Chapitre 4: Séries bivariées

Arnout Van Messem

Bachelier en Sciences informatiques

LIÈGE université

Introduction

Nom	Prov.	Parti	Nombre	IRich.	Chom
Amay	Liège	Ecolo	2	93	16.51
Amel	Liège	Autre	1	85	4.4
Andenne	Namur	PS	2	89	15.79
Anderlues	Hainaut	PS	1	85	20
Anhée	Namur	CdH	1	91	12.33
Ans	Liège	PS	3	96	17.14
Anthisnes	Liège	PS	1	110	8.81
Antoing	Hainaut	PS	1	88	14.49
Arlon	Luxembourg	CdH	2	118	11.14
Assesse	Namur	CdH	3	107	7.99
Ath	Hainaut	PS	2	101	12.33
Attert	Luxembourg	CdH	1	124	6.35
⋮					

LIÈGE université

Séries bivariées

Individus	Variables X	Y
1	x_1	y_1
2	x_2	y_2
$\vdots$	$\vdots$	$\vdots$
i	x_i	y_i
$\vdots$	$\vdots$	$\vdots$
n	x_n	y_n

ou

$$S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}.$$

X et Y peuvent être toutes les deux qualitatives ou toutes les deux quantitatives ou mixtes.

Plan du chapitre

1. Présentation et organisation des données: le tableau de contingence
2. Représentations graphiques de la distribution des effectifs/fréquences et de la série brute
3. Résumés statistiques (covariance et corrélation)
4. Modélisation: droite de régression linéaire

Tableau de contingence

$x_1 < x_2 < \ldots < x_J$: valeurs distinctes observées de la variable X

$y_1 < y_2 < \ldots < y_K$: valeurs distinctes observées de la variable Y

n_{jk} = effectif associé au couple (x_j, y_k)
tels que

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

Tableau de contingence

Valeurs de X	Valeurs de Y				
	y_1	$\dots$	y_k	$\dots$	y_K
x_1	n_{11}	$\dots$	n_{1k}	$\dots$	n_{1K}
$\vdots$					
x_j	n_{j1}	$\dots$	n_{jk}	$\dots$	n_{jK}
$\vdots$					
x_J	n_{J1}	$\dots$	n_{Jk}	$\dots$	n_{JK}

A. Parti – Provinces

Partis	Provinces				
	Brabantwallon	Hainaut	Liège	Lux	Namur
Autre	2	0	7	3	1
CdH	3	13	16	23	12
Ecolo	2	1	1	1	0
MR	14	16	32	10	17
PS	6	39	28	7	8

B. Indice de Richesse groupé – Chômage groupé

	Chômage					
Indice	$[0,5]$	$]5,10]$	$]10,15]$	$]15,20]$	$]20,25]$	$]25,30]$
$]70,80]$	0	0	0	0	1	1
$]80,90]$	3	2	4	2	2	2
$]90,100]$	1	5	16	8	1	0
$]100,110]$	0	16	12	0	0	0
$]110,120]$	0	3	2	0	0	0
$]120,130]$	0	2	1	0	0	0

LIÈGE université

Quelques remarques

- Au lieu d'indiquer les effectifs dans le tableau de contingence, on peut calculer les fréquences des couples observés:

$$f_{jk} = \frac{n_{jk}}{n}$$

vérifiant

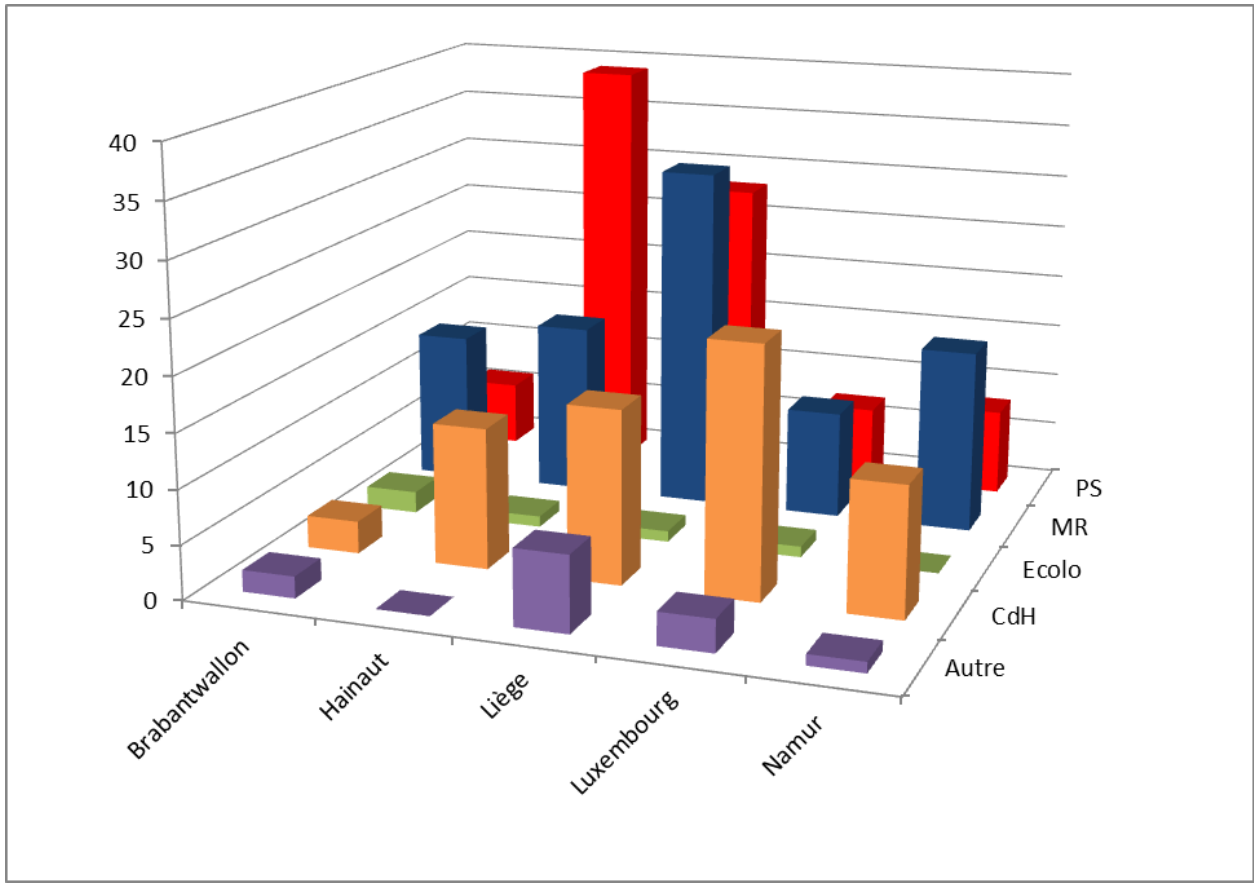
$$\sum_{j=1}^{J} \sum_{k=1}^{K} f_{jk} = 1.$$

- Il n'est plus aussi évident de définir des effectifs et fréquences cumulés dans le contexte bivarié.
- Comme dans le cas univarié, on peut utiliser des diagrammes en barres (cas quali), diagrammes en bâtons (cas quanti discret) et des histogrammes (cas quanti continu) pour représenter la distribution des effectifs (ou des fréquences) bivariés.

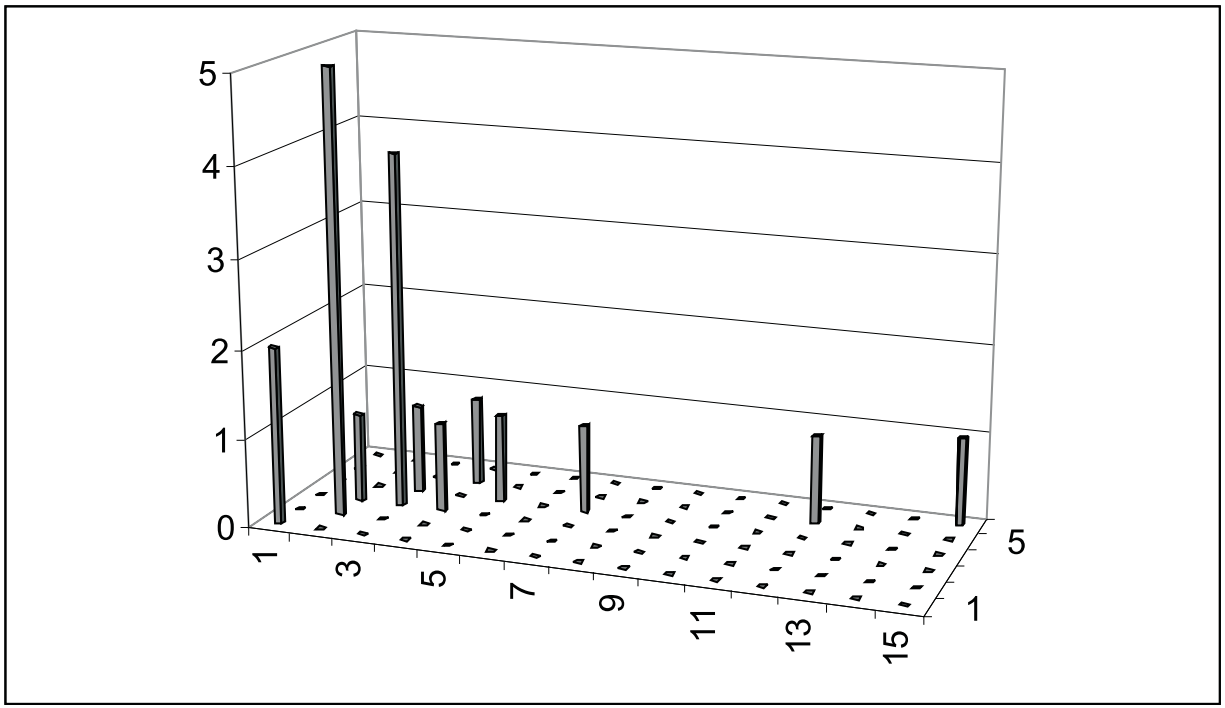
LIÈGE université

Représentations graphiques (du tableau de contingence)

1) Diagramme en barres si X et Y sont qualitatives :



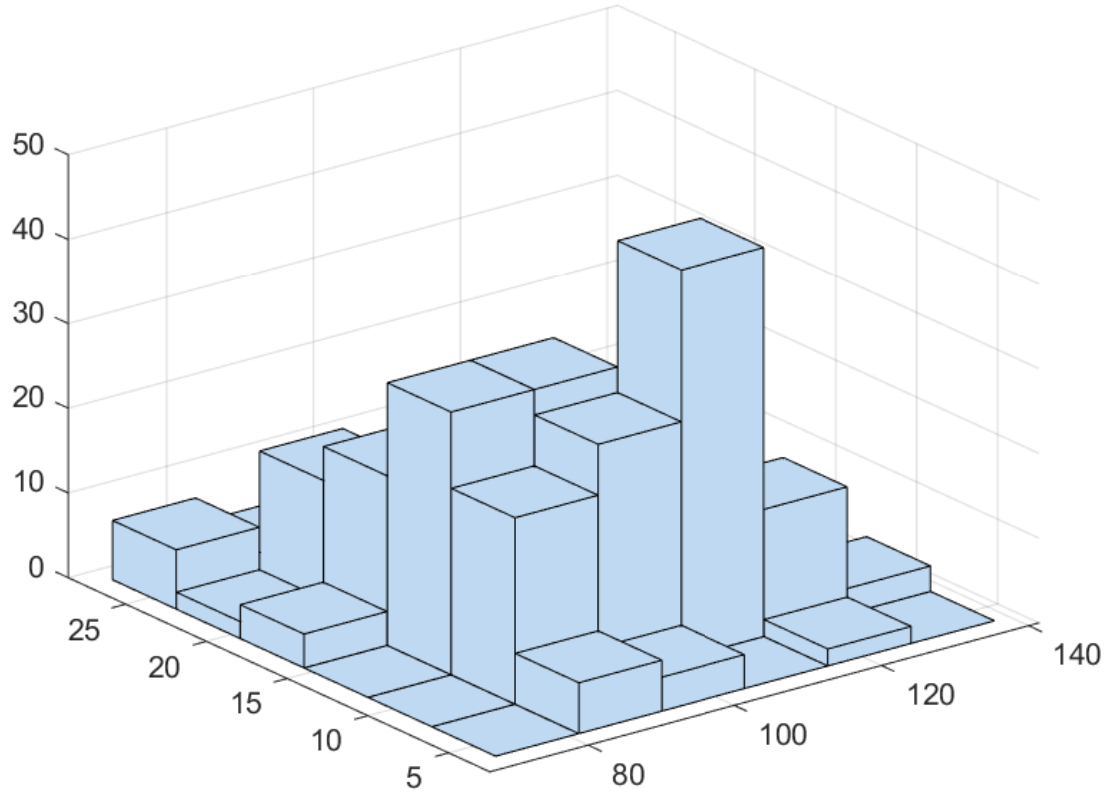
2) Diagramme en bâtons si X et Y sont quantitatives discrètes



X: nombres de personnes d'un service universitaire

Y: nombre de téléphones (fixes) disponibles

3) Histogramme si X et Y sont quantitatives continues réparties en classes (chaque classe est caractérisée par un parallélipipède dont la base est définie par le couple de classes et dont le volume est proportionnel à l'effectif ou à la fréquence du couple de classes).



Représentations graphiques (de la série brute)

4) Boîtes à moustaches si une des variables est quantitative et l'autre qualitative

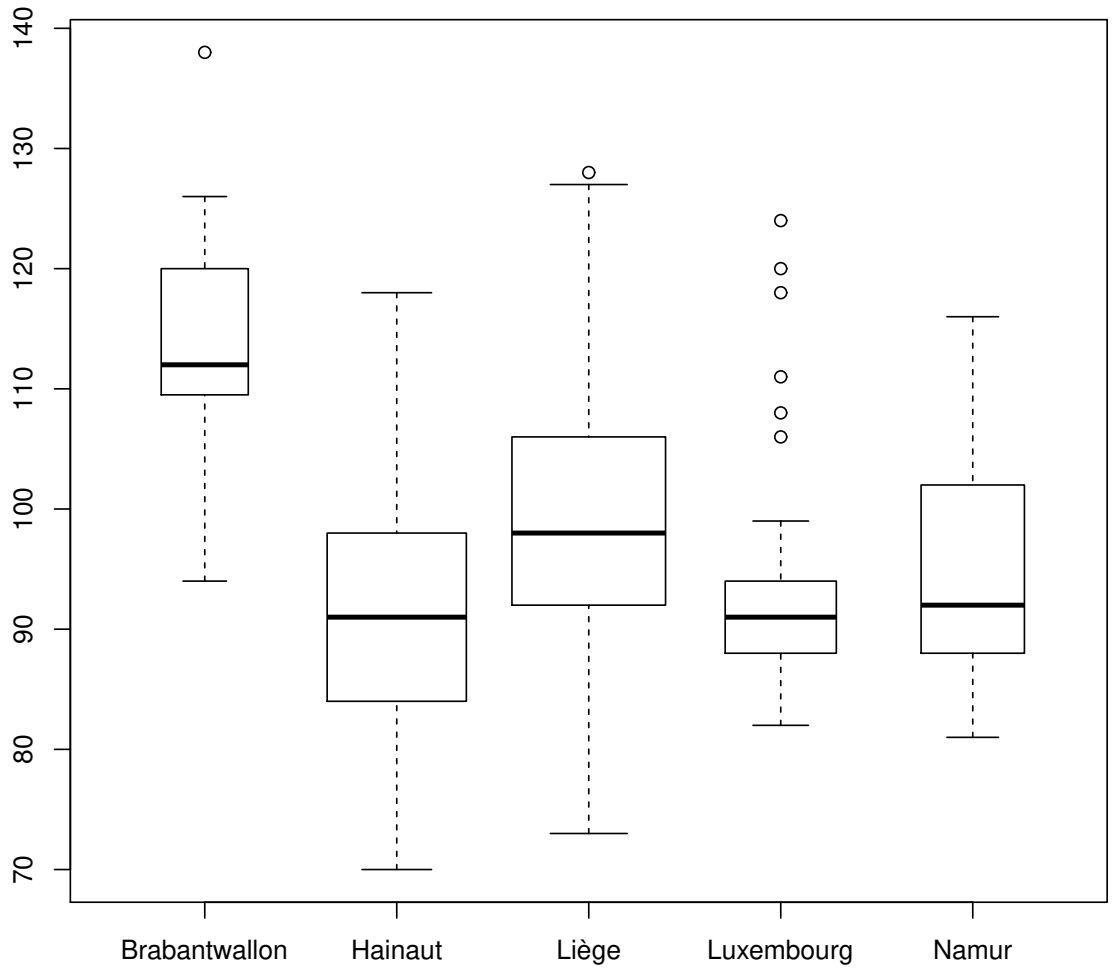
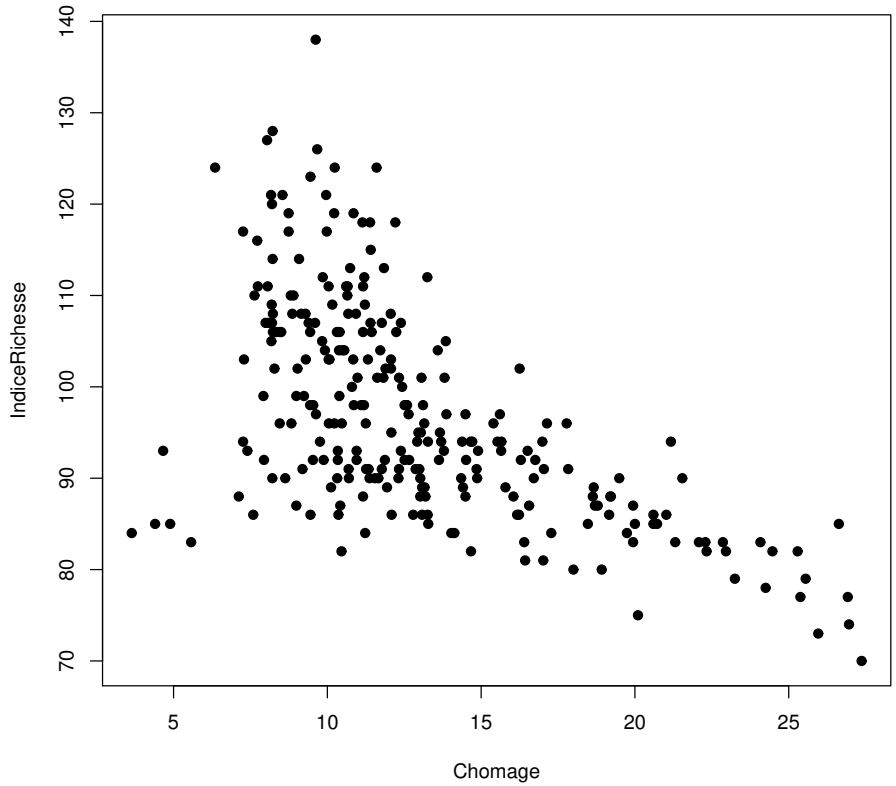
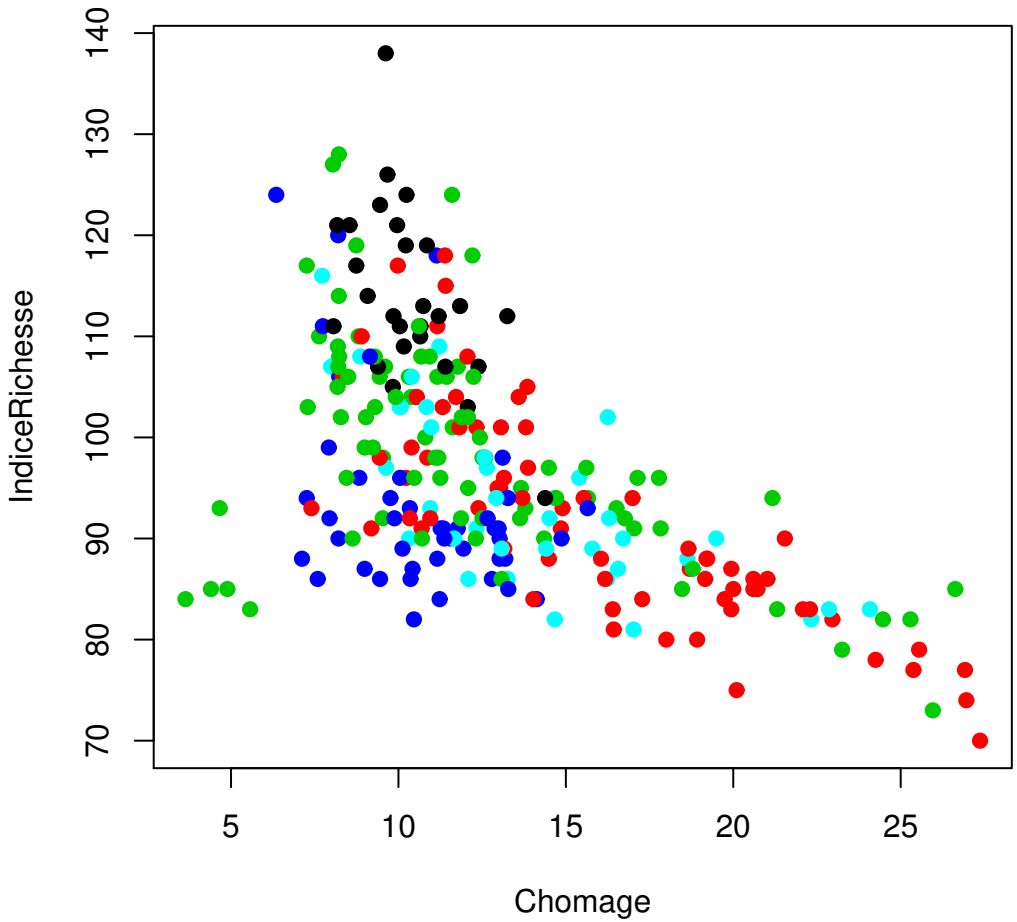


