRUXELLES

Régression linéaire simple

- Attention : dépendance **linéaire** !
- Différents types de dépendance :

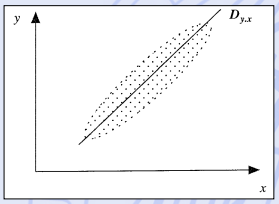


ULB UNIVERSITÉ LIBRE DE BRUXELLES

Droite de régression de y en x

- Equation :

$$D_{y.x} : \quad y = a + bx \quad a,b \in \tilde{}$$



- Problème : déterminer a et b à partir d'une série bivariée :

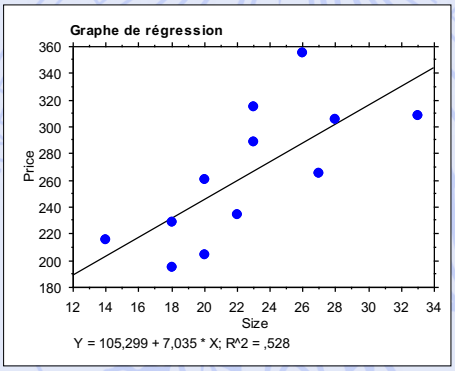
$$\{(x_i, y_i); i = 1,2,\ldots,n\}$$

ULB UNIVERSITÉ LIBRE DE BRUXELLES

Exemple 1

- Lien entre le prix de vente d'une maison et sa surface habitable

Size	*Price* ($000s)
23	315
18	229
26	355
20	261
22	234
14	216
33	308
28	306
23	289
20	204
27	265
18	195



ULB UNIVERSITÉ LIBRE DE BRUXELLES

Résidus

- Définition :

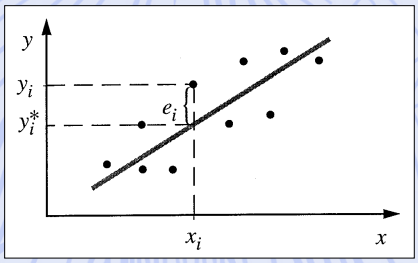
Valeurs ajustées

$$y_i^* = a + bx_i$$

$$\Downarrow$$

$$e_i = y_i - y_i^* = y_i - a - bx_i$$

Résidus



ULB UNIVERSITÉ LIBRE DE BRUXELLES

Principe des moindres carrés

- Idée : choisir a et b de façon à rendre les résidus les plus petits possible.
- Minimiser $\sum_{i=1}^{n} |e_i|$?

  Peu pratique d'un point de vue mathématique.
- Minimiser $\sum_{i=1}^{n} e_i^2$

  → Principe des moindres carrés !

ULB UNIVERSITÉ LIBRE DE BRUXELLES

Principe des moindres carrés

- Objectif :

$$\underset{a,b}{\text{Min}} \quad Q(a,b) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

- Solution :

$$\frac{\partial Q(a,b)}{\partial a} = 0 \qquad \frac{\partial Q(a,b)}{\partial b} = 0$$

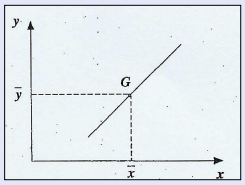
ULB UNIVERSITÉ LIBRE DE BRUXELLES

Calcul

- 1ère dérivée partielle :

$$\frac{\partial Q(a,b)}{\partial a} = 0 \quad \Leftrightarrow \quad 2\times(-1)\times\sum_{i=1}^{n}(y_i - a - bx_i) = 0$$

$$\Leftrightarrow \quad \sum_{i=1}^{n}(y_i - a - bx_i) = 0$$



$$\Leftrightarrow \quad \sum_{i=1}^{n} y_i - na - b\sum_{i=1}^{n} x_i = 0$$

$$\Leftrightarrow \quad \sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i$$