Diagramme de dispersion (repère orthogonal): pour deux variables quantitatives



Avec l'indication des valeurs d'une variable qualitative en plus



Quelques remarques

A. Variables discrètes

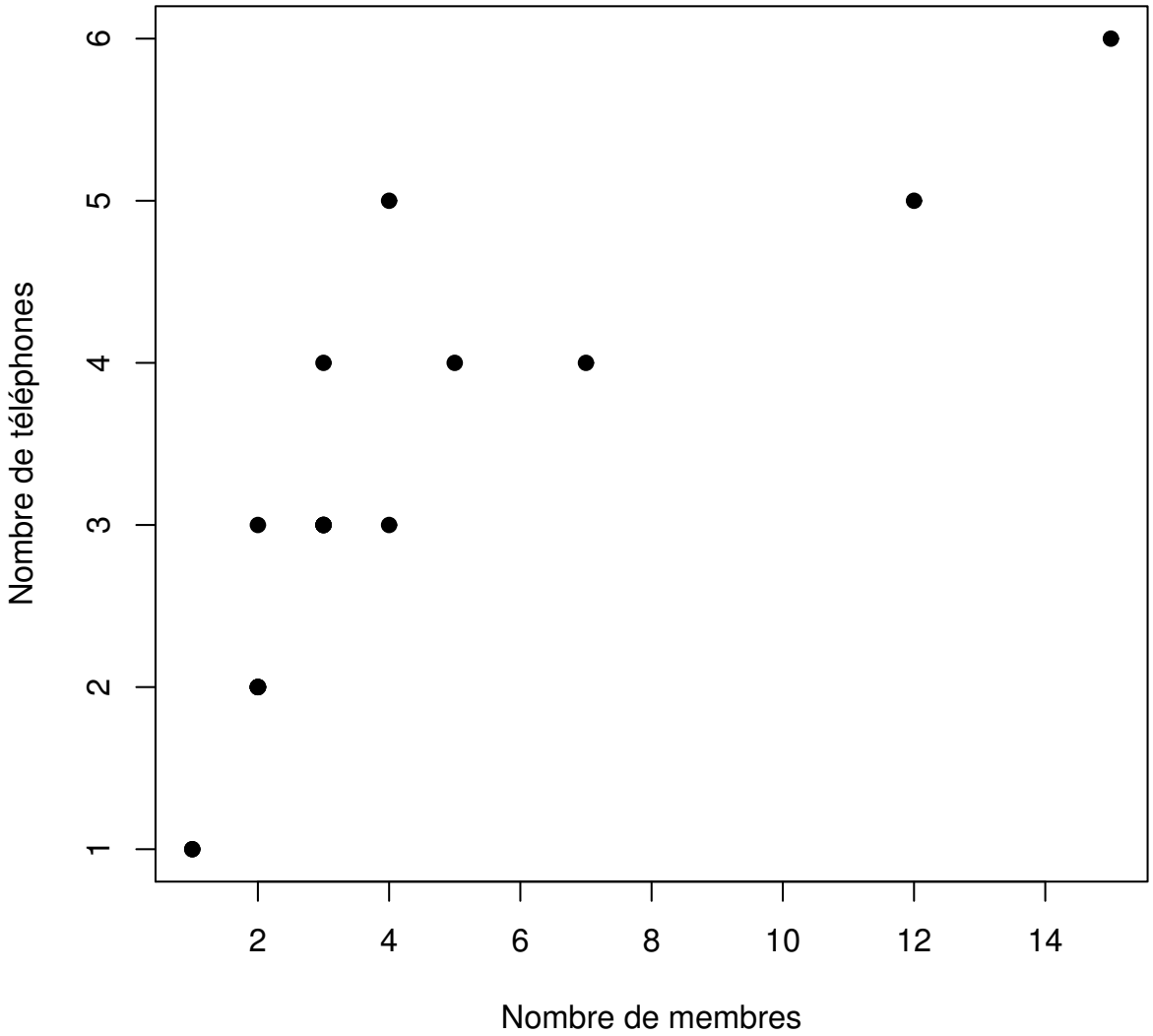
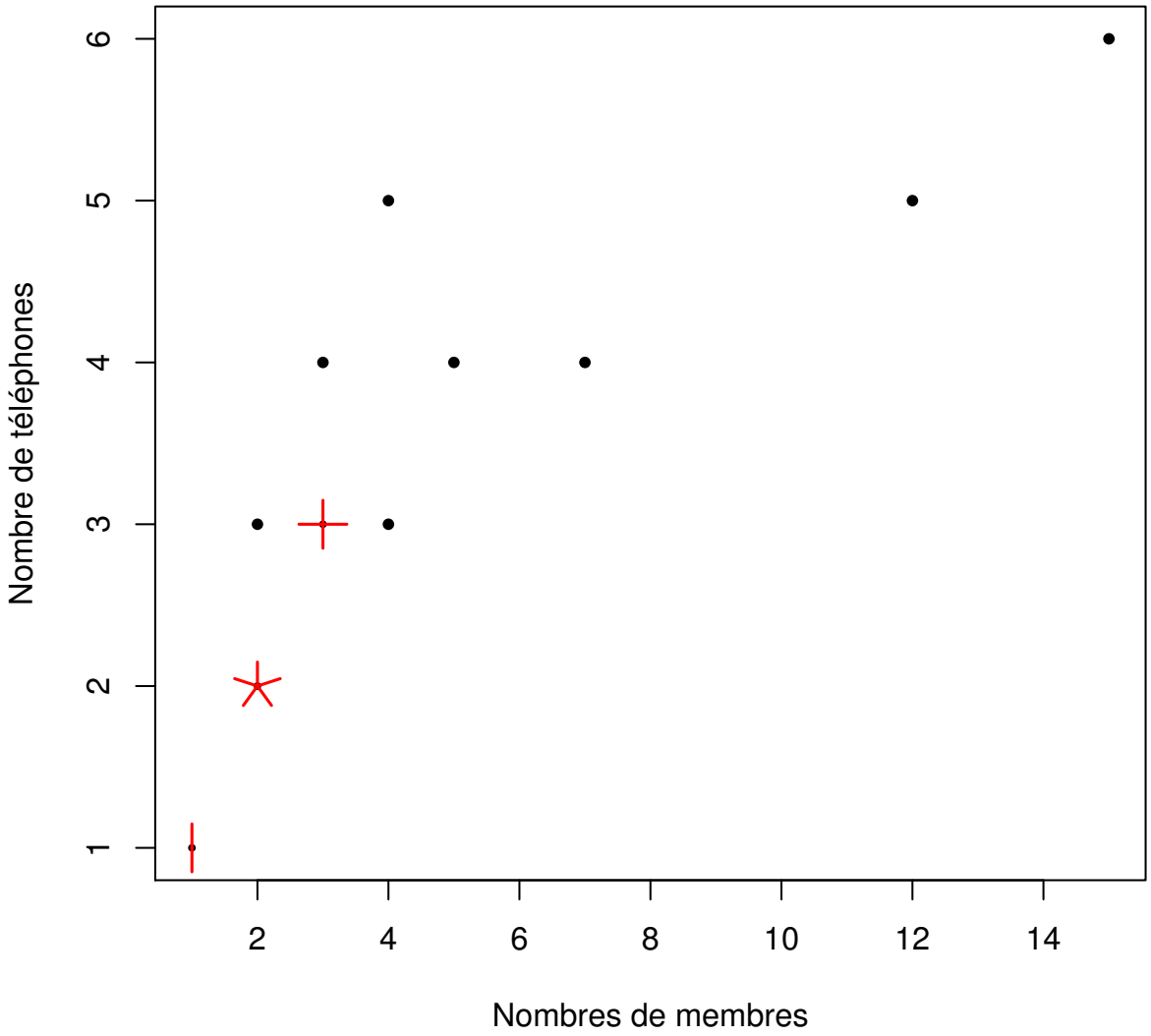


Diagramme de dispersion “sunflower”



B. Différence d'unités sur les axes

Habituellement, lorsque l'on travaille dans le plan réel, on ne se contente pas de prendre des axes orthogonaux mais plutôt des axes orthonormés. Cela signifie que les unités doivent être les mêmes en abscisse et en ordonnée.

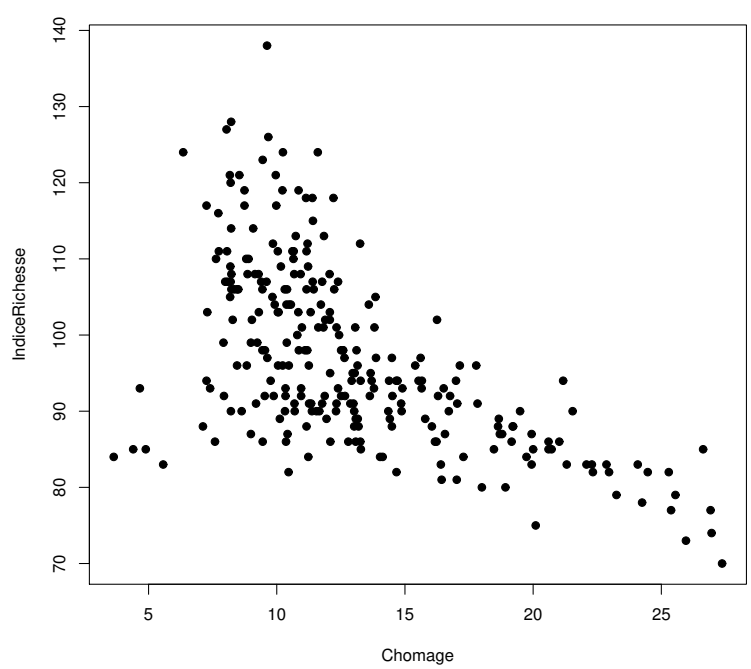
En toute généralité, il n'y a pas de raison pour que les deux variables aient les mêmes unités. Et même dans ce cas-là, le fait que la dispersion soit différente pour X et pour Y implique qu'un même écart, placé horizontalement ou verticalement, n'a pas la même importance.

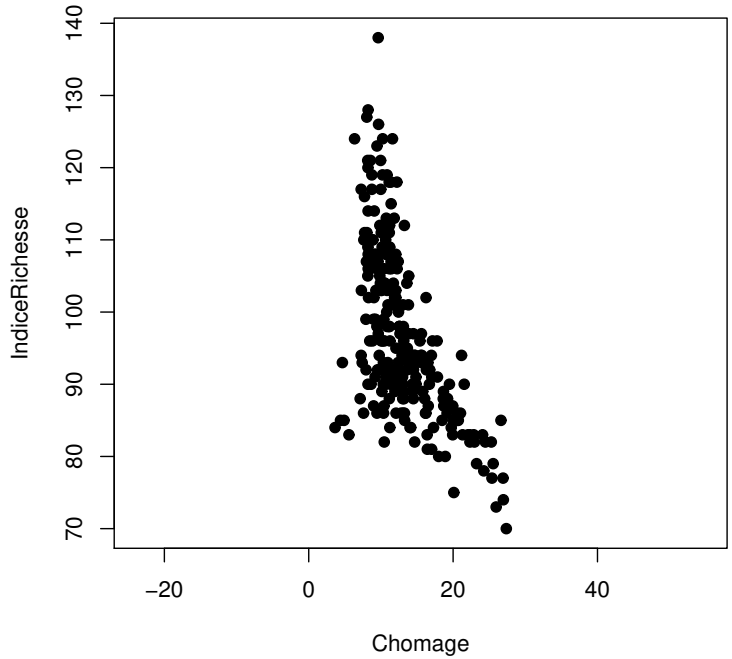
Lorsqu'il est important de visualiser les données dans un espace orthonormé, il convient de standardiser les données:

$$x_i \longrightarrow x_i' = \frac{x_i - \bar{x}}{s}$$

Exemple avec le chômage et l'indice de richesse

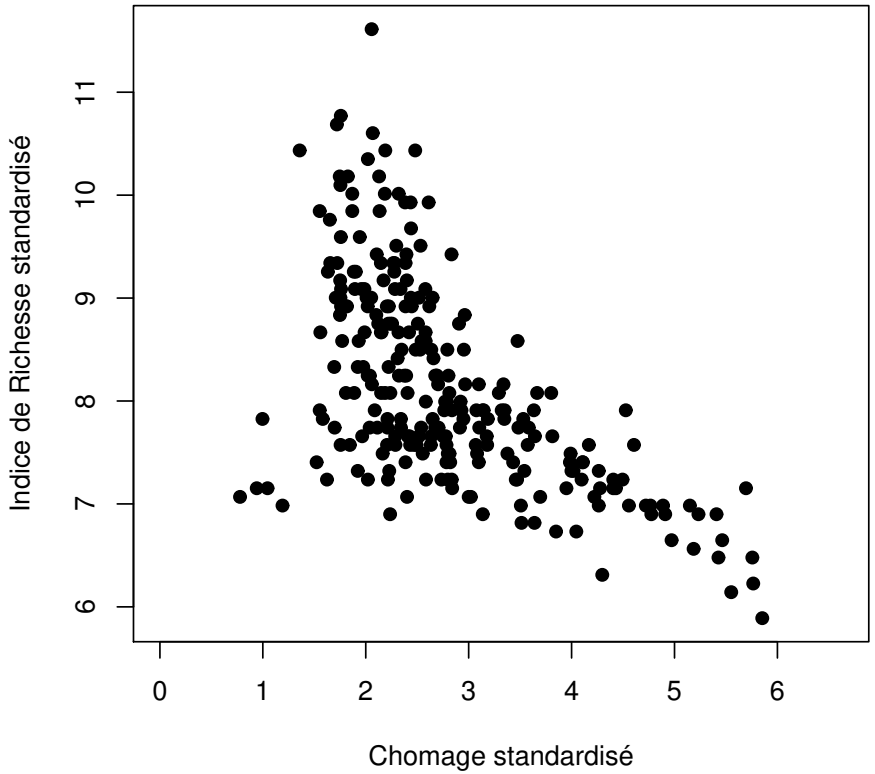
Repère orthogonal vs repère orthonormé





MAIS Les écarts horizontaux et verticaux ne sont pas comparables car les dispersions sont différentes.

Exemple avec le chômage et l'indice de richesse

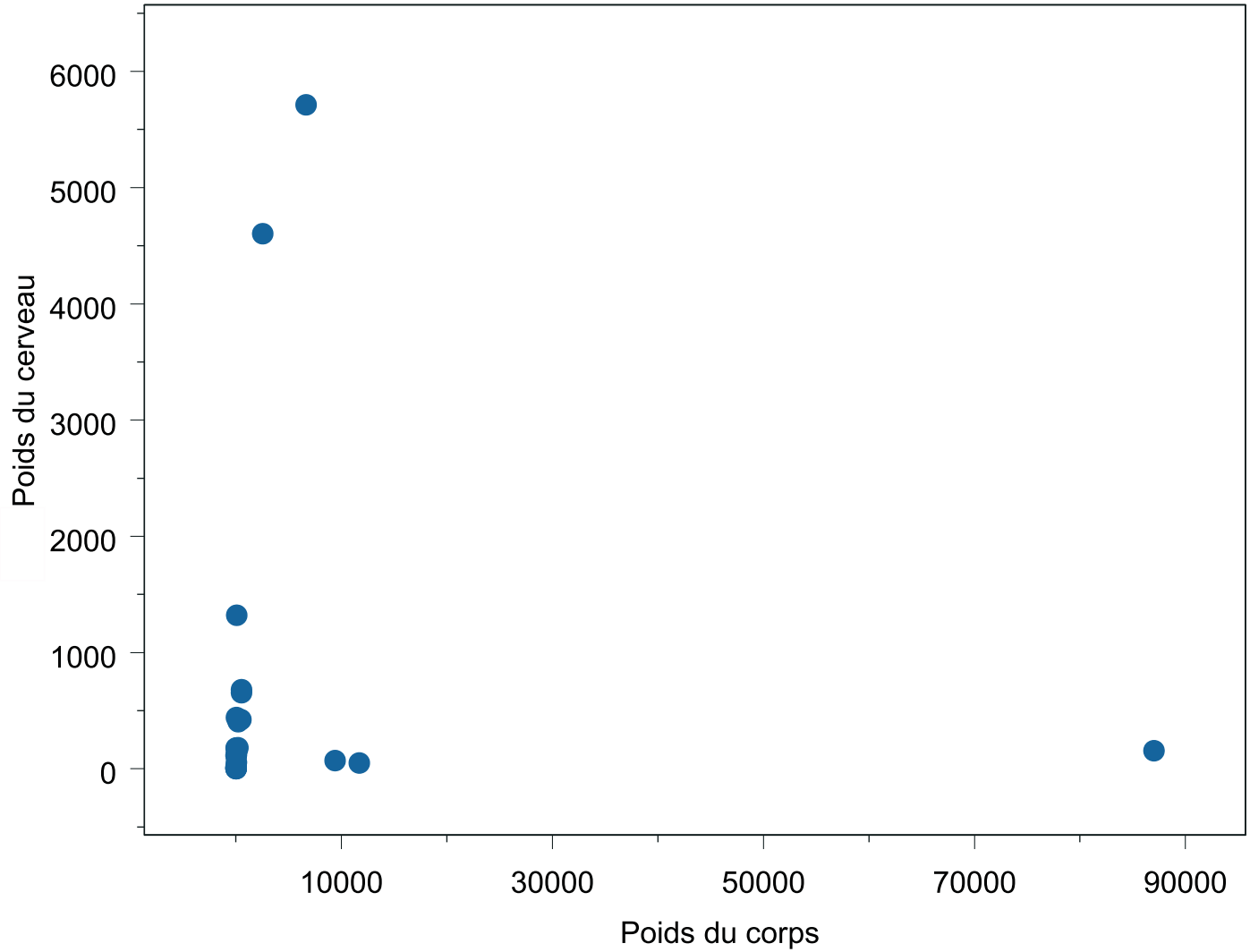


En "standardisant" les variables, la géométrique euclidienne peut être exploitée.

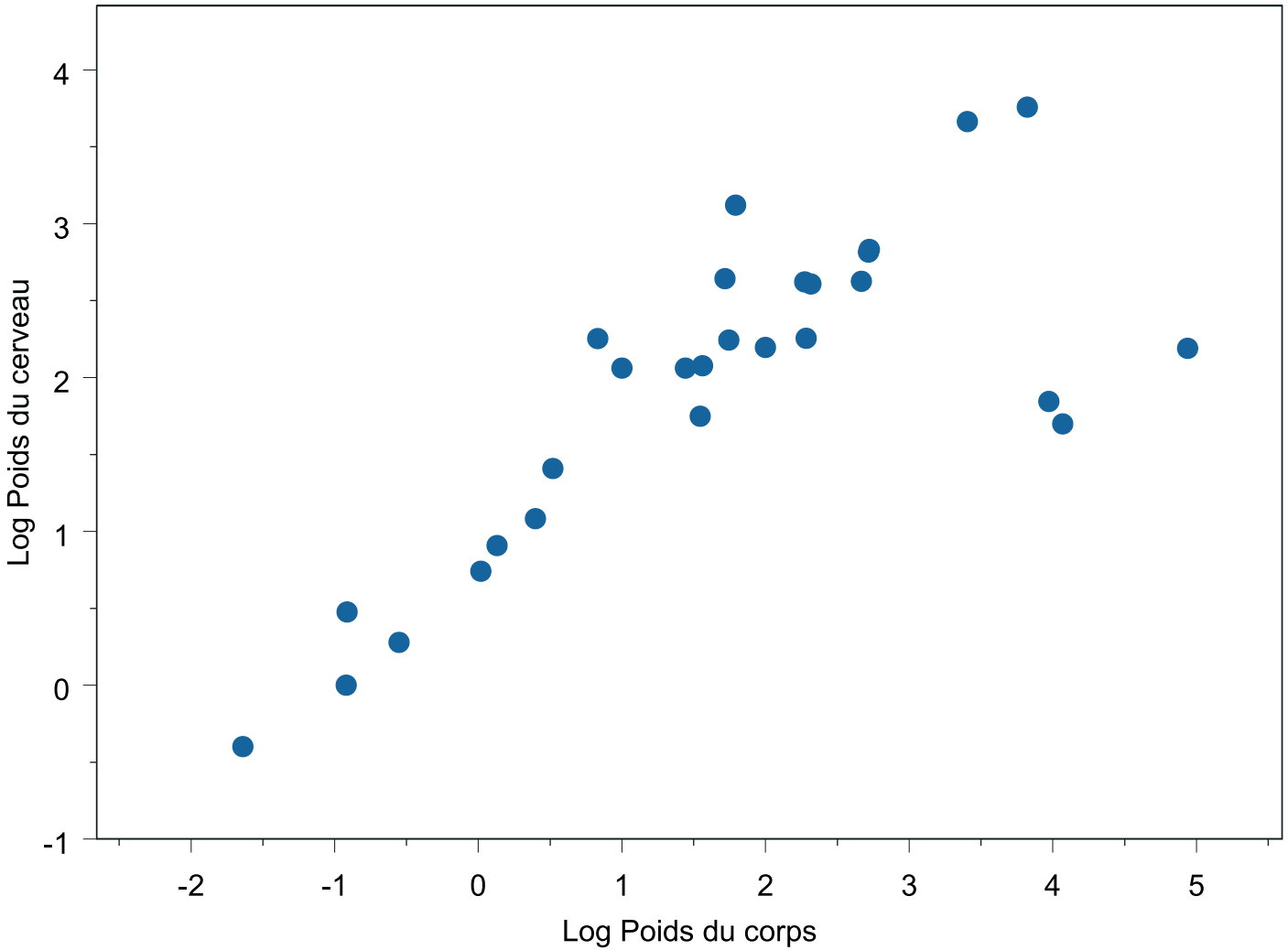
C. Transformation de données

i	Nom	Poids du corps (Kg)	Poids du cerveau (g)
1	Castor	1.35	8.1
2	Vache	465	423
3	Loup gris	36.33	119.5
4	Chèvre	27.66	115
5	Cobaye	1.04	5.5
6	Diplodocus	11700	50
7	Éléphant d'Asie	2547	4603
8	Âne	187.1	419
9	Cheval	521	655
10	Babouin	10	115
11	Chat	3.3	25.6
12	Girafe	529	680
13	Gorille	207	406
14	Humain	62	1320
15	Éléphant d'Afrique	6654	5712
16	Triceratops	9400	70
17	Singe Rhesus	6.8	179
18	Kangourou	35	56
19	Hamster	0.12	1
20	Souris	0.023	0.4
21	Lapin	2.5	12.1
22	Mouton	55.5	175
23	Jaguar	100	157
24	Chimpanzé	52.16	440
25	Brachiosaurus	87000	154.5
26	Rat	0.280	1.9
27	Taupe	0.122	3
28	Cochon	192	180

LIÈGE université


Poids du cerveau
6000
5000
4000
3000
2000
1000
0
10000
30000
50000
70000
90000
Poids du corps

En coordonnées doublement logarithmiques



Séries marginales

A partir de la série bivariée "brute"

$$S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$$

- Série marginale en X : $S_X = \{x_i, 1 \leq i \leq n\}$.
- Série marginale en Y : $S_Y = \{y_i, 1 \leq i \leq n\}$.

LIÈGE université

A partir du tableau de contingence $\{(x_j, y_k), n_{jk}\}$ on peut retrouver la série

$$S_X = \{(x_j, n_{j\bullet}), 1 \leq j \leq J\}$$

où $n_{j\bullet}$ est l'effectif marginal de x_j

Valeurs de X	Valeurs de Y					
	y_1	y_2	$\dots$	y_k	$\dots$	y_K
x_1	n_{11}	n_{12}	$\dots$	n_{1k}	$\dots$	n_{1K}
$\vdots$						
x_j	n_{j1}	n_{j2}	$\dots$	n_{jk}	$\dots$	n_{jK}
$\vdots$						
x_J	n_{J1}	n_{J2}	$\dots$	n_{Jk}	$\dots$	n_{JK}

D'où, $n_{j\bullet} = \sum_{k=1}^{K} n_{jk}$.

De la même manière,

$$S_Y = \{(y_k, n_{\bullet k}), 1 \leq k \leq K\}$$

où $n_{\bullet k}$ est l'effectif marginal de y_k donné par $n_{\bullet k} = \sum_{j=1}^{J} n_{jk}$.

	Valeurs de Y					
Valeurs de X	y_1	$\dots$	y_k	y_k	$\dots$	y_K
x_1	n_{11}	$\dots$	n_{1k}	n_{1k}	$\dots$	n_{1K}
				$+$		
$\vdots$				$\vdots$		
x_j	n_{j1}	$\dots$	n_{jk}	n_{jk}	$\dots$	n_{jK}
				$+$		
$\vdots$				$\vdots$		
x_J	n_{J1}	$\dots$	n_{Jk}	n_{Jk}	$\dots$	n_{JK}

Les distributions marginales sont ajoutées au tableau de contingence dans une ligne et une colonne:

	Valeurs de Y					Dist Marg
Valeurs de X	y_1	$\dots$	y_k	$\dots$	y_K	de X
x_1	n_{11}	$\dots$	n_{1k}	$\dots$	n_{1K}	$n_{1\bullet}$
$\vdots$						$\vdots$
x_j	n_{j1}	$\dots$	n_{jk}	$\dots$	n_{jK}	$n_{j\bullet}$
$\vdots$						$\vdots$
x_J	n_{J1}	$\dots$	n_{Jk}	$\dots$	n_{JK}	$n_{J\bullet}$
Dist Marg Y	$n_{\bullet 1}$	$\dots$	$n_{\bullet k}$	$\dots$	$n_{\bullet K}$	n

A. Parti – Provinces

Partis	Provinces					
	Brabant	Hainaut	Liège	Lux	Namur	Dist Marg
Autre	2	0	7	3	1	13
CdH	3	13	16	23	12	67
Ecolo	2	1	1	1	0	5
MR	14	16	32	10	17	89
PS	6	39	28	7	8	88
Dist Marg	27	69	84	44	38	262

C. Indice de richesse groupé – Chômage groupé: tableau complet

Indice	Chômage						D. Marg Indice
	$[0,5]$	$]5,10]$	$]10,15]$	$]15,20]$	$]20,25]$	$]25,30]$	
$]70,80]$	0	0	0	0	1	1	2
$]80,90]$	3	2	4	2	2	2	15
$]90,100]$	1	5	16	8	1	0	31
$]100,110]$	0	16	12	0	0	0	28
$]110,120]$	0	3	2	0	0	0	5
$]120,130]$	0	2	1	0	0	0	3
Dist. Marg	4	28	35	10	4	3	84

Fréquences marginales

$$\begin{aligned} f_{j\bullet} &= \text{fréquence de la valeur } x_j \\ &= \frac{n_{j\bullet}}{n} \end{aligned}$$

et, de même,

$$\begin{aligned} f_{\bullet k} &= \text{fréquence de la valeur } y_k \\ &= \frac{n_{\bullet k}}{n} \end{aligned}$$

On a $\sum_{j=1}^{J} f_{j\bullet} = \sum_{k=1}^{K} f_{\bullet k} = \sum_{j=1}^{J} \sum_{k=1}^{K} f_{jk} = 1$.

Série bivariée $\rightarrow$ Séries marginales

MAIS

Séries marginales $\nrightarrow$ Série bivariée

L'étude de la série double permet de disposer des deux séries statistiques simples.

	Sport (oui-non)		Dist Marg
Sexe	OUI	NON	Sexe
F	21	21	42
M	69	9	78
Dist Marg de Sport	90	30	120

Série bivariée $\rightarrow$ Séries marginales

MAIS

Séries marginales $\nrightarrow$ Série bivariée

L'inverse n'est par contre pas vrai. A partir des distributions des effectifs des deux séries simples S_X et S_Y, on ne peut pas construire le tableau de contingence relatif à l'observation simultanée des deux variables X et Y.

	Sport (oui-non)		Dist Marg
Sexe	OUI	NON	Sexe
F	12 22	3020	42
M	78 68	0 10	78
Dist Marg de Sport	90	30	120

Distributions conditionnelles

Quelle est la distribution du parti du bourgmestre dans chacune des provinces?

	Provinces					
Partis	Brabant	Hainaut	Liège	Lux	Namur	Dist Marg
Autre	2	0	7	3	1	13
CdH	3	13	16	23	12	67
Ecolo	2	1	1	1	0	5
MR	14	16	32	10	17	89
PS	6	39	28	7	8	88
Dist Marg	27	69	84	44	38	262

Distributions conditionnelles

Quel est le taux de chômage des communes dont l'indice de richesse est juste en dessous de la valeur de référence?

	Chômage						D. Marg
Indice	$[0,5]$	$]5,10]$	$]10,15]$	$]15,20]$	$]20,25]$	$]25,30]$	Indice
$]70,80]$	0	0	0	0	1	1	2
$]80,90]$	3	2	4	2	2	2	15
$]90,100]$	1	5	16	8	1	0	31
$]100,110]$	0	16	12	0	0	0	28
$]110,120]$	0	3	2	0	0	0	5
$]120,130]$	0	2	1	0	0	0	3
D. Marg	4	28	35	10	4	3	84

Définition et notations

Une distribution conditionnelle consiste à fixer a priori la valeur d'une variable et à examiner les variations de l'autre variable compte tenu de cette contrainte.

Par exemple, fixons la variable X à x_j.

On ne s'intéresse alors qu'aux couples $(x_j, y_1), \ldots, (x_j, y_K)$, pour lesquels les effectifs sont connus: $n_{j1}, \ldots, n_{jK}$.

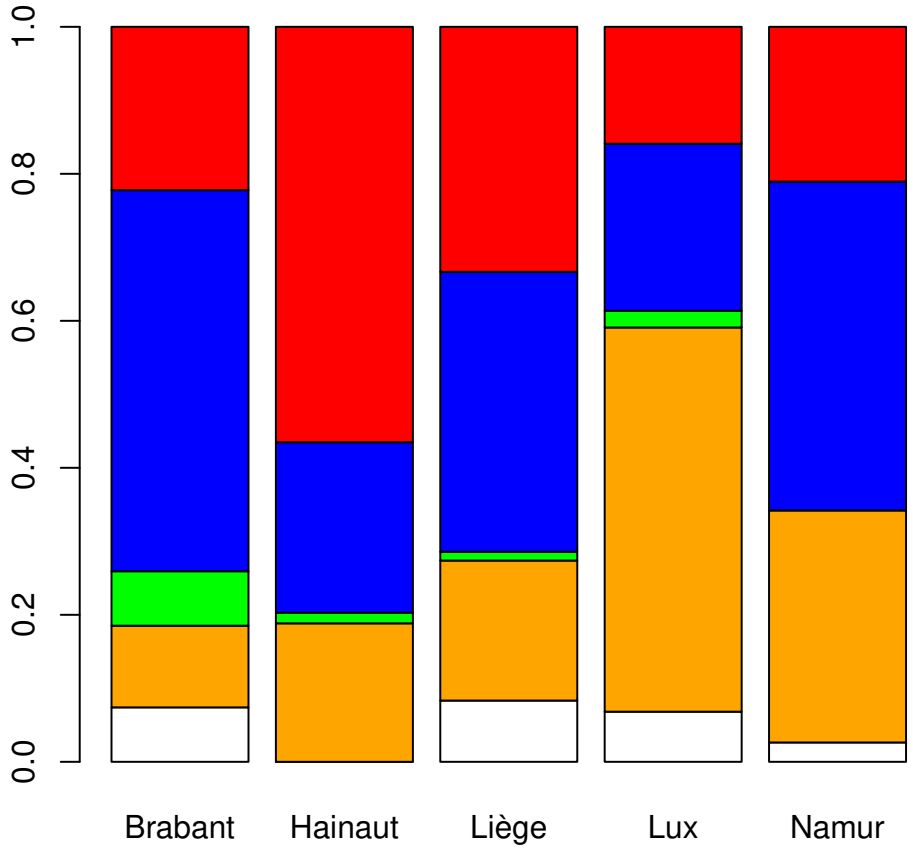
D'où, la distribution de Y conditionnelle à $X = x_j$ est

$$S_{Y|x_j} = \{(y_k, n_{jk}), 1 \leq k \leq K\}.$$

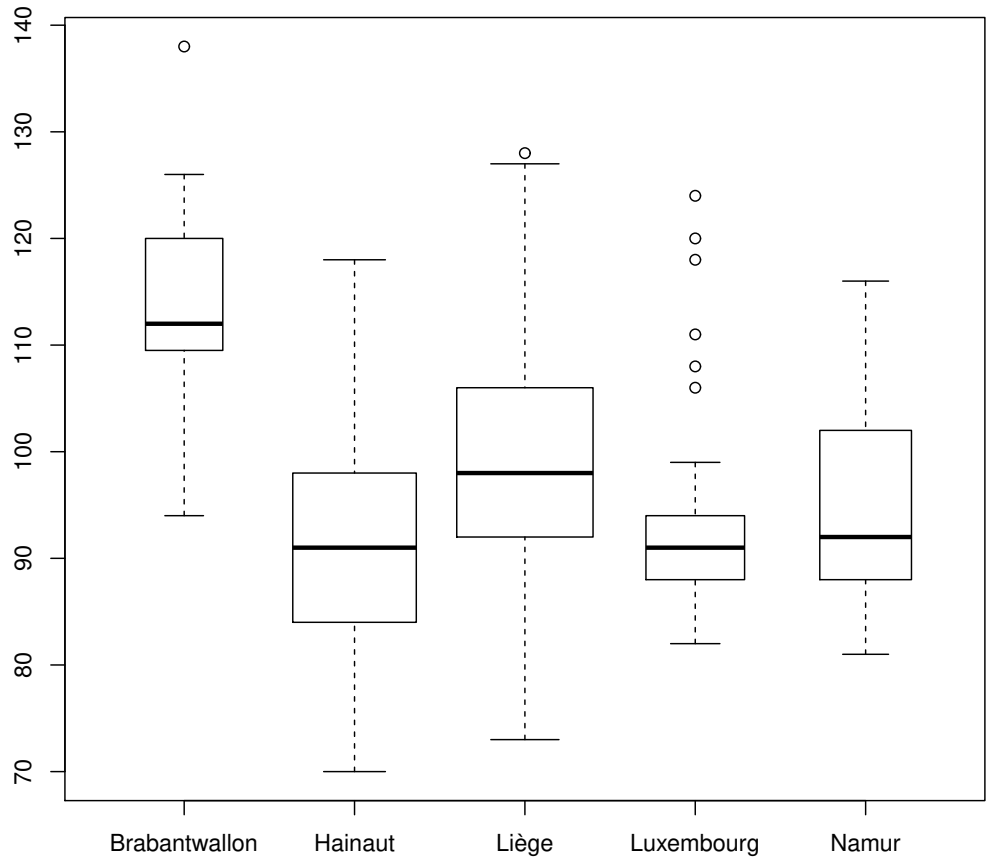
Effectif total de $S_{Y|x_j}$ (= somme des effectifs):
$n_{j1} + \ldots + n_{jK} = \sum_{k=1}^{K} n_{jk} = n_{j\bullet}$.

Les fréquences conditionnelles (= effectif divisé par l'effectif total):
$f_{k|x_j} = \frac{n_{jk}}{n_{j\bullet}}$.

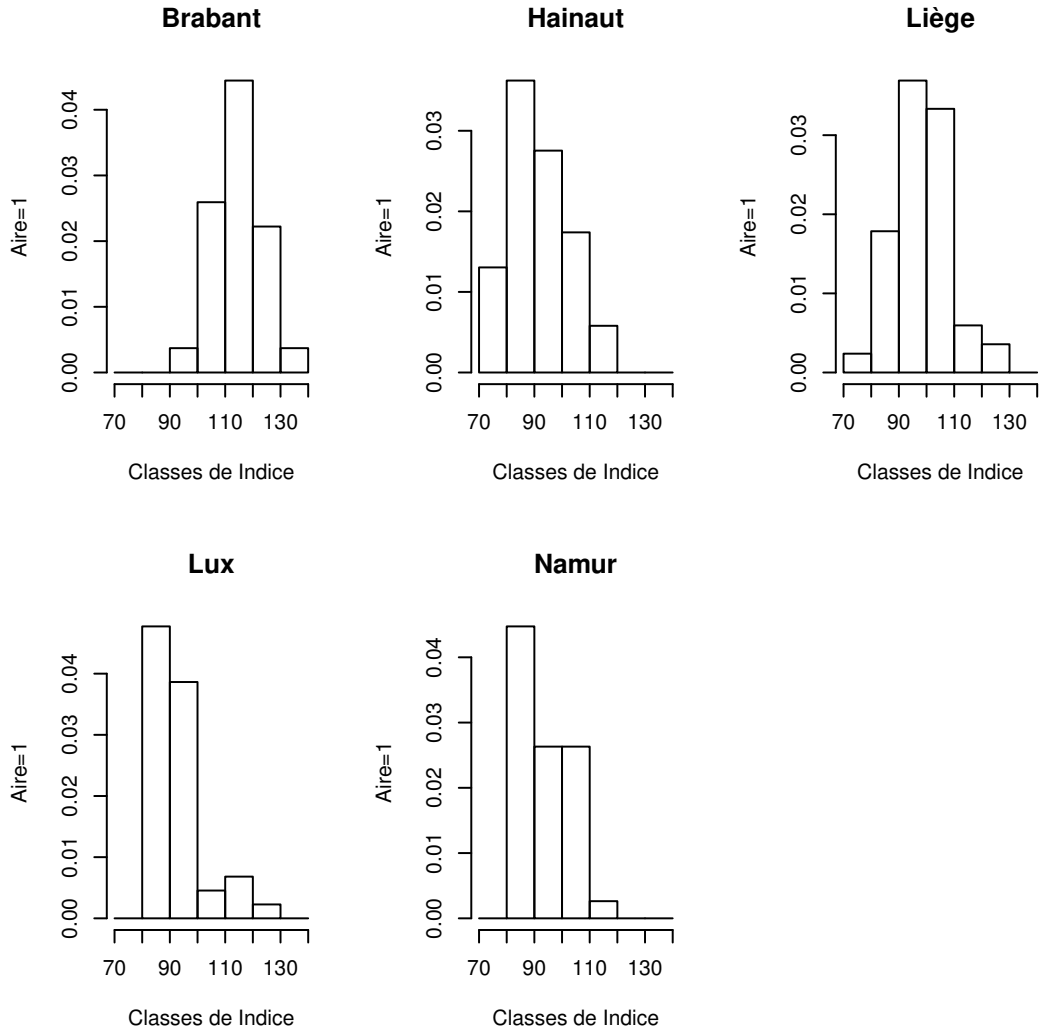
Distributions conditionnelles du parti du bourgmestre en fonction de la province



Distributions conditionnelles de la variable indice de Richesse en fonction de la province



Distributions conditionnelles de la variable indice de Richesse en fonction de la province



Réduction des données

- Réduction des séries univariées marginales ou conditionnelles (moyennes, variances, médianes,...)
- Analyse bivariée

Réduction des séries marginales

Moyenne marginale $\bar{x}$ de S_X : $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ ou $\bar{x} = \frac{1}{n}\sum_{j=1}^{J} n_{j\bullet} x_j$.
Variance marginale de S_X :

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \text{ ou } s_x^2 = \frac{1}{n}\sum_{j=1}^{J} n_{j\bullet}(x_j - \bar{x})^2.$$

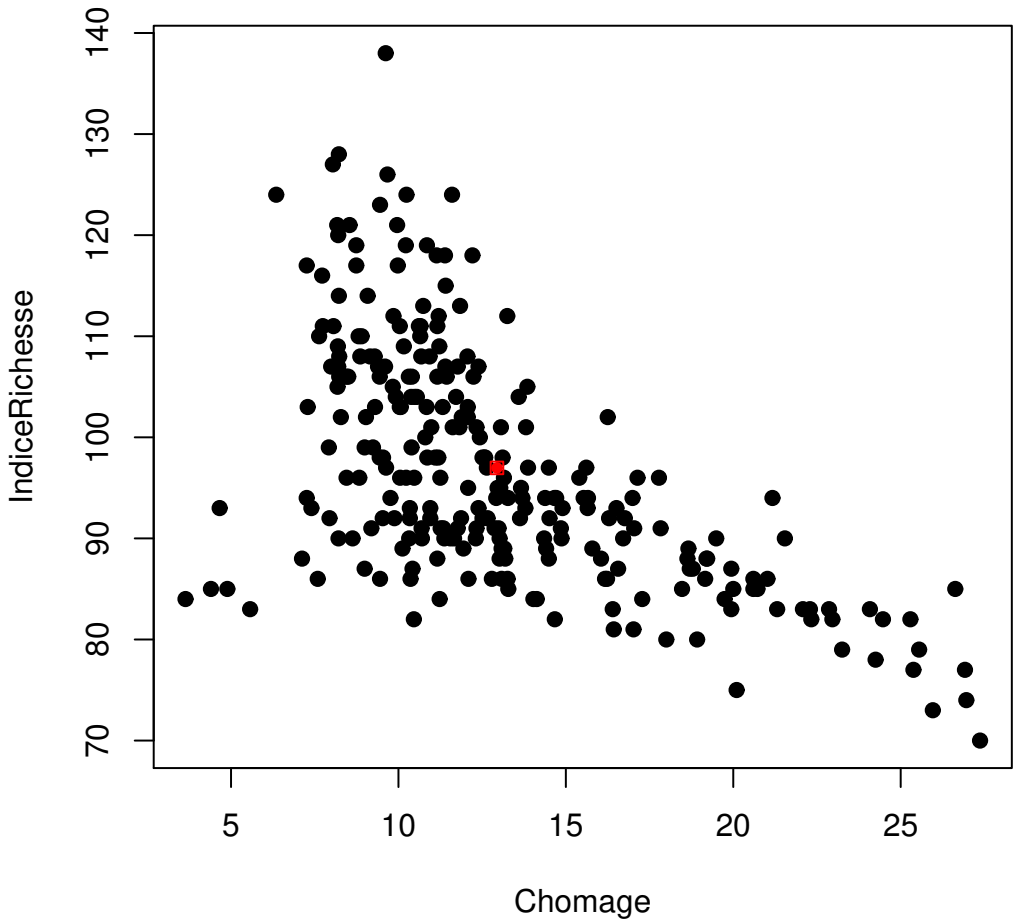
Moyenne marginale $\bar{y}$ de S_Y : $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{x} = \frac{1}{n}\sum_{k=1}^{K} n_{\bullet k} y_k$.
Variance marginale de S_Y

$$s_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 \text{ ou } s_y^2 = \frac{1}{n}\sum_{k=1}^{K} n_{\bullet k}(y_k - \bar{y})^2.$$

De la même manière, on peut définir $\tilde{x}, \tilde{y}$,...