$$\Leftrightarrow \quad \bar{y} = a + b\bar{x}$$

ULB UNIVERSITÉ LIBRE DE BRUXELLES

Calcul

- 2ème dérivée partielle :

$$\frac{\partial Q}{\partial b} = 0 \quad \Leftrightarrow \quad 2\times(-1)\sum_{i=1}^{n}(y_i - a - bx_i)x_i = 0$$

$$\Leftrightarrow \quad \sum_{i=1}^{n} x_i y_i - a\sum_{i=1}^{n} x_i - b\sum_{i=1}^{n} x_i^2 = 0$$

$$\Leftrightarrow \quad \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \underbrace{(\bar{y} - b\bar{x})}_{a}\bar{x} - b\frac{1}{n}\sum_{i=1}^{n} x_i^2 = 0$$

$$\Leftrightarrow \quad \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\,\bar{y}\right)}_{s_{xy}} - b\underbrace{\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2\right)}_{s_x^2} = 0$$

ULB UNIVERSITÉ LIBRE DE BRUXELLES

Solution

$$a = \bar{y} - b\bar{x} \qquad b = \frac{s_{xy}}{s_x^2}$$

- Remarques :
  - Il s'agit bien d'un minimum (Cf. dérivées secondes).
  - Autres expressions de la droite de régression :

$$y = \bar{y} + b(x - \bar{x})$$

$$y - \bar{y} = \frac{s_{xy}}{s_x^2}(x - \bar{x})$$

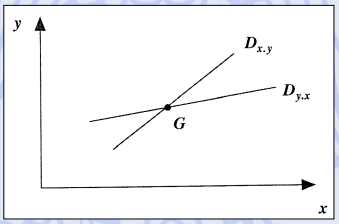
ULB UNIVERSITÉ LIBRE DE BRUXELLES

2 droites de régression !

- Régression de y en x :
  - Coefficient de régression de y en x : $b = \frac{s_{xy}}{s_x^2} = b_{y.x}$
- Régression de x en y :
  - Coefficient de régression de x en y : $b_{x.y} = \frac{s_{xy}}{s_y^2}$

$$D_{x.y} : \quad x = \bar{x} + \frac{s_{xy}}{s_y^2}(y - \bar{y})$$

$$\Leftrightarrow \quad y = \bar{y} + \frac{s_y^2}{s_{xy}}(x - \bar{x})$$



Variances résiduelle et de régression

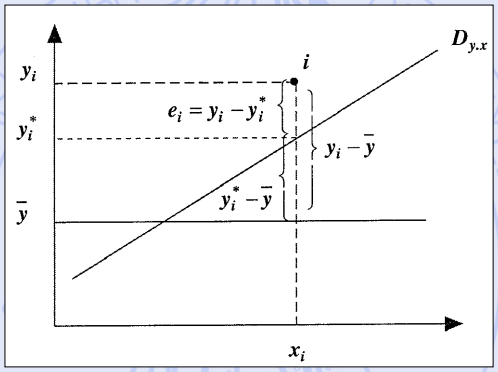
- Décomposition de la variance de y :

$$s_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 =$$

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n}(y_i - y_i^*)^2}_{s^2_{y.x}} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}(y_i^* - \bar{y})^2}_{s^2_{reg}}$$

Variance résiduelle

Variance de régression



Démonstration

$$s_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n}\sum_{i=1}^{n}(\underbrace{y_i - y_i^*}_{} + \underbrace{y_i^* - \bar{y}}_{})^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i^*)^2 + \frac{1}{n}\sum_{i=1}^{n}(y_i^* - \bar{y})^2$$

$$+ \underbrace{\frac{2}{n}\sum_{i=1}^{n}(y_i - y_i^*)(y_i^* - \bar{y})}_{=0 \quad \text{(à démontrer)}}$$