LIÈGE université

Moyenne = centre de gravité du nuage de points



Formules

La moyenne conditionnelle de Y sachant que $X = x_j$ est

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

La variance conditionnelle de Y sachant que $X = x_j$ est

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

Les notations sont du même style pour les distributions conditionnelles de X en fonction de Y.

Moyennes et écarts-types des séries d'indices de richesse conditionnelles aux provinces

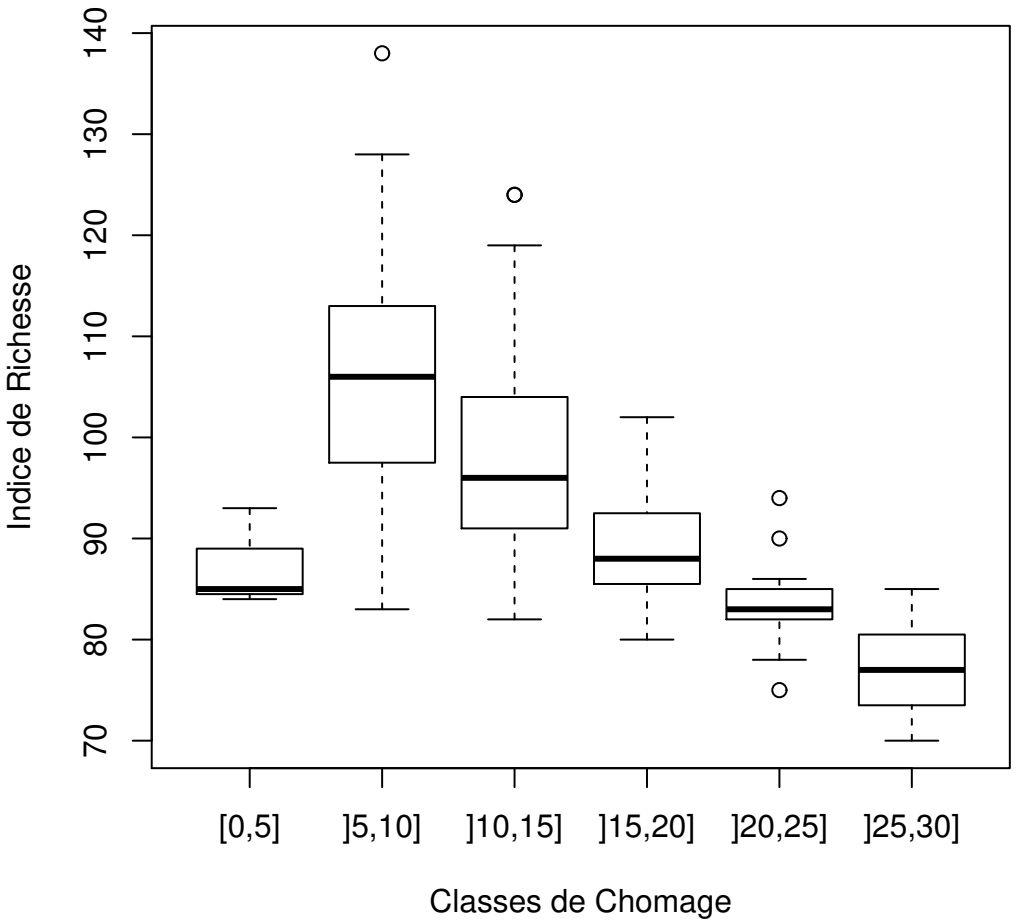
Illustration de la propriété de décomposition de la variance:

Provinces	Moyennes conditionnelles	Ecarts-types conditionnels
Brabant wallon	114.1	8.63
Hainaut	91.7	10.56
Liège	98.9	10.66
Luxembourg	93.4	9.55
Namur	94.2	8.87

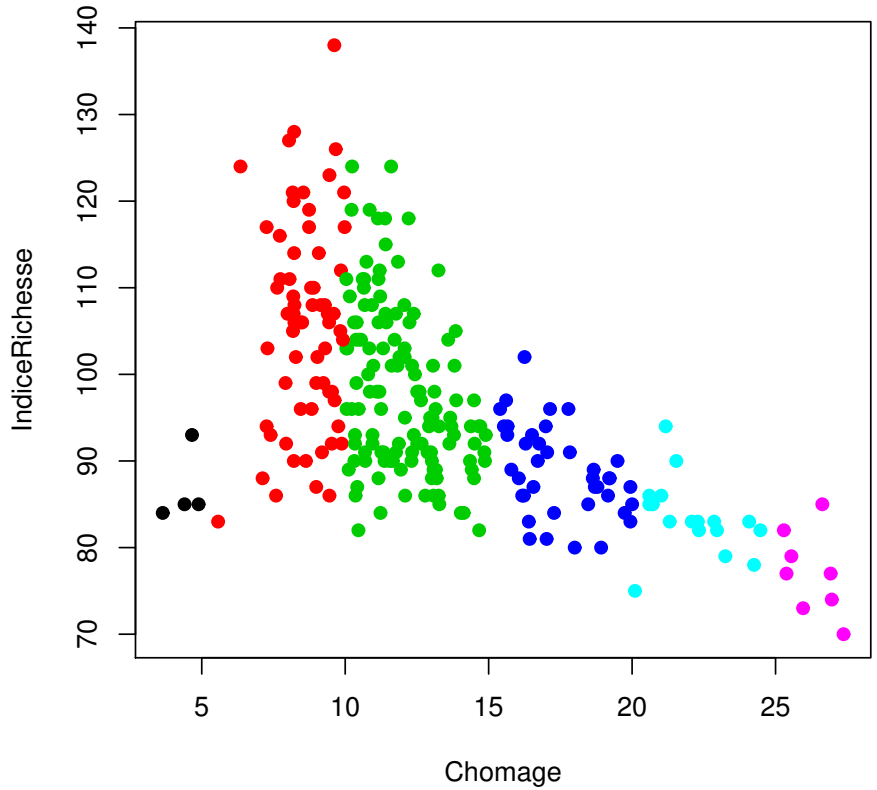
Utilité: première analyse du lien éventuel existant entre les variables

	Chômage						D. Marg
Indice	$[0,5]$	$]5,10]$	$]10,15]$	$]15,20]$	$]20,25]$	$]25,30]$	Indice
$]70,80]$	0	0	0	0	1	1	2
$]80,90]$	3	2	4	2	2	2	15
$]90,100]$	1	5	16	8	1	0	31
$]100,110]$	0	16	12	0	0	0	28
$]110,120]$	0	3	2	0	0	0	5
$]120,130]$	0	2	1	0	0	0	3
D. Marg	4	28	35	10	4	3	84

Médianes de l'indice de richesse en fonction des classes de chômage



Moyennes de l'indice de richesse en fonction du chômage



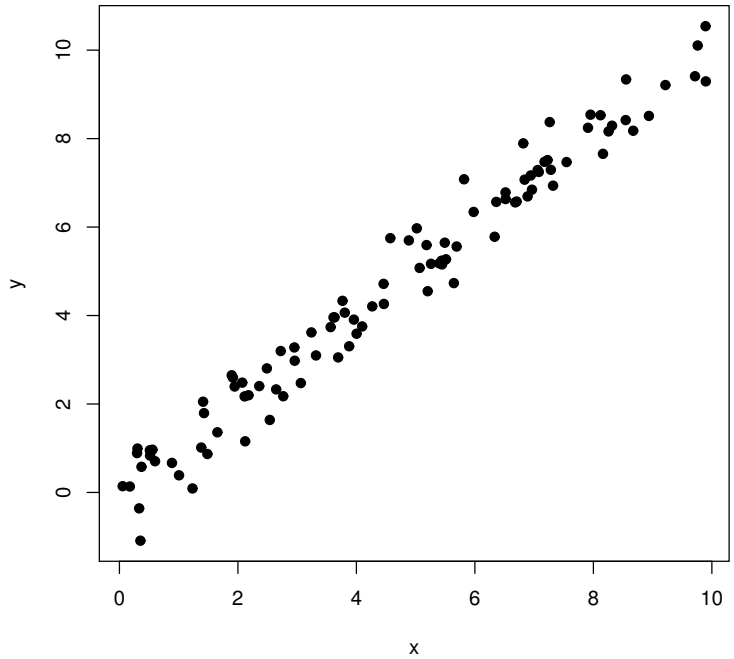
Classes Chômage	[0, 5]	]5, 10]	]10, 15]	]15, 20]	]20, 25]	]25, 30]
Moy. cond. IR	x	105.9	98.1	88.8	83.5	77.1

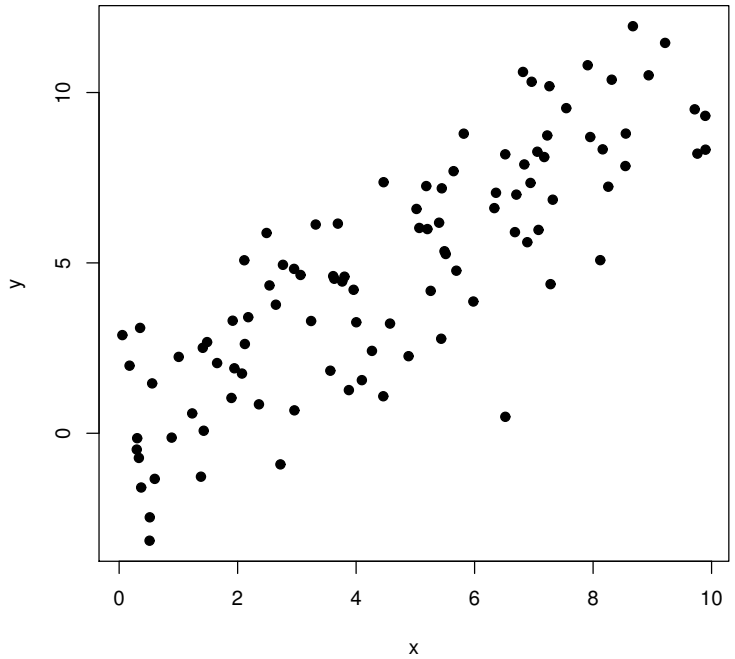
Analyse bivariée

Lorsqu'une des variables du couple est qualitative, l'analyse bivariée se fait à partir des moyennes et variances conditionnelles.

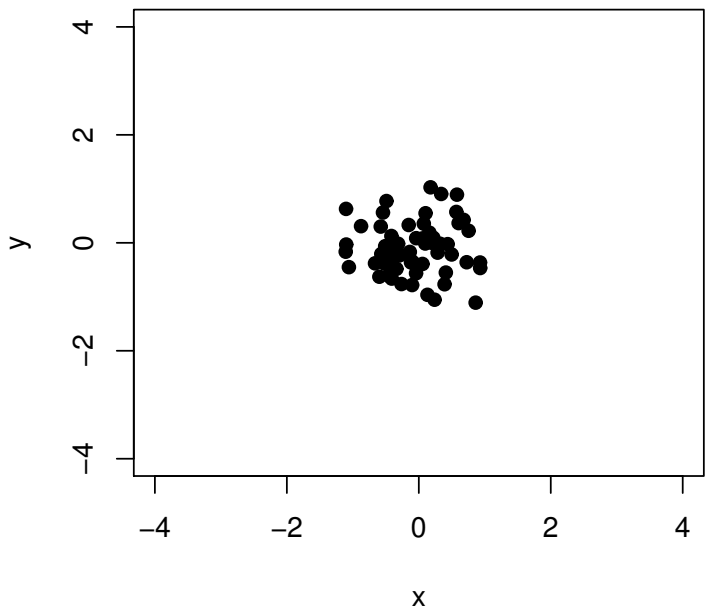
Si X et Y sont quantitatives, la première démarche consiste à analyser la forme du diagramme de dispersion.

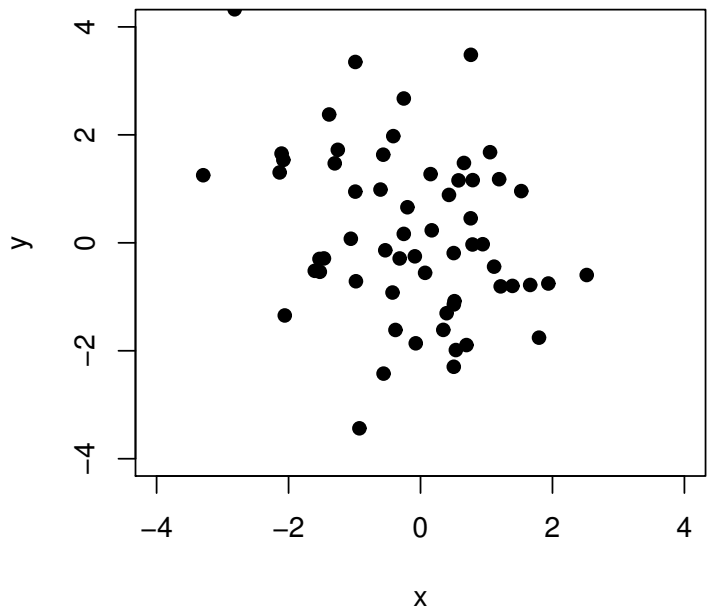
Le graphique est-il concentré ou au contraire dispersé?



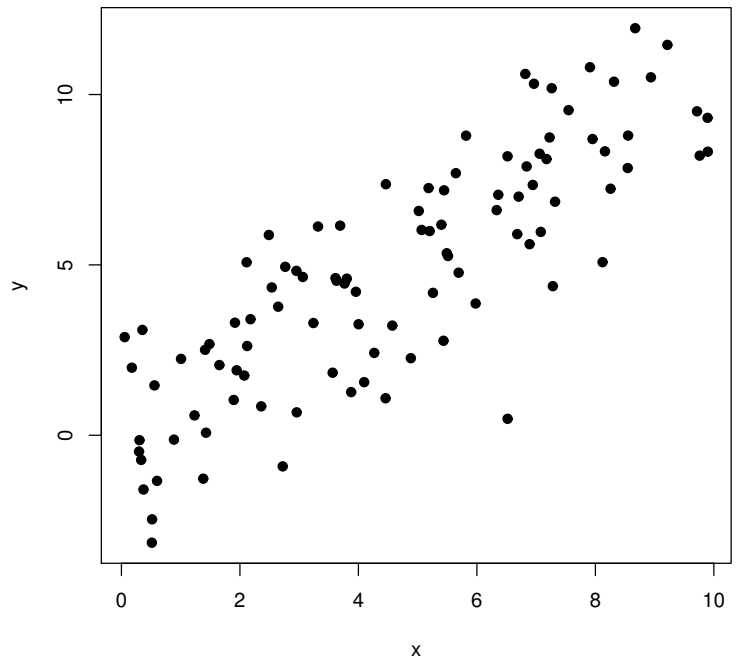


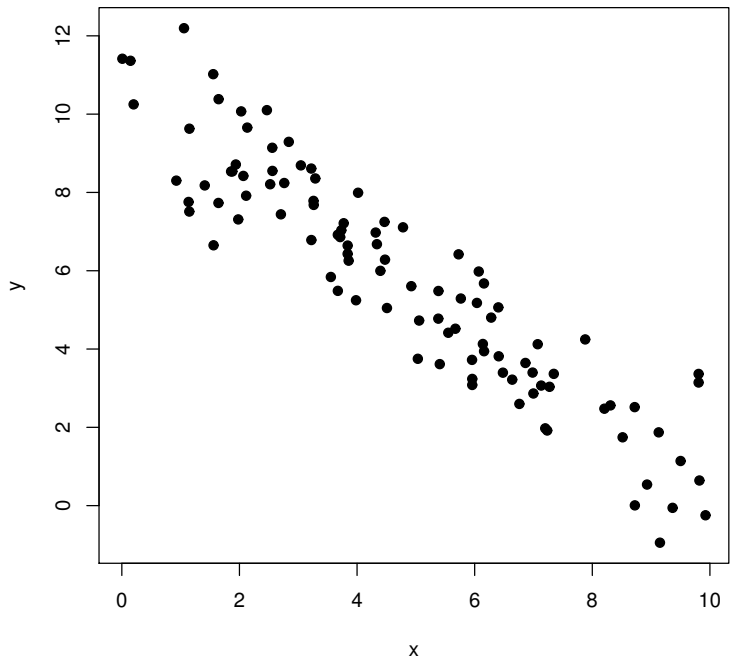
Le graphique est-il concentré ou au contraire dispersé?



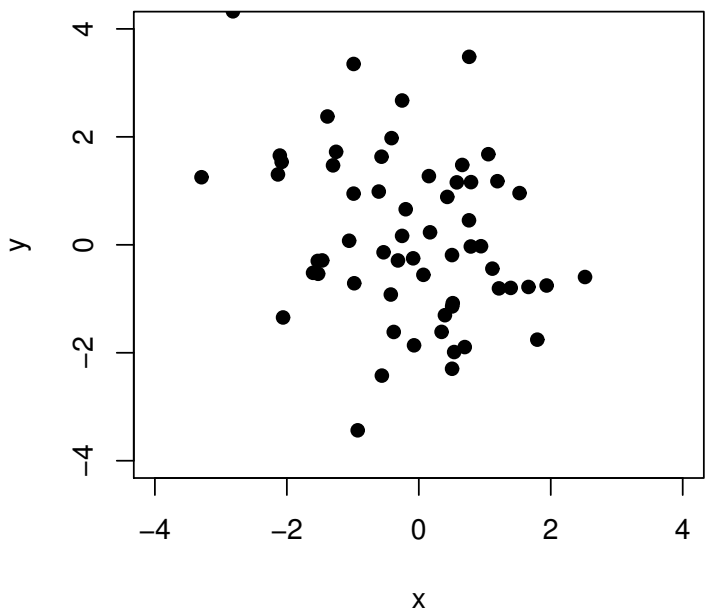


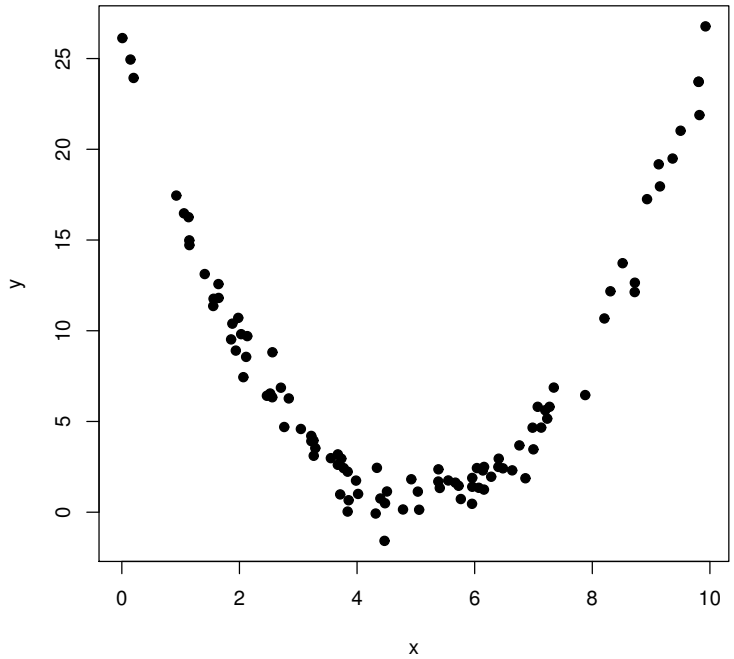
Le diagramme de dispersion présente-t-il une “structure” ?



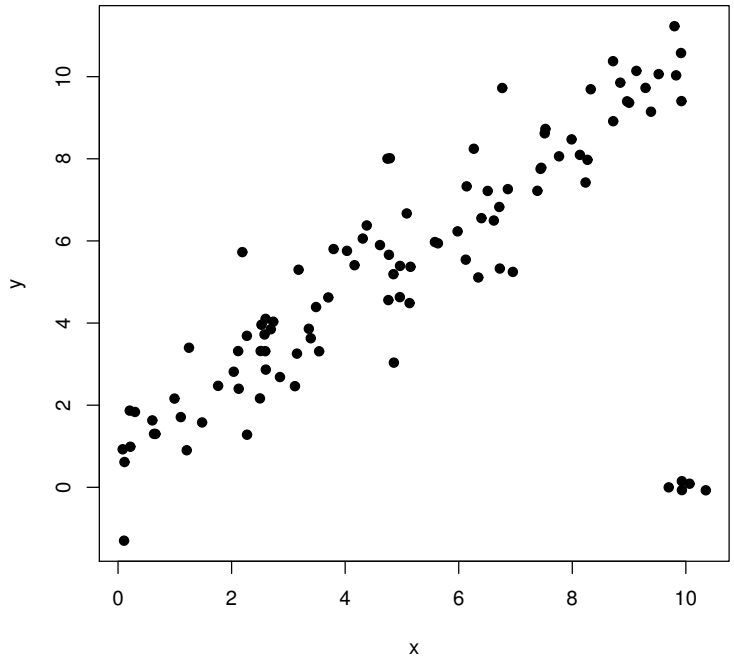


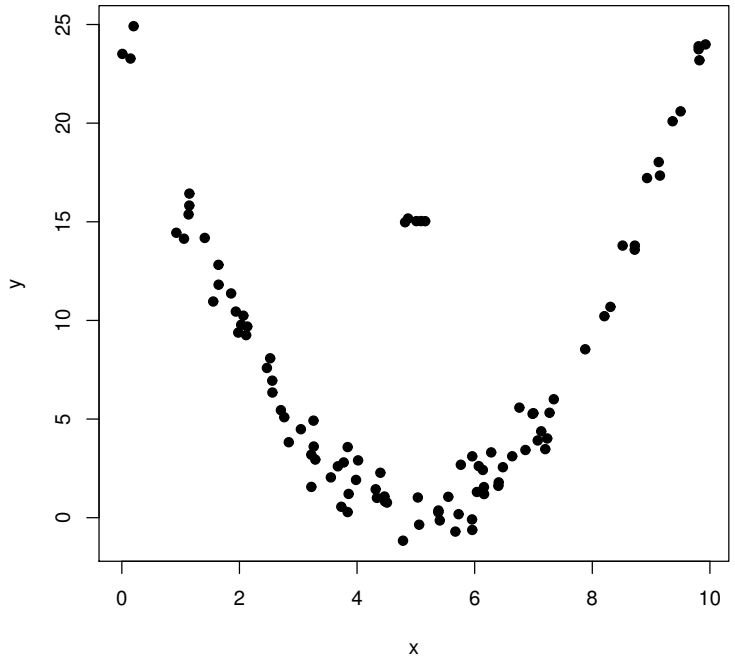
Le diagramme de dispersion présente-t-il une “structure” ?





Y a-t-il des observations atypiques?





Indépendance ou dépendance/association

Soit X l'indice de richesse.

Les valeurs prises par X sont-elles dépendantes

- du taux de chômage
- du prix moyen des terrains à bâtir
- de la qualité de l'air de la commune
- ...

En cas de dépendance supposée...

- La dépendance fonctionnelle: il existe une fonction f telle que $y = f(x)$.
- La dépendance statistique: association moins nette que la dépendance fonctionnelle.

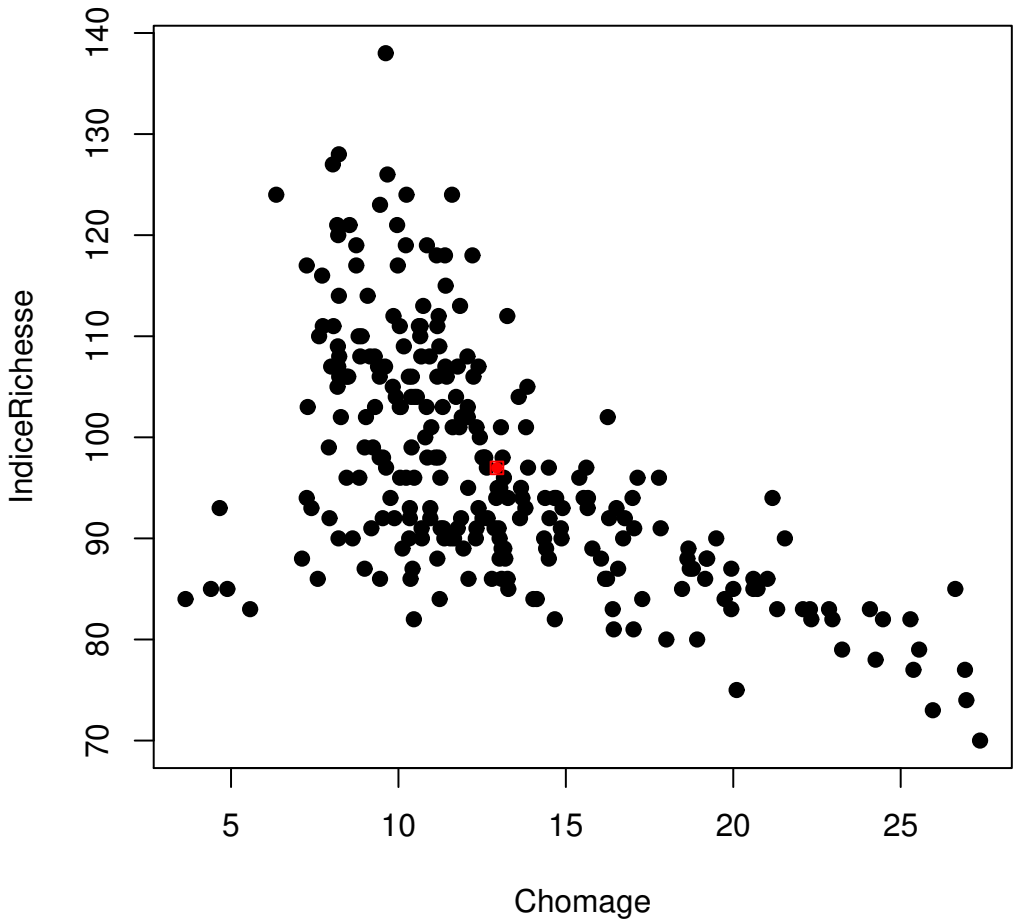
Pour détecter la présence d'une dépendance statistique, on calcule la covariance entre les variables X et Y:

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

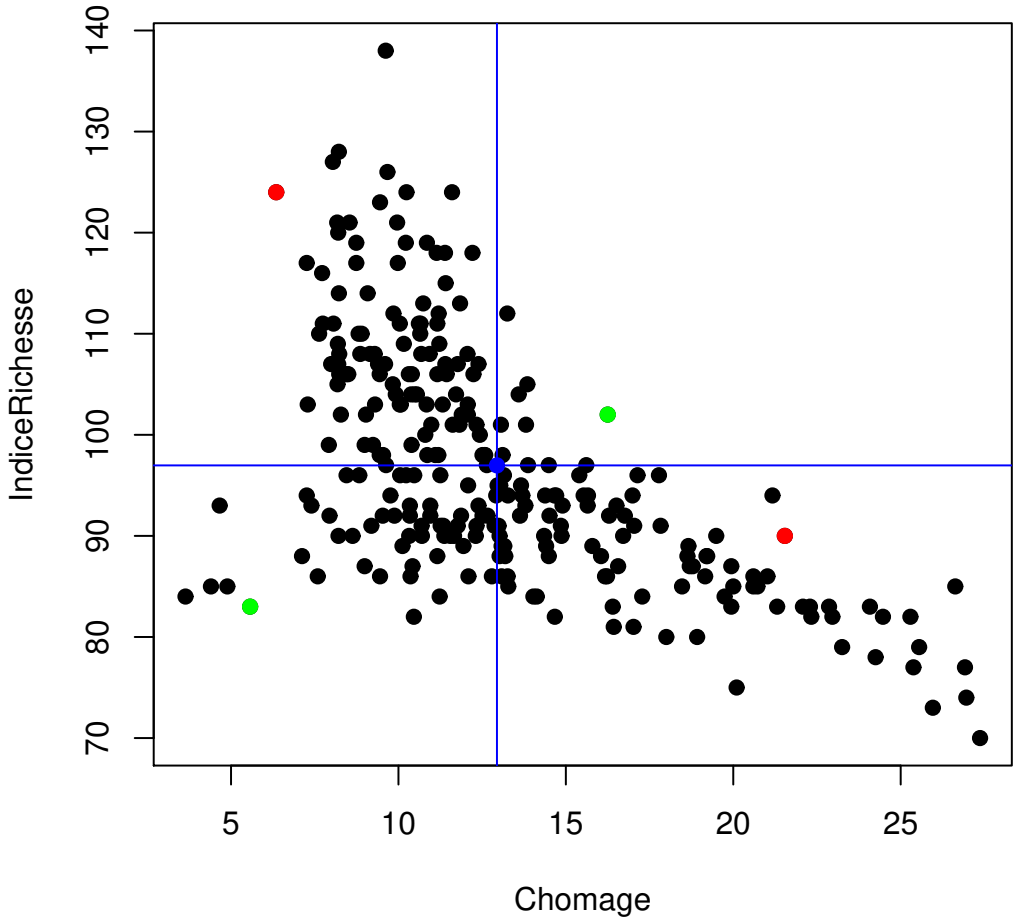
A partir de la distribution des effectifs,

$$s_{xy} = \frac{1}{n} \sum_{j=1}^{J} \sum_{k=1}^{K} n_{jk} (x_j - \bar{x})(y_k - \bar{y}).$$

Interprétation du signe de la covariance

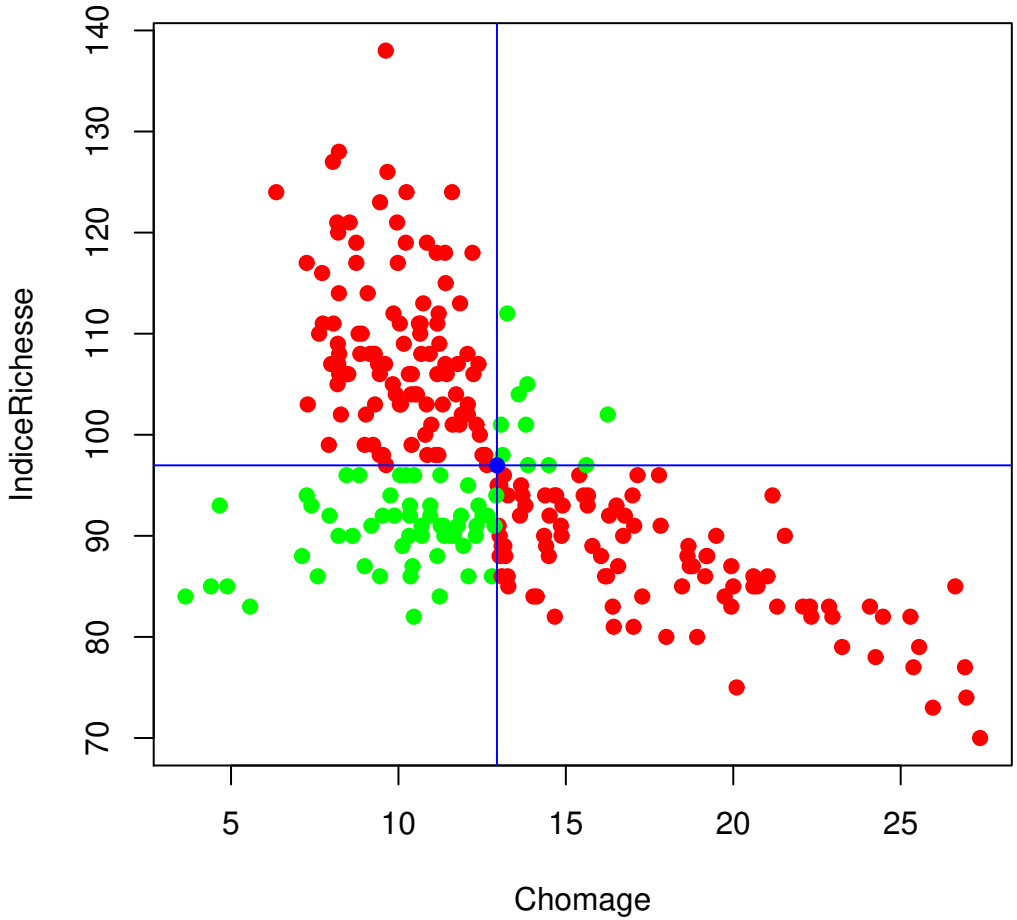


Interprétation du signe de la covariance



$s_{xy} = -33.5$

Interprétation du signe de la covariance



$s_{xy} =$
$\frac{1}{n}\{\text{Somme des contributions} > 0 + \text{Somme des contributions} < 0\} =$
-33.5

Propriétés de la covariance

Proposition

1) $s_{xy} = s_{yx}$ *et* $s_{xx} = s_x^2$.
2) $s_{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}$ *ou* $s_{xy} = \frac{1}{n}\sum_{j=1}^{J}\sum_{k=1}^{K} n_{jk} x_j y_k - \bar{x}\bar{y}$.
3) *Effet du changement d'échelle et d'origine sur la covariance: si on effectue une transformation affine sur les observations pour obtenir la nouvelle série double* $S' = \{(x_i', y_i') : 1 \leq i \leq n\}$ *avec*

$$x_i' = ax_i + b \text{ et } y_i' = cy_i + d, \quad \text{avec } a, b, c, d \in \boldsymbol{R},$$

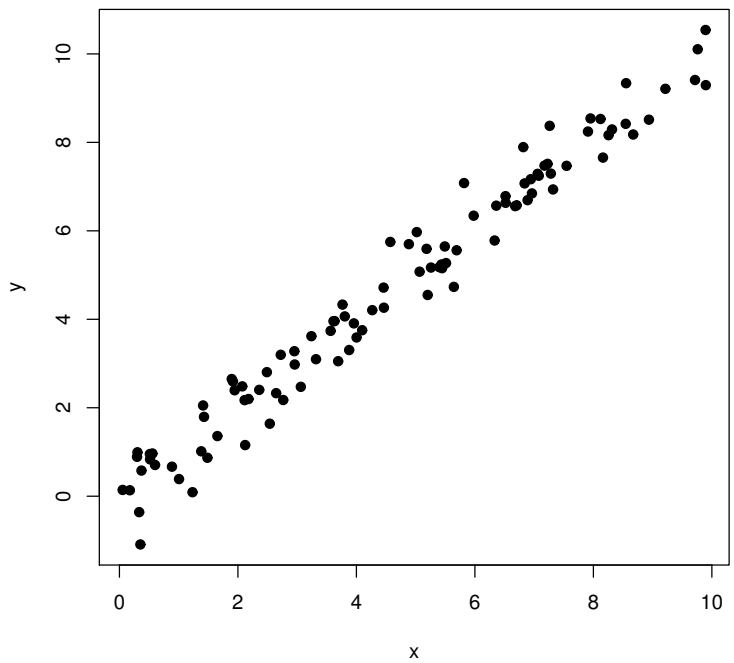
alors la covariance de la nouvelle série vérifie $s_{xy}' = acs_{xy}$.

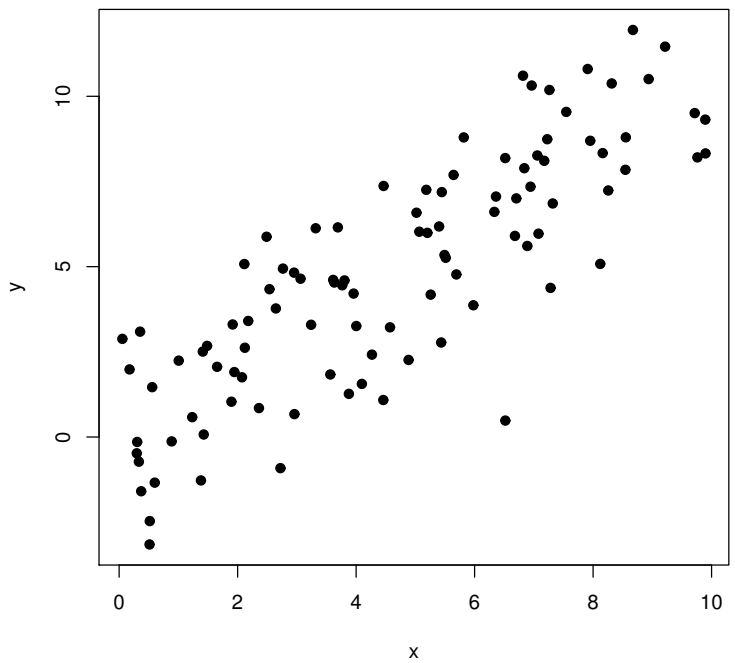
Démonstrations: au tableau

Outils: définition de s_{xy}, distributivité, exploitation des transformations de moyennes et variances.

Interprétation

Le signe de la covariance est informatif mais la valeur absolue de s_{xy} ne permet pas de se faire une idée de l'intensité de la relation.





Pour le premier graphique, $s_{xy} = 7.78$ et pour le second, $s_{xy} = 8.03$

Propriétés de la covariance (fin)

4) Borne sur la covariance:

Proposition

La covariance est toujours, en valeur absolue, inférieure ou égale au produit des écarts-types marginaux s_x, s_y :

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

L'égalité n'est possible que si tous les points observés sont alignés.

5) Matrice de variances-covariances $S = \begin{pmatrix} s_x^2 & s_{xy} \\ s_{xy} & s_y^2 \end{pmatrix}$.

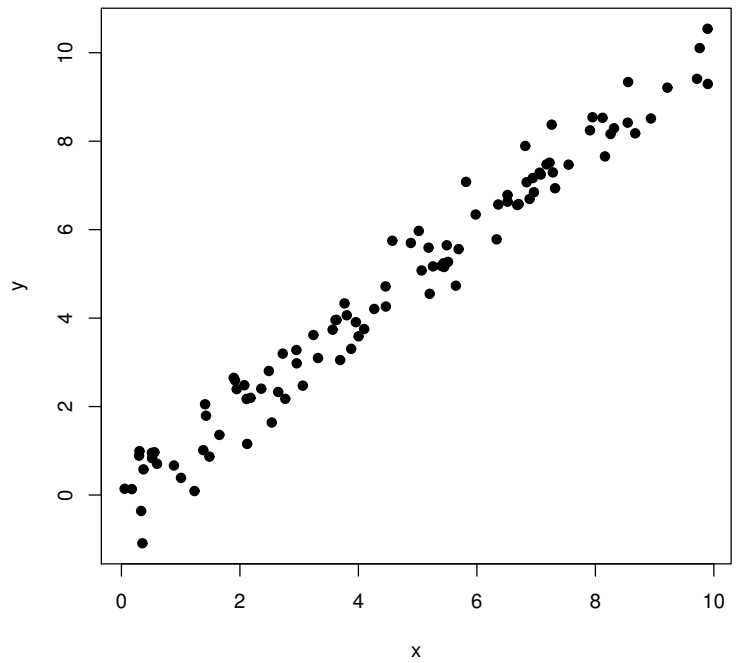
Proposition

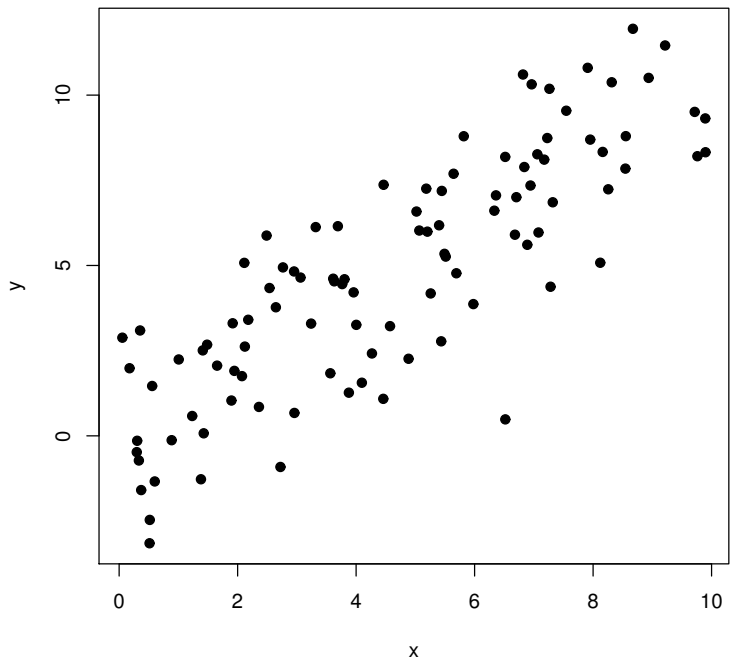
Cette matrice est 2-carrée, symétrique et habituellement inversible (sinon on parle d'un cas dégénéré).

Résultats admis.

Interprétation

Pour le premier graphique, $s_{xy} = 7.78$ et pour le second, $s_{xy} = 8.03$





avec $s_y \times s_y = 7.92$ dans le premier graphique et $s_y \times s_y = 9.66$ dans le second.

Coefficient de corrélation

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

si les variances marginales s_x et s_y diffèrent de 0.