Rappel : $y_i^* = \bar{y} + b_{y.x}(x_i - \bar{x})$

ULB UNIVERSITÉ LIBRE DE BRUXELLES

Corrélation et régression

- Lien entre r et le coefficient de régression :

$$b = \frac{s_{xy}}{s_x^2} = b_{y.x} \quad \Rightarrow \quad b_{y.x} = r\frac{s_y}{s_x}$$

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

264

ULB UNIVERSITÉ LIBRE DE BRUXELLES

Interprétation de r

$$\begin{aligned}
s_{y.x}^2 &= \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i^*)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y} - b_{y.x}(x_i - \bar{x}))^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 + b_{y.x}^2\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\
&\quad -2b_{y.x}\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \\
&= s_y^2 + \frac{s_{xy}^2}{s_x^4}s_x^2 - 2\frac{s_{xy}}{s_x^2}s_{xy} = s_y^2 - \frac{s_{xy}^2}{s_x^2}\frac{s_y^2}{s_x^2} \\
&= s_y^2\left(1 - \frac{s_{xy}^2}{s_x^2 s_y^2}\right) = s_y^2(1 - r^2)
\end{aligned}$$

265

Interprétation de r

- Variance résiduelle :

$$s^2_{y.x} = s^2_y\left(1-r^2\right)$$

- Variance de régression :

$$s^2_{reg} = s^2_y - s^2_{y.x} = s^2_y r^2$$

- Coefficient de détermination :

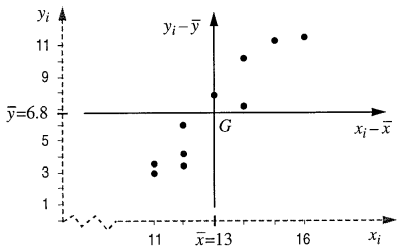
$$r^2 = \frac{s^2_{reg}}{s^2_y}$$

  - Pourcentage de la variance de y expliqué par x.

Exemple 4

- Argent de poche hebdomadaire (y) en fonction de l'âge (x) – 14 observations (ados)

x_i	12	12	15	14	16	14	12	13	11	11
y_i	4.1	3.4	11.3	10.2	11.5	7.2	6.0	7.8	3.5	3.0



Exemple 4

x_i	y_i	$x_i-\bar{x}$	$y_i-\bar{y}$	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$	$(x_i-\bar{x})(y_i-\bar{y})$
12	4.1	- 1	- 2.7	1	7.29	2.7
12	3.4	- 1	- 3.4	1	11.56	3.4
15	11.3	2	4.5	4	20.25	9.0
14	10.2	1	3.4	1	11.56	3.4
16	11.5	3	4.7	9	22.09	14.1
14	7.2	1	0.4	1	0.16	0.4
12	6.0	- 1	- 0.8	1	0.64	0.8
13	7.8	0	1.0	0	1.00	0
11	3.5	- 2	- 3.3	4	10.89	6.6
11	3.0	- 2	- 3.8	4	14.44	7.6
130	68.0	0	0	26	99.88	48.0

$$g=\begin{pmatrix}\bar{x}\\ \bar{y}\end{pmatrix}=\begin{pmatrix}13\\ 6{,}8\end{pmatrix}$$

$$V=\begin{pmatrix}s_x^2 & s_{xy}\\ s_{xy} & s_y^2\end{pmatrix}=\begin{pmatrix}2{,}6 & 4{,}8\\ 4{,}8 & 9{,}99\end{pmatrix}$$

r = 0,94
r^2 = 0,89

$$D_{y.x}: \quad y=6{,}8+\frac{4{,}8}{2{,}6}(x-13)=-17{,}2+1{,}85x$$

Exemples d'Ascombe

- Importance de l'examen graphique des données :
  - Forme de la relation entre x et y,
  - Analyse des résidus (valeurs anormales).
- 4 jeux de données donnant lieu à la même droite de régression !