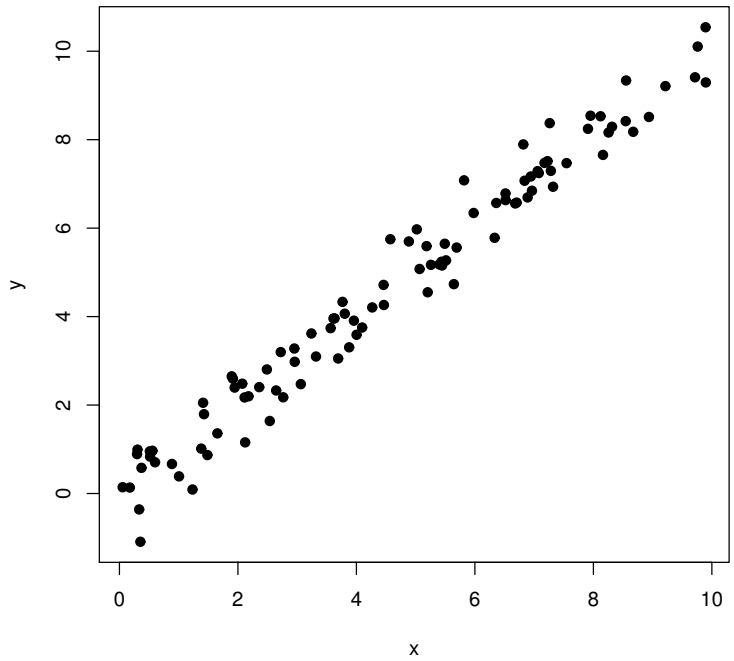
Proposition

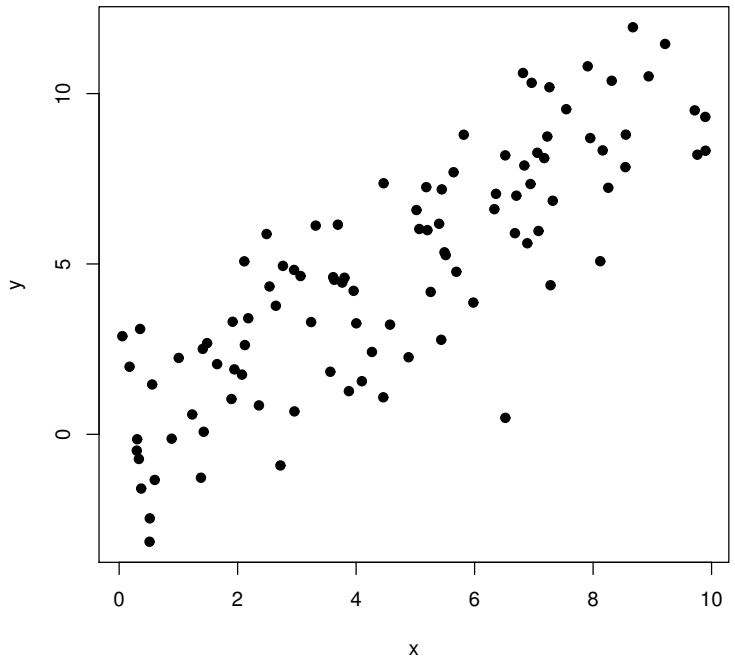
- *Le signe de* r_{xy} *est celui de* s_{xy}*.*
- $-1 \leq r_{xy} \leq 1$ *et* r_{xy} *est égal à 1 ou* -1 *uniquement lorsque les points* (x_i, y_i) *sont tous situés sur une même droite.*
- $r'_{xy} = \frac{ac}{|ac|} r_{xy}$*.*

Démonstrations: au tableau

Outils: définition de r_{xy} et exploitation des propriétés précédentes.

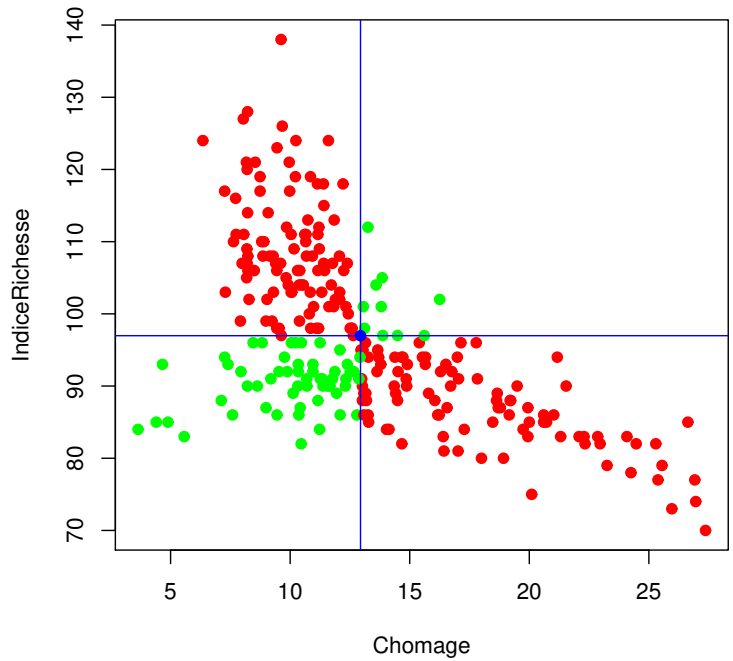
Interprétation de la valeur de r_{xy}





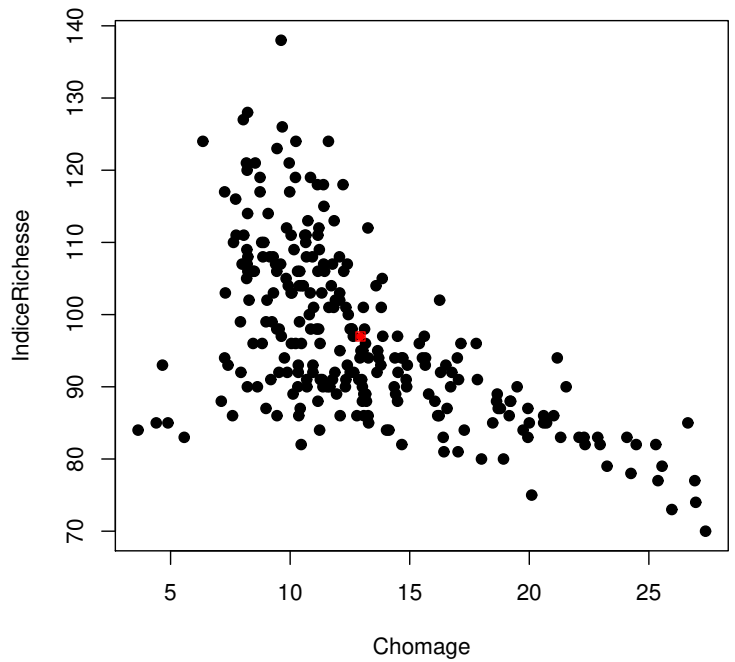
Pour le premier graphique, $r_{xy} = 0.98$ et pour le second, $s_{xy} = 0.83$

Interprétation de la valeur de r_{xy}



$r_{xy} = -0.61$

Corrélation et causalité



Un taux de chômage élévé dans une commune implique, assez logiquement, un indice de richesse faible.
Mais ce n'est pas toujours aussi évident...

Cigognes - bébés

Country	Area (km^2)	Storks (pairs)	Humans (10^6)	Birth rate (10^3/yr)
Albania	28,750	100	3.2	83
Austria	83,860	300	7.6	87
Belgium	30,520	1	9.9	118
Bulgaria	111,000	5000	9.0	117
Denmark	43,100	9	5.1	59
France	544,000	140	56	774
Germany	357,000	3300	78	901
Greece	132,000	2500	10	106
Holland	41,900	4	15	188
Hungary	93,000	5000	11	124
Italy	301,280	5	57	551
Poland	312,680	30,000	38	610
Portugal	92,390	1500	10	120
Romania	237,500	5000	23	367
Spain	504,750	8000	39	439
Switzerland	41,290	150	6.7	82
Turkey	779,450	25,000	56	1576

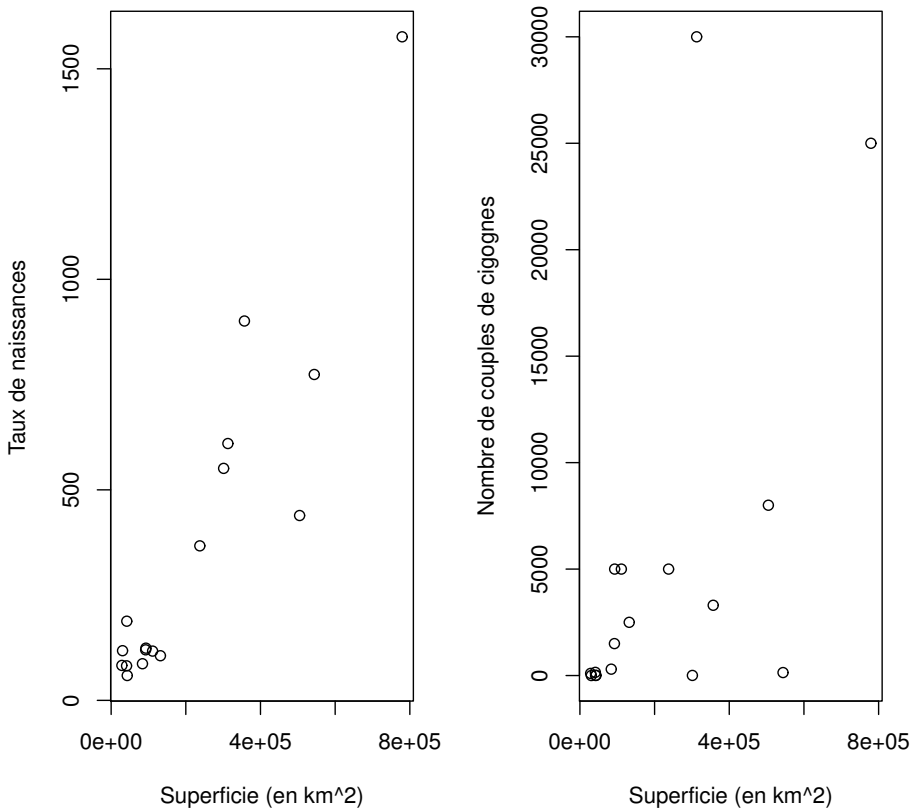
cor(Storks,Birth)= 0.62; la turquie a beaucoup de bébés car elle a beaucoup de cigognes.

$\Rightarrow$ les cigognes produisent les bébés???

NON: Corrélation n'implique pas causalité

Cigognes - bébés

Un autre facteur explique la forte corrélation entre les deux variables



cor(Area,Birth)= 0.92 et cor(Area,Storcks)= 0.58

Chocolat et Nombre de Prix Nobels

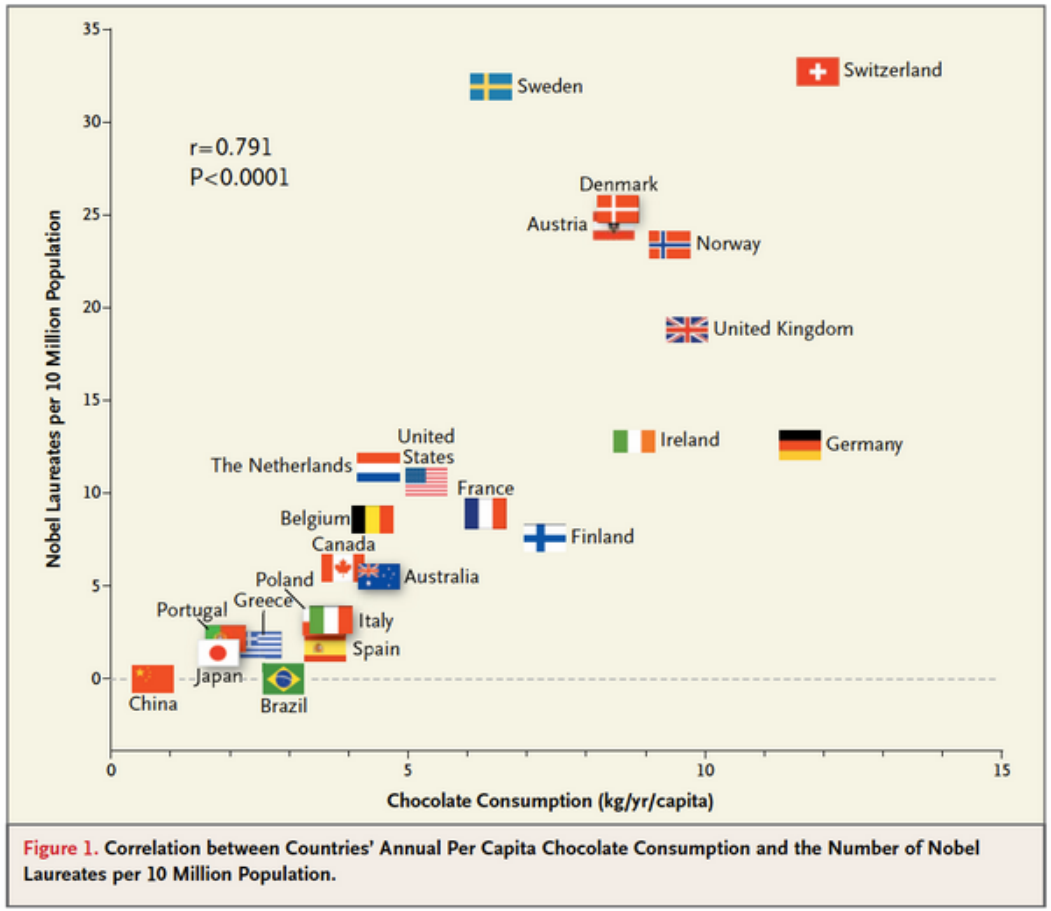


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

Correlation of chocolate consumption with Nobel Laureates (Image credit: New England Journal of Medicine)

http://www.rtbf.be/info/societe/detail_consommation-de-chocolat-et-prix-nobel-aucun-lien-selon-des-c id=8017229

Spurrious correlation

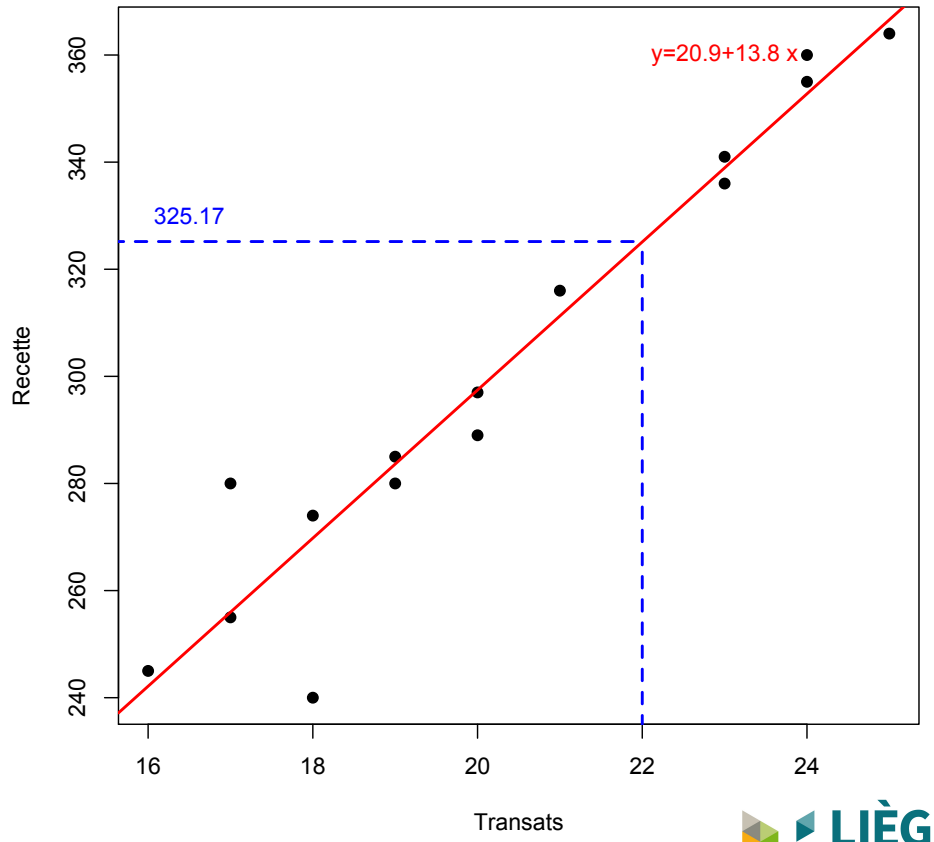
http://www.tylervigen.com/

Régression linéaire

Si le diagramme de dispersion est de tendance linéaire, on peut envisager de décrire la relation entre les deux variables à l'aide d'une droite (dite "droite de régression").

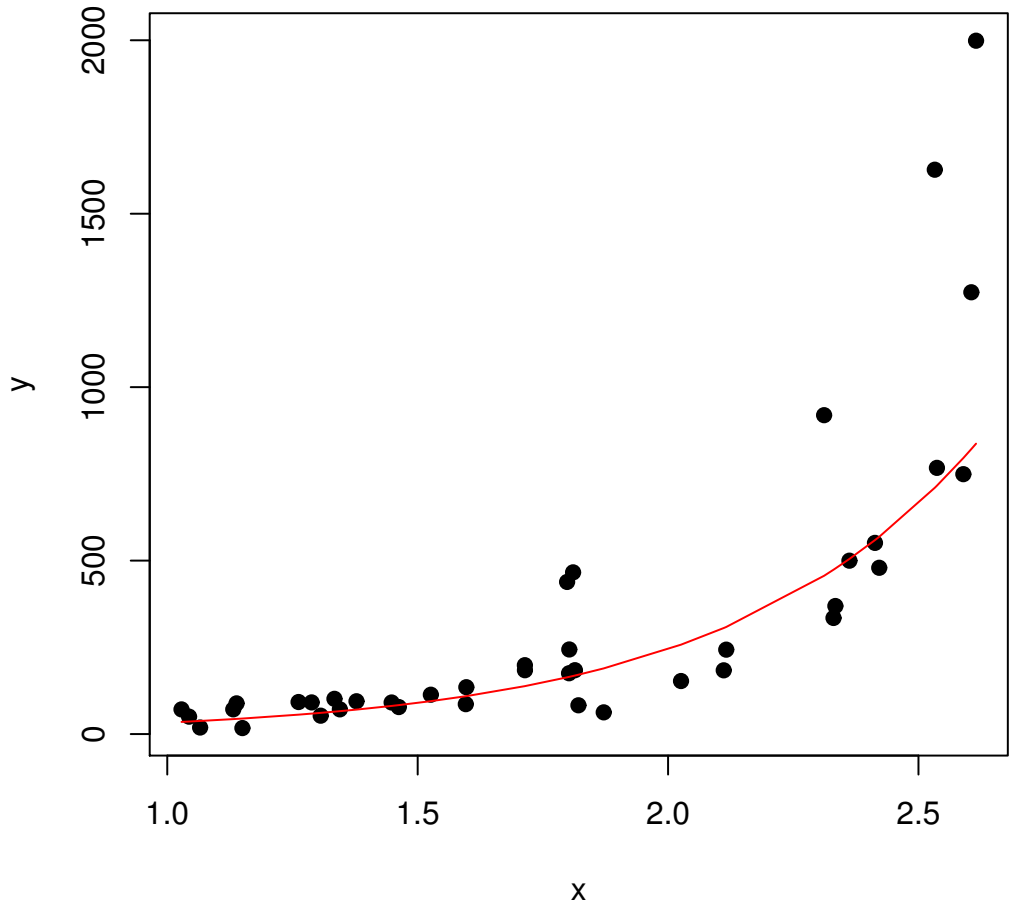
Date	Transats	Recette
01 août	17	255
02 août	19	280
03 août	18	274
04 août	20	289
05 août	23	336
06 août	25	364
07 août	19	285
08 août	16	245
09 août	24	360
10 août	23	341
11 août	21	316
12 août	17	280
13 août	18	240
14 août	20	297
15 août	24	355

$r_{xy} = 0.96$



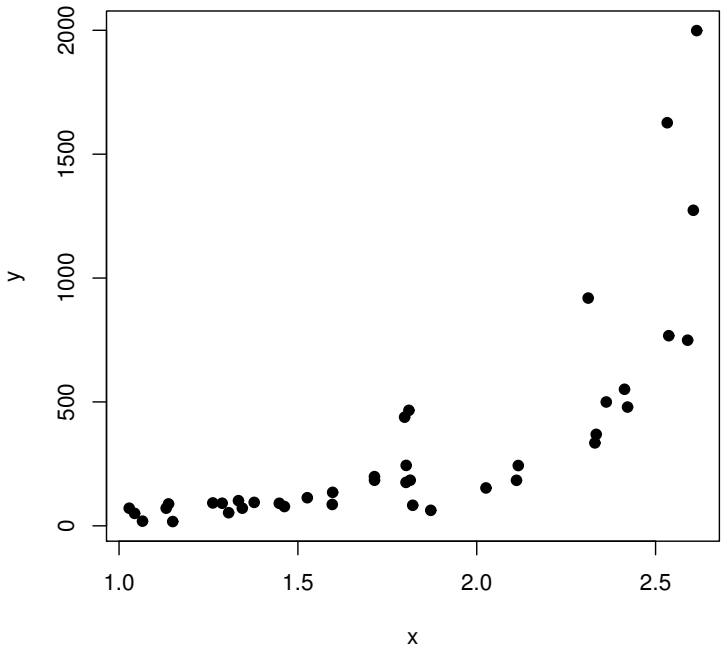
Utilité? Déterminer la valeur de Y *attendue* pour une valeur donnée de X

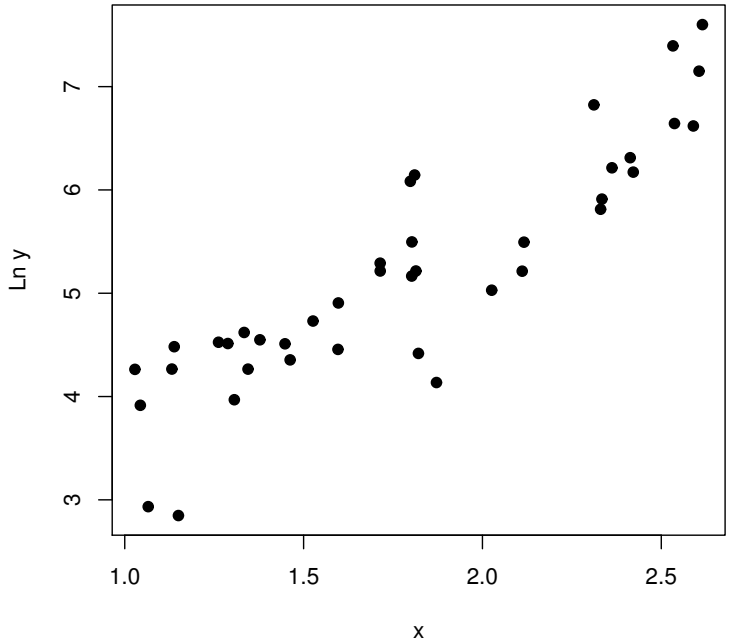
Uniquement en cas de tendance linéaire?



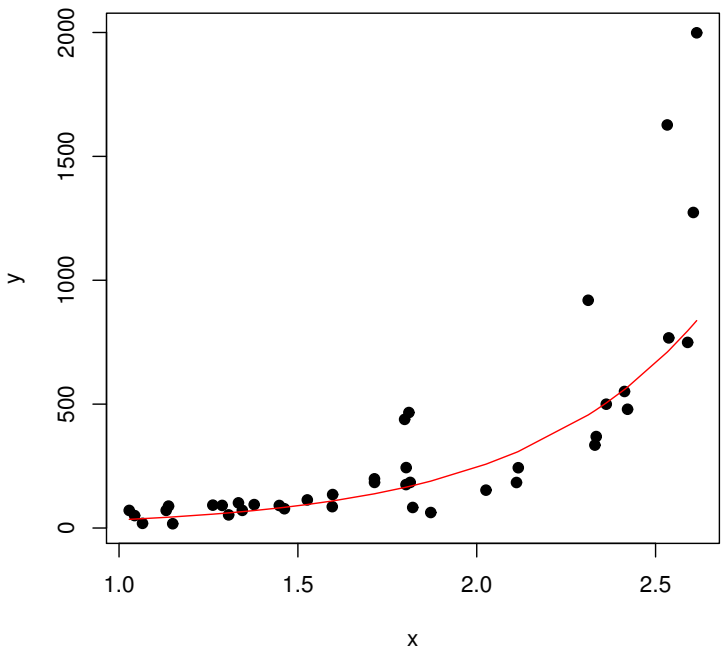
Le lien entre X et Y n'est visiblement pas linéaire...
Plutôt du type $y = ba^x \Leftrightarrow \ln y = \ln b + x \ln a$, relation linéaire entre $\ln Y$ et X!

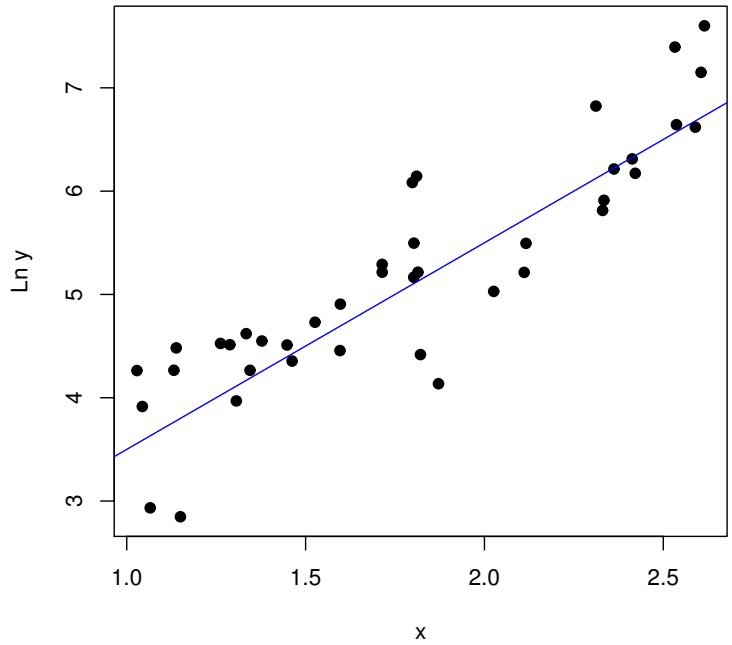
Possibilité de *linéariser* la relation





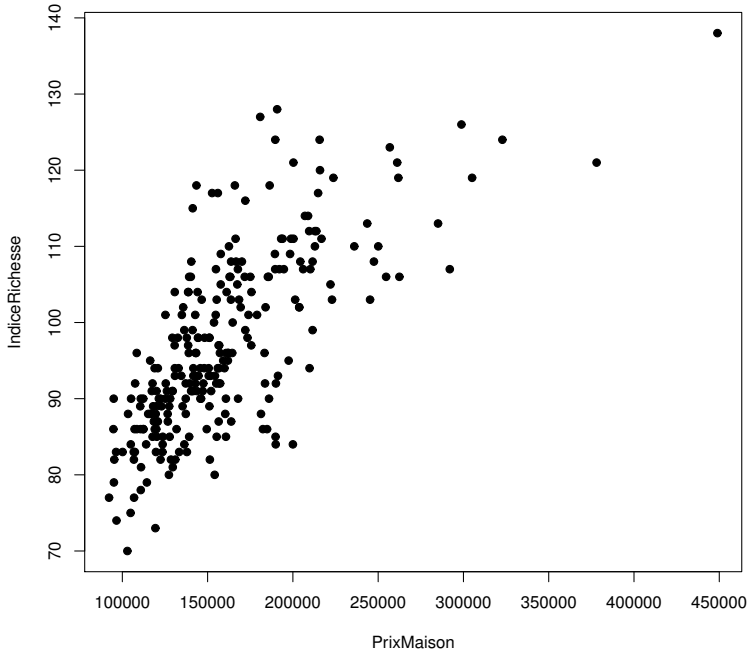
Possibilité de *linéariser* la relation

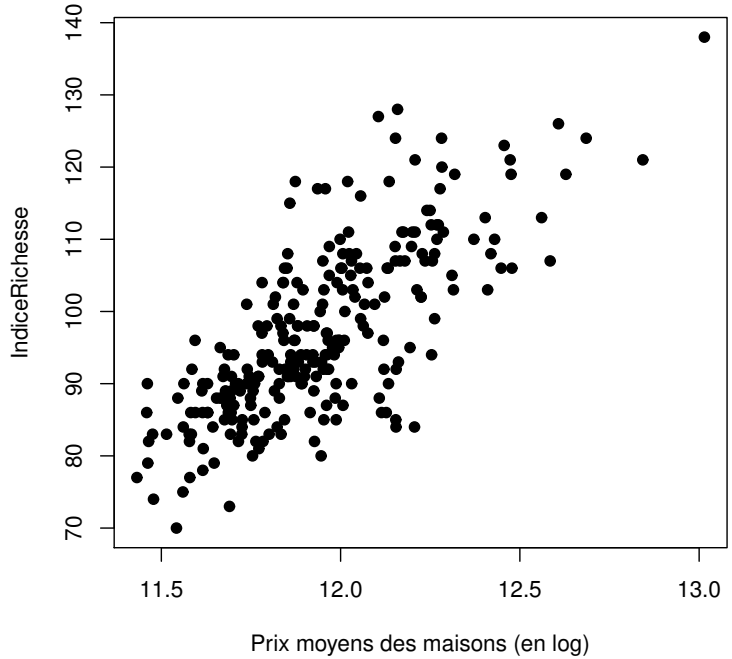




Bien entendu, le coefficient de corrélation change sous de telles transformations!

Exemple: Indice Richesse versus Prix Moyens des maisons





En passant au logarithme sur le prix moyens des maisons, la relation linéaire entre les deux variables est plus nette.

Définitions

Droite de régression de Y en X: $y = ax + b$
X variable explicative et Y variable dépendante

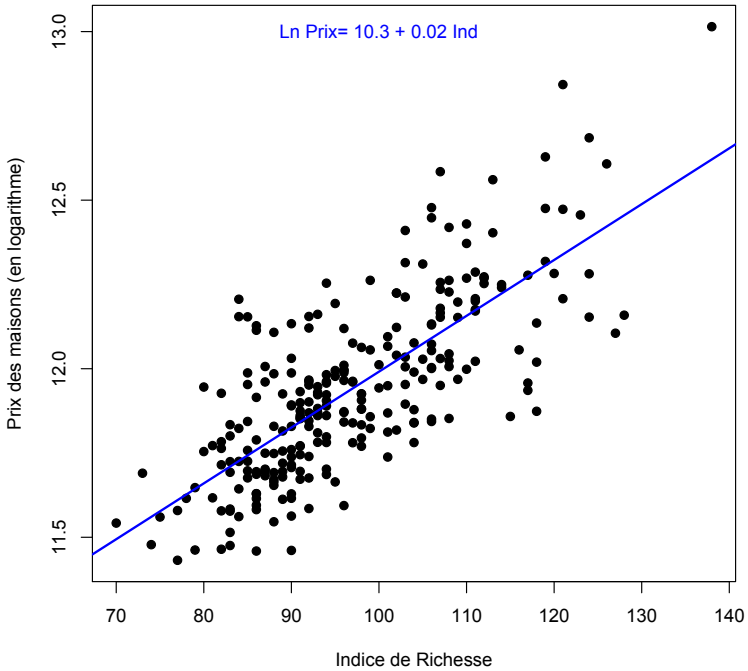
ou

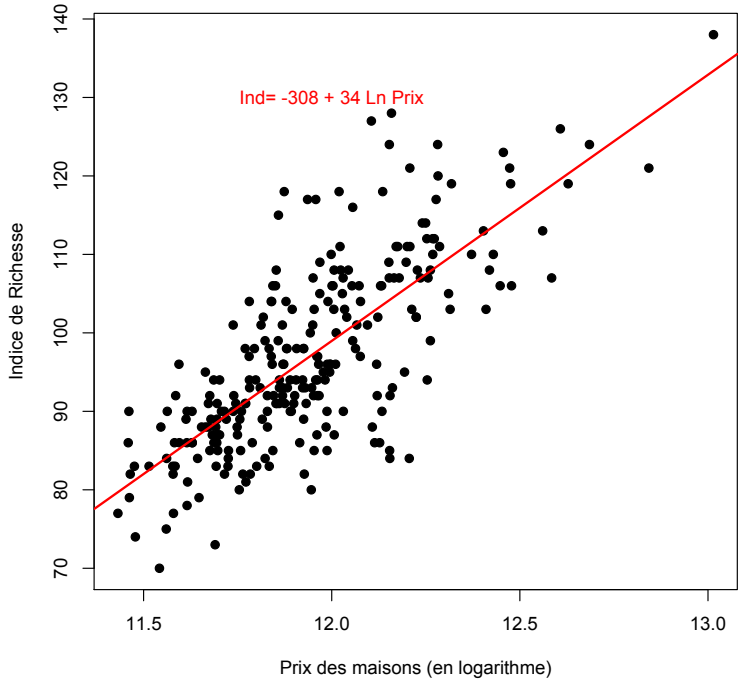
Droite de régression de X en Y: $x = a'y + b'$
X variable dépendante et Y variable explicative

Les variables X et Y ne jouent plus des rôles symétriques.

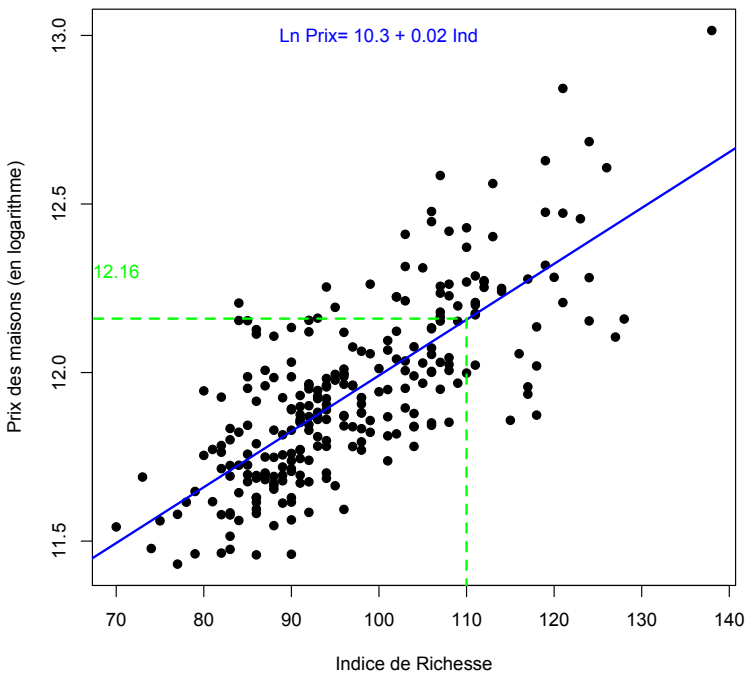
LIÈGE université

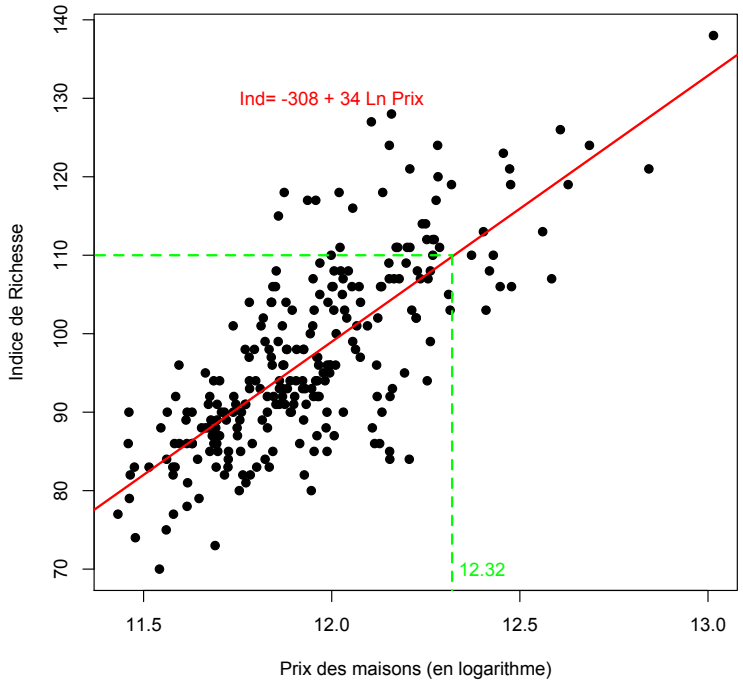
Exemple: Indice - Ln Prix $\longleftrightarrow$ Ln Prix - Indice





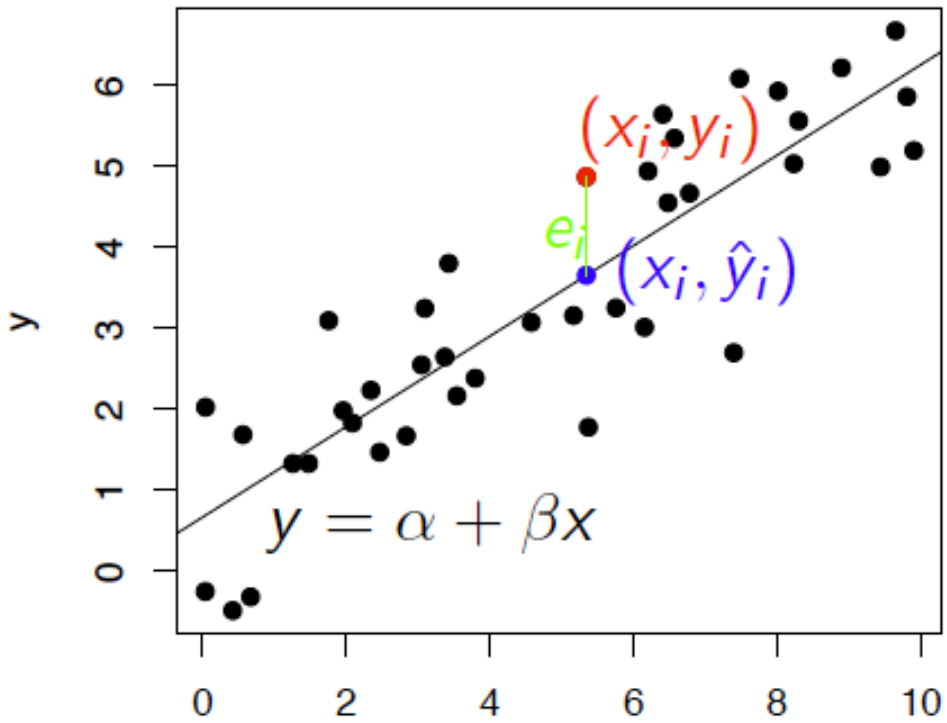
Exemple: Indice - Ln Prix $\longleftrightarrow$ Ln Prix - Indice





A un indice de richesse égal à 110 correspond un prix (en log) de 12.16 dans le premier modèle et un prix (en log) de 12.32 dans le second.

Critère des moindres carrés pour déterminer la droite de régression de Y en X

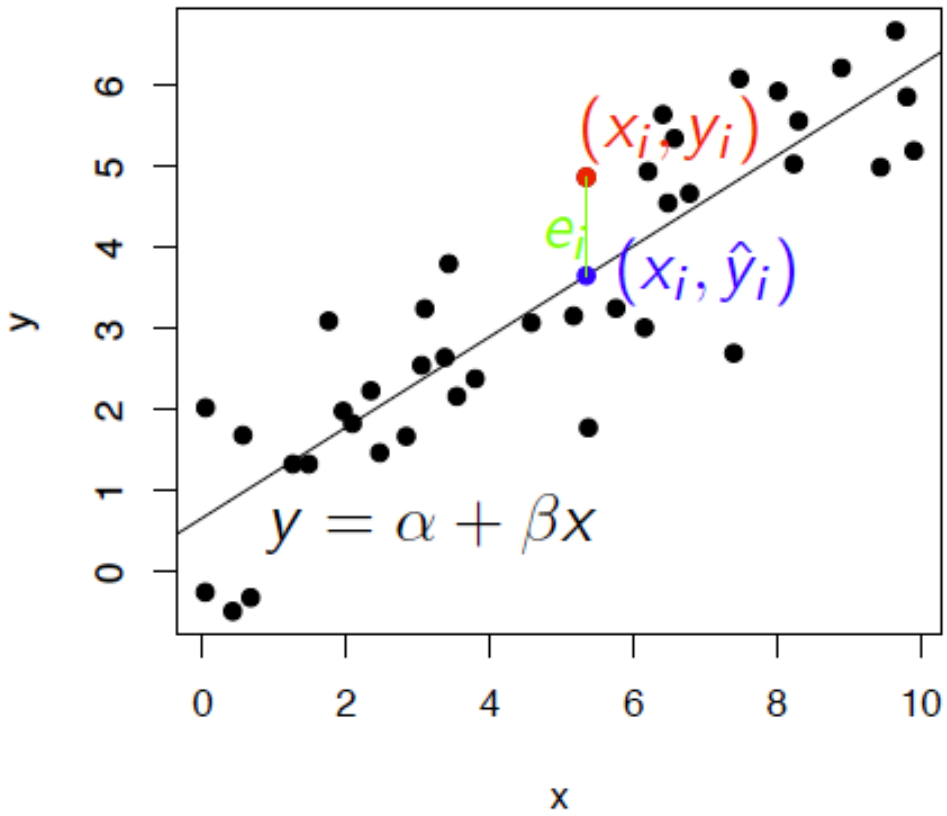


$\hat{y}_i = ax_i + b$ valeurs estimée ou ajustée par la régression

$\delta_i = y_i - \hat{y}_i$ résidu de la $i$ème observation

Critère des moindres carrés pour déterminer la droite de régression de Y en X

Déterminer a et b de manière à minimiser la somme des carrés des écarts entre les valeurs observées y_i et les valeurs correspondant aux abscisses x_i et situées sur la droite.



Droite des moindres carrés de Y en X

Trouver a et b de manière à

$$\min_{a,b} \sum_{i=1}^{n} \delta_i^2 = \min_{a,b} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Proposition

Soit $S = \{(x_i, y_i), 1 \leq i \leq n\}$ une série statistique (quantitative) bivariée. L'équation de la droite de régression par la méthode des moindres carrés est donnée par

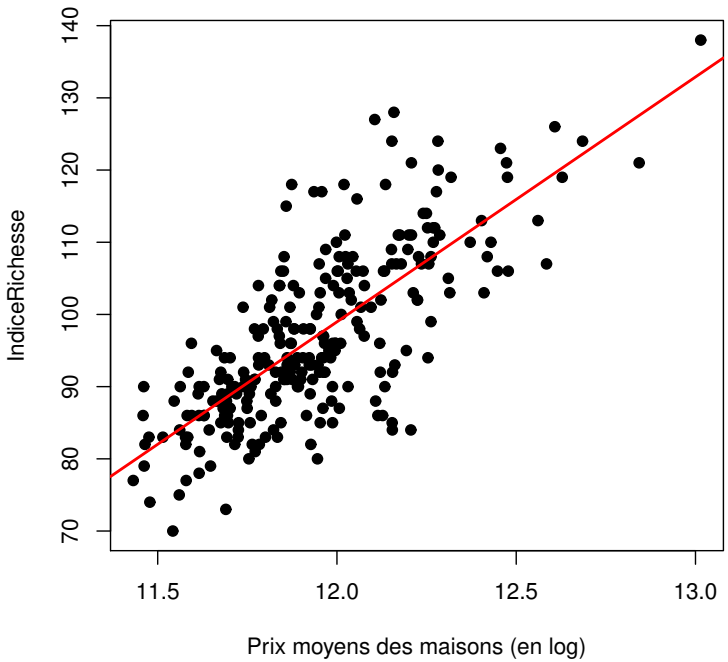
$$y = \hat{a}x + \hat{b} \text{ où } \hat{a} = \frac{s_{xy}}{s_x^2} \text{ et } \hat{b} = \bar{y} - \hat{a}\bar{x}.$$

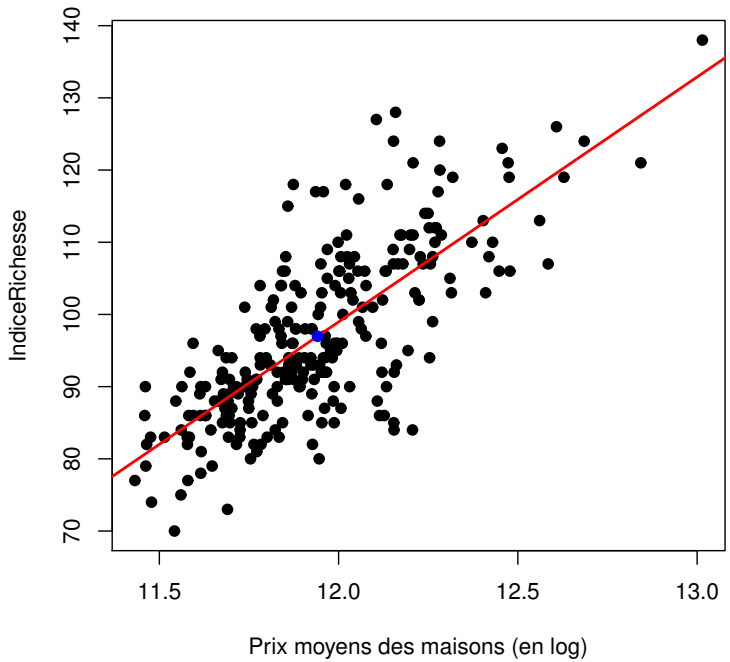
Résultat admis.