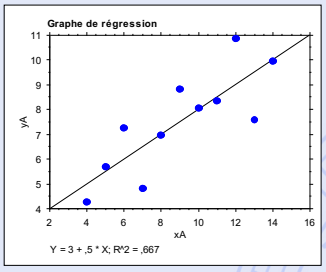
Exemples d'Ascombe

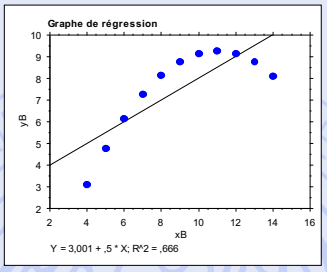
Ensemble A		Ensemble B		Ensemble C		Ensemble D	
x	y	x	y	x	y	x	y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.10	14	8.84	8	7.14
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.10	4	5.39	19	12.50
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

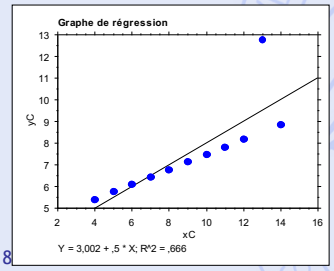
$$n = 11\,;\, \bar{x} = 9\,;\, \bar{y} = 7.5\,;\, s_x^2 = 10\,;\, s_y^2 = 3.75\,;\, s_{xy} = 5$$

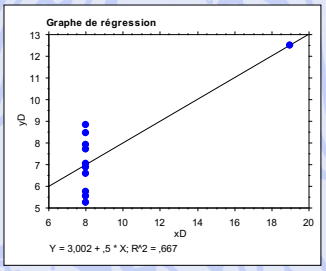
$$r = 0.816\,;\, D_{y.x} \equiv y = 3 + 0.5\,x\,;\, r^2 = 0.667$$

Exemples d'Ascombe



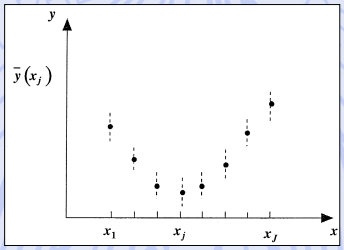






Régression non linéaire

- Dépendance non linéaire ? Forme ?
  - Analyse graphique :



  - Ajustement d'une courbe non linéaire.
  - Mesure de l'intensité de la dépendance non linéaire de y en x.

Ajustements non linéaires

- Ajustements linéarisables :
  - Régression exponentielle

$$y = 10^{a+bx} \quad \Rightarrow \quad \log_{10} y = a + bx$$

Régression linéaire de z en x, avec :

$$z = \log_{10} y \qquad z_i = \log_{10} y_i$$

ULB UNIVERSITÉ LIBRE DE BRUXELLES

Ajustements non linéaires

- Ajustements linéarisables :
  - Régression hyperbolique

$$y = \frac{1}{a+bx} \Rightarrow \frac{1}{y} = a + bx$$

Régression linéaire de z en x, avec :

$$z = \frac{1}{y} \qquad z_i = \frac{1}{y_i}$$

ULB UNIVERSITÉ LIBRE DE BRUXELLES

Ajustement polynomial

$$y = b_0 + b_1 x + b_2 x^2 + \ldots + b_p x^p$$

- Critères des moindres carrés :

$$\text{Min } Q(b_0, b_1, \ldots, b_p) = \sum_{i=1}^{n} e_i^2$$

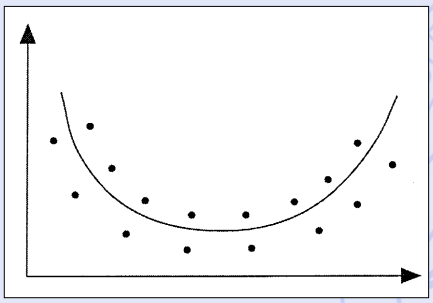
$$= \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i - b_2 x_i^2 - \ldots - b_p x_i^p)^2$$

- Condition nécessaire pour un minimum :

$$\frac{\partial Q}{\partial b_j} = 0 \qquad j = 0, 1, 2, \ldots, p$$

Exemple : parabole

$$y = b_0 + b_1 x + b_2 x^2$$



$$\text{Min } Q(b_0, b_1, b_2) = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i - b_2 x_i^2)^2$$

$$\rightarrow \begin{cases} \frac{\partial Q}{\partial b_0} = -2\sum_{i=1}^{n} (y_i - b_0 - b_1 x_i - b_2 x_i^2) = 0 \\ \frac{\partial Q}{\partial b_1} = -2\sum_{i=1}^{n} (y_i - b_0 - b_1 x_i - b_2 x_i^2) x_i = 0 \\ \frac{\partial Q}{\partial b_2} = -2\sum_{i=1}^{n} (y_i - b_0 - b_1 x_i - b_2 x_i^2) x_i^2 = 0 \end{cases}$$

Rapport de corrélation

- Mesure de l'intensité de la dépendance non linéaire de y en x.
- Décomposition de la variance marginale :

$$\begin{aligned} s_y^2 &= \frac{1}{n}\sum_{k=1}^{K} n_{.k}(y_k - \bar{y})^2 \\ &= \frac{1}{n}\sum_j \sum_k n_{jk}\left(y_k - \bar{y}(x_j) + \bar{y}(x_j) - \bar{y}\right)^2 \\ &= \frac{1}{n}\sum_j \sum_k n_{jk}\left(y_k - \bar{y}(x_j)\right)^2 + \frac{1}{n}\sum_j n_{j.}\left(\bar{y}(x_j) - \bar{y}\right)^2 \\ &= \text{moyenne des variances conditionnelles} \\ &+ \text{variance des moyennes conditionnelles} \end{aligned}$$

ULB UNIVERSITÉ LIBRE DE BRUXELLES

Rapport de corrélation

- Définition :

$$\theta^2_{y.x} = \frac{\frac{1}{n}\sum_j n_{j.}\left(\bar{y}(x_j) - \bar{y}\right)^2}{s^2_y}$$

- A comparer avec :

$$r^2 = \frac{\text{variance de régression}}{\text{variance marginale de } y}$$

ULB UNIVERSITÉ LIBRE DE BRUXELLES

Propriétés

- Indépendant de l'origine et de l'unité.
- $$0 \le \theta^2_{y.x} \le 1$$

$$s^2_y(x_j) = 0,\ j = 1,2,\ldots,J \Rightarrow \theta^2_{y.x} = 1$$

$$\bar{y}(x_j) = \bar{y},\ j = 1,2,\ldots,J \Rightarrow \theta^2_{y.x} = 0$$

$$r^2 \le \theta^2_{y.x}$$

- Indice de non linéarité de la régression :

$$\theta^2_{y.x} - r^2$$

ULB UNIVERSITÉ LIBRE DE BRUXELLES

Cas 2 - Variables ordinales

- Série : 2 variables ordinales

$$\{(x_i, y_i); i = 1, 2, \ldots, n\}$$

- Rangs :

$R(x_i)$ rang de x_i dans $\{x_{(1)}, x_{(2)}, \ldots, x_{(n)}\}$

$R(y_i)$ rang de y_i dans $\{y_{(1)}, y_{(2)}, \ldots, y_{(n)}\}$

ULB UNIVERSITÉ LIBRE DE BRUXELLES

Exemple

- Grades de 5 étudiants en 1ère et 2ème BA (1ère session) :

1ère BA	Sat.	GD	Aj.	LPGD	D
2ème BA	Sat.	LPGD	D	GD	Aj.