Illustration: droite de régression de l'indice de richesse Y sur le log des prix X

$s_{xy} = 2.339$ et $s_x^2 = 0.069 \Rightarrow \hat{a} = \frac{2.339}{0.069} = 33.9$

$\bar{y} = 96.973$ et $\bar{x} = 11.941 \Rightarrow \hat{b} = -307.9.$



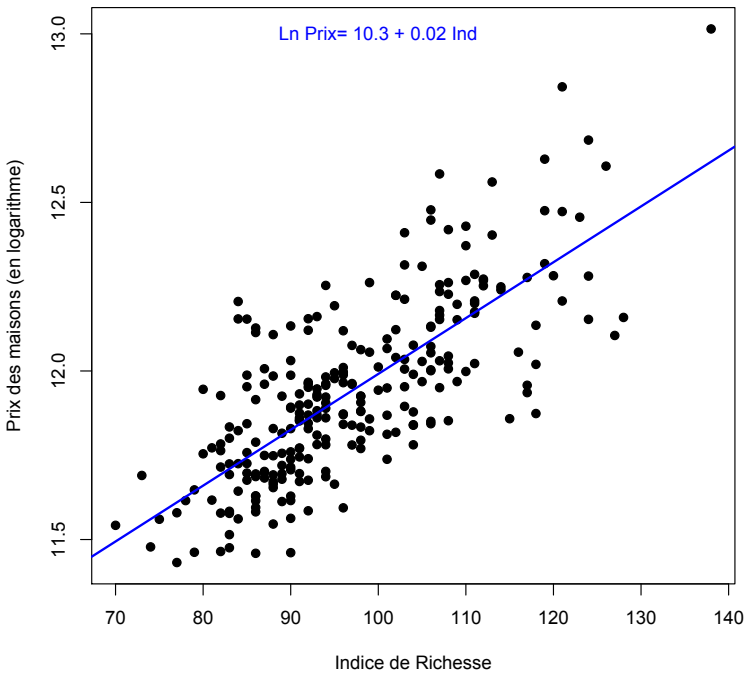


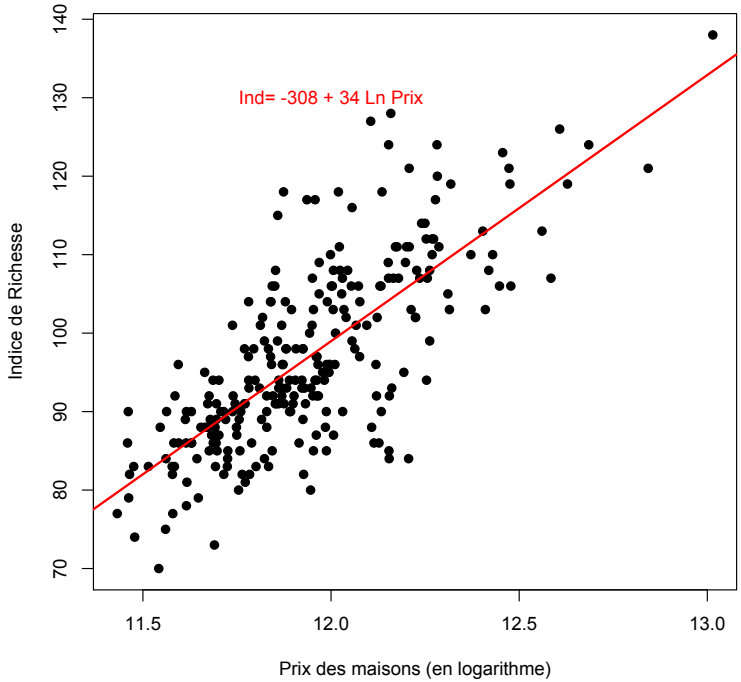
Toute droite de régression passe par le centre de gravité des données!

Droite de régression de Y en X et droite de régression de X en Y

Droite de régression de Y en X: $y = \hat{a}x + \hat{b}$ avec $\hat{a} = \frac{s_{xy}}{s_x^2}$ et $\hat{b} = \bar{y} - \hat{a}\bar{x}$

Droite de régression de X en Y: $x = \hat{a}'y + \hat{b}'$ avec $\hat{a}' = \frac{s_{xy}}{s_y^2}$ et $\hat{b}' = \bar{x} - \hat{a}'\bar{y}$





Comparaison des deux droites

Afin de comparer les estimations basées sur les deux équations de droite, il serait plus simple de les exprimer dans le même repère (par exemple sous la forme de deux équations exprimant y en fonction de x).

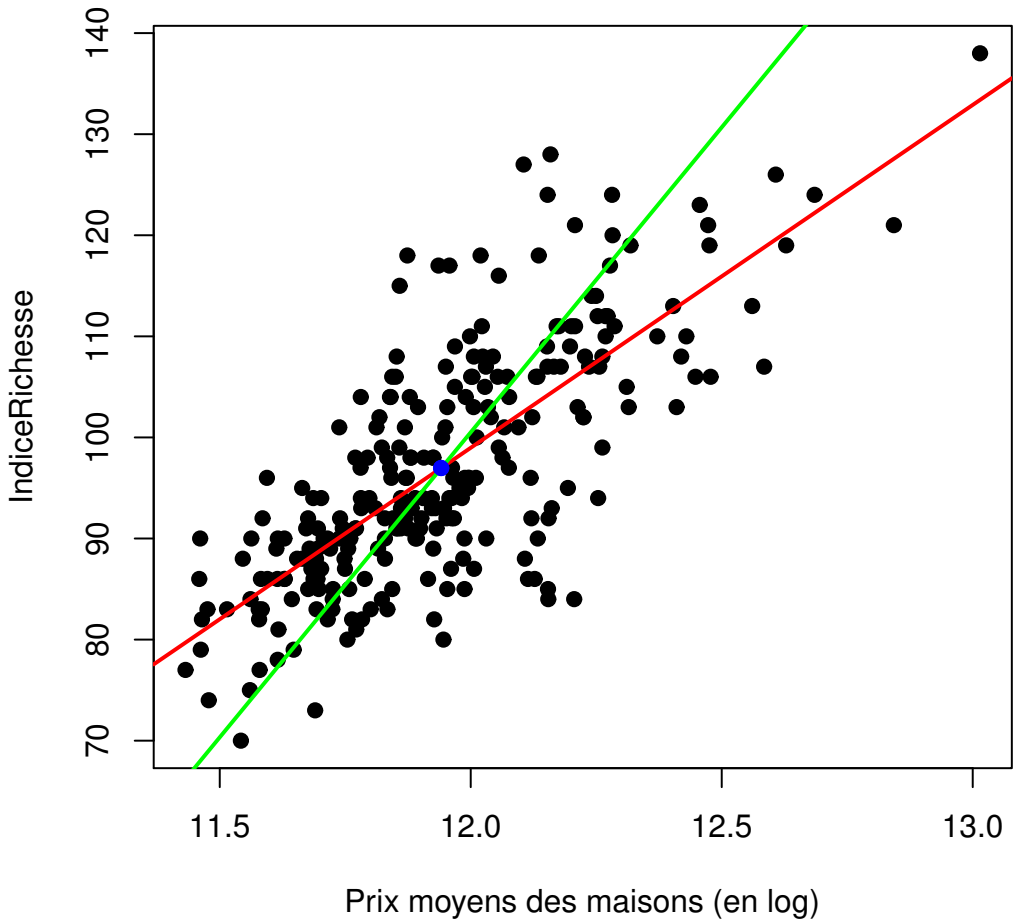
$$x = \hat{a}'y + \hat{b}' \Leftrightarrow \hat{a}'y = x - \hat{b}' \Leftrightarrow y = \frac{1}{\hat{a}'}x - \frac{\hat{b}'}{\hat{a}'}$$

Sous cette forme, les deux droites de régression

$$y = \hat{a}x + \hat{b} \text{ et } y = \frac{1}{\hat{a}'}x - \frac{\hat{b}'}{\hat{a}'}$$

peuvent être représentées dans le même repère. Elles passent par le même point, $(\bar{x}, \bar{y})$, mais ont des coefficients angulaires égaux à $\hat{a}$ et $\frac{1}{\hat{a}'}$ respectivement

Illustration: les deux droites de régression entre l'indice de richesse et le prix des maisons (en log)



Quand les deux droites de régression sont-elles confondues?

Deux droites passant par le même point sont confondues si leurs coefficients angulaires sont les mêmes.

$$\hat{a} = \frac{1}{\hat{a}'} \Leftrightarrow \frac{s_{xy}}{s_x^2} = \frac{s_y^2}{s_{xy}} \Leftrightarrow s_{xy}^2 = s_x^2 s_y^2 \Leftrightarrow |r_{xy}| = 1$$

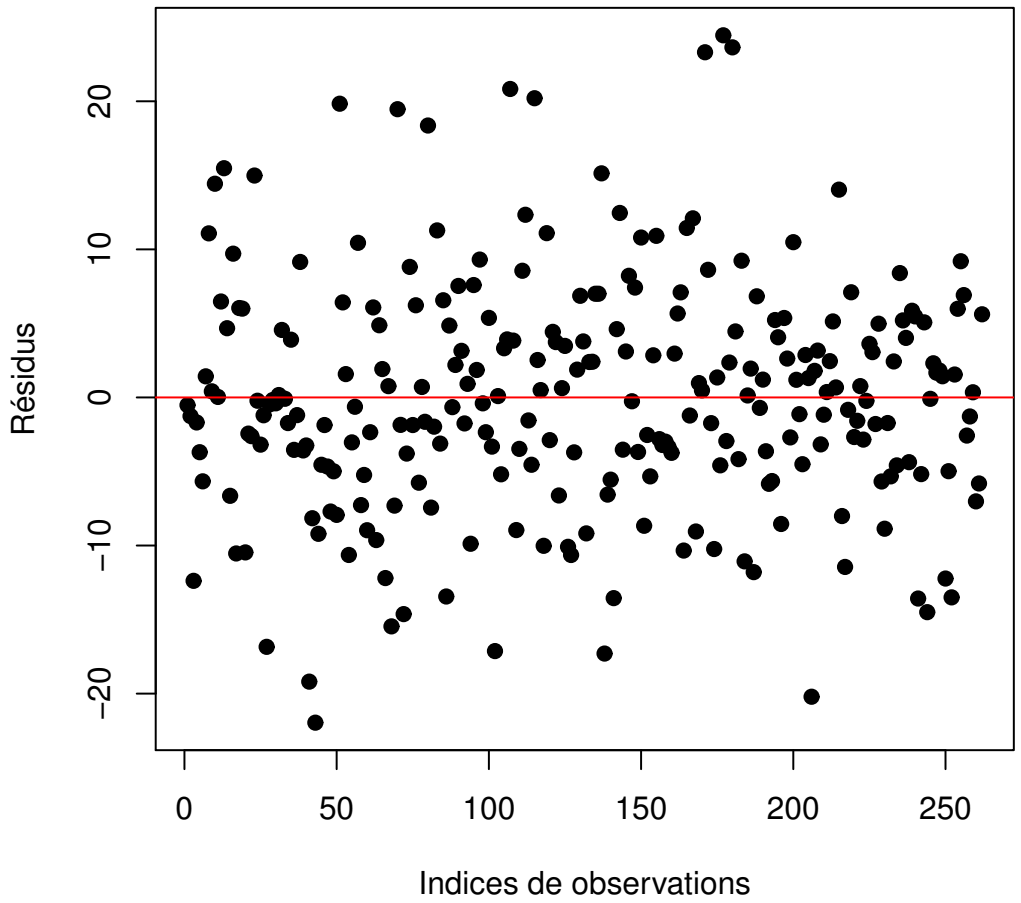
Analyse des résidus et des valeurs ajustées

1. La série des résidus: $S_r = \{\delta_1, \ldots, \delta_n\}$ où $\delta_i = y_i - \hat{y}_i$.
2. La série des valeurs ajustées (ou estimées): $S_{\text{est}} = \{\hat{y}_1, \ldots, \hat{y}_n\}$ où $\hat{y}_i = \hat{a}x_i + \hat{b}$.

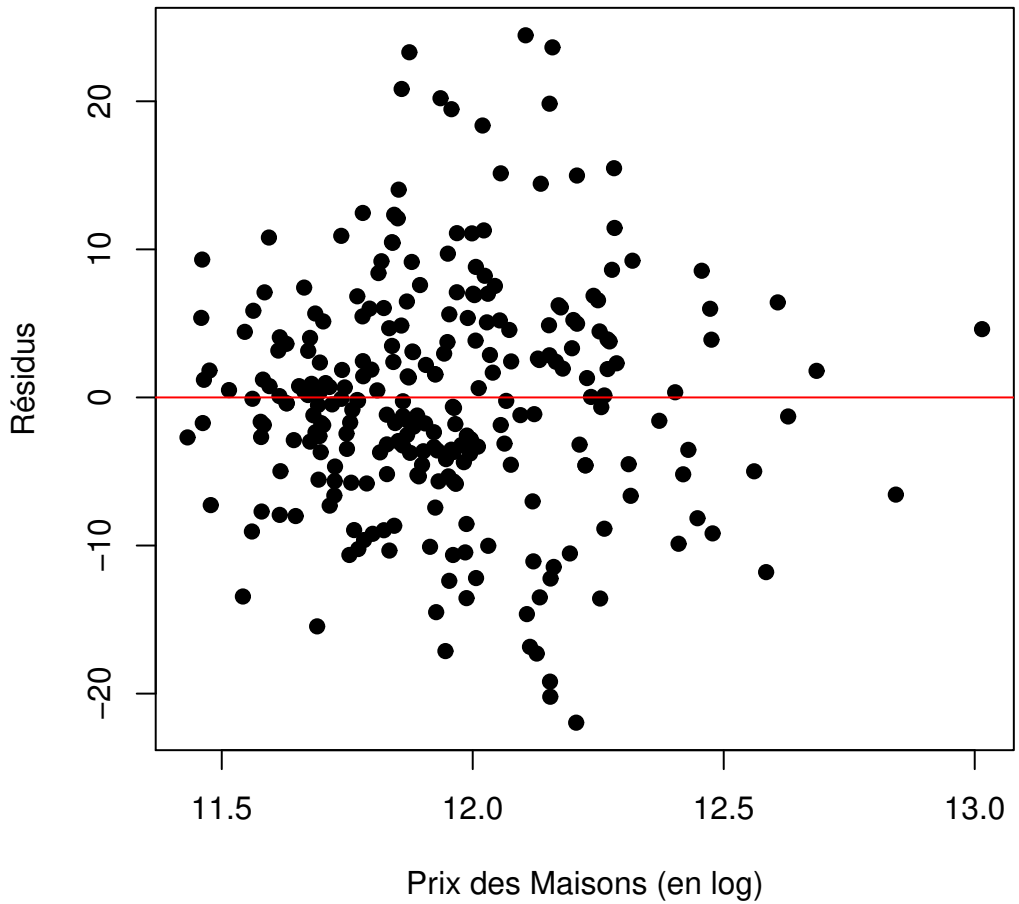
La qualité de l'ajustement peut se mesurer par une analyse graphique des résidus.

C'est une absence de structure dans les résidus que l'on aimerait observer!

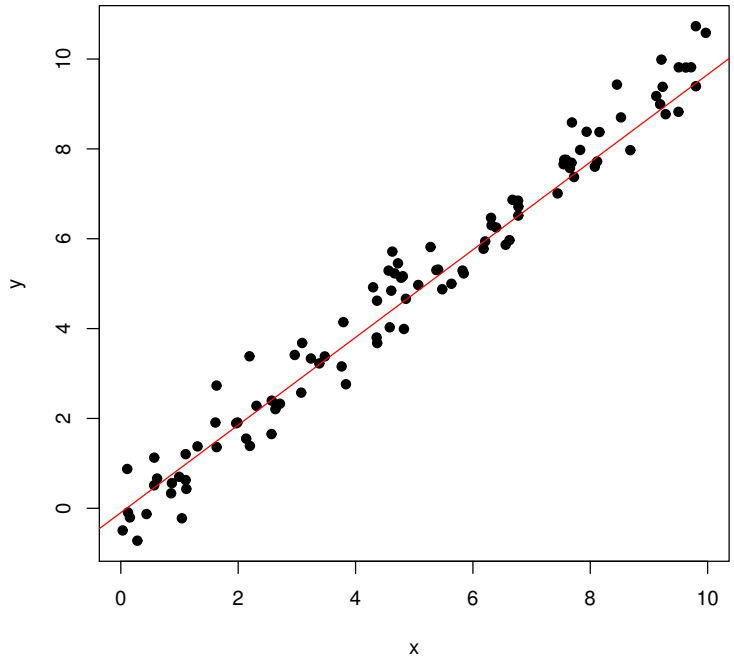
Résidus de la droite de régression de l'IR sur le Prix en fonction des indices des observations

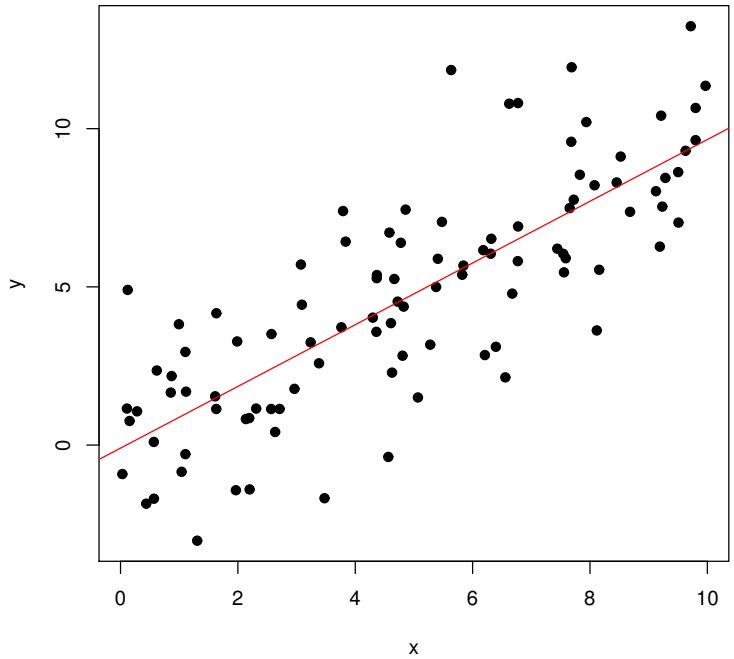


Résidus de la droite de régression de l'IR sur le Prix en fonction de la variable X (log du prix)

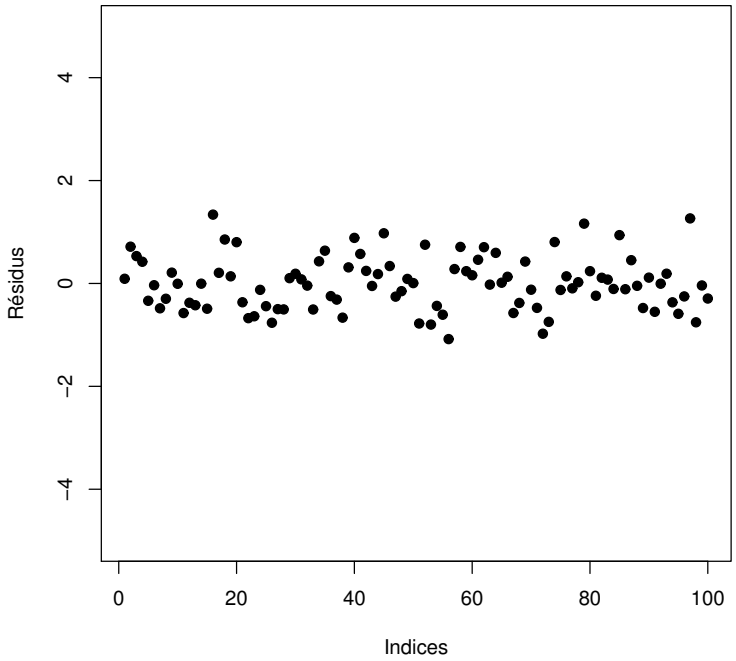