- Rangs :

1ère BA	2	4	1	5	3
2ème BA	2	5	3	4	1

ULB UNIVERSITÉ LIBRE DE BRUXELLES

Coefficient de corrélation des rangs de Spearman

- Définition : r_S, coefficient de corrélation (Pearson) calculé sur les rangs.
- Exemple : $\{R(x_i), R(y_i)\} =$

$$\{(2,2),(4,5),(1,3),(5,4),(3,1)\}$$

$$\Rightarrow r_S = 0,50$$

- Propriété : si tous les x_i et y_i sont distincts :

$$r_S = 1 - \frac{6\sum_i [R(x_i) - R(y_i)]^2}{n(n^2-1)}$$

ULB UNIVERSITÉ LIBRE DE BRUXELLES

Cas 3 - Variables nominales

- Tableau de contingence

$$\{(x_j, y_k, n_{jk}); j = 1,\ldots,J; k = 1,\ldots,K\}$$

- Exemple : enquête sur 100 personnes
  - x = sexe, y = attitude vis-à-vis du sport

x_j	y_k P	D	I	$n_{j.}$
F	21	15	9	45
H	39	13	3	55
$n_{.k}$	60	28	12	100

ULB UNIVERSITÉ LIBRE DE BRUXELLES

Profils lignes et profils marginaux

$$f_{k|j} = \frac{n_{jk}}{n_{j.}} \qquad f_{.k} = \frac{n_{.k}}{n} \qquad (j \text{ fixé}; k = 1,\ldots,K)$$

$f_{k\|j}$	P	D	I	
F	0.47	0.33	0.20	1
H	0.71	0.24	0.05	1
$f_{.k}$	0.60	0.28	0.12	1

ULB UNIVERSITÉ LIBRE DE BRUXELLES

Profils colonnes et profils marginaux

$$f_{j|k} = \frac{n_{jk}}{n_{.k}} \qquad f_{j.} = \frac{n_{j.}}{n} \qquad (k \text{ fixé}; j = 1,\ldots,J)$$

$f_{j\|k}$	P	D	I	$f_{j.}$
F	0.35	0.54	0.75	0.45
H	0.65	0.46	0.25	0.55
	1	1	1	1

ULB UNIVERSITÉ LIBRE DE BRUXELLES

Indépendance

- Si les profils lignes et les profils colonnes sont égaux aux profils marginaux.

$$f_{k|j} = f_{.k} \Leftrightarrow \frac{n_{jk}}{n_{j.}} = \frac{n_{.k}}{n}$$

$$f_{j|k} = f_{j.} \Leftrightarrow \frac{n_{jk}}{n_{.k}} = \frac{n_{j.}}{n}$$

- Condition d'indépendance :

$$n_{jk} = \frac{n_{j.} n_{.k}}{n}$$

ULB UNIVERSITÉ LIBRE DE BRUXELLES

Effectifs théoriques

- Définition :

$$n^*_{jk} = \frac{n_{j.} n_{.k}}{n}$$

- Exemple :

n^*_{jk}	P	D	I	
F	27	12.6	5.4	45
H	33	15.4	6.6	55
	60	28	12	100

ULB UNIVERSITÉ LIBRE DE BRUXELLES

Ecart à l'indépendance

- Différence entre effectifs observés et effectifs théoriques :

$$e_{jk} = n_{jk} - n^*_{jk}$$

- Exemple :

e_{jk}	P	D	I
F	- 6	2.4	3.6
H	6	- 2.4	- 3.6

ULB UNIVERSITÉ LIBRE DE BRUXELLES

Mesure du chi-carré (χ^2)

- Mesure de « distance » entre effectifs observés et effectifs théoriques :

$$D^2 = \sum_j \sum_k \frac{\left(n_{jk} - n^*_{jk}\right)^2}{n^*_{jk}} = \sum_j \sum_k \frac{e^2_{jk}}{n^*_{jk}}$$

- Exemple :

e^2_{jk}/n^*_{jk}	P	D	I	
F	1.33	0.46	2.40	4.19
H	1.09	0.37	1.96	3.42
	2.42	0.83	4.36	7.61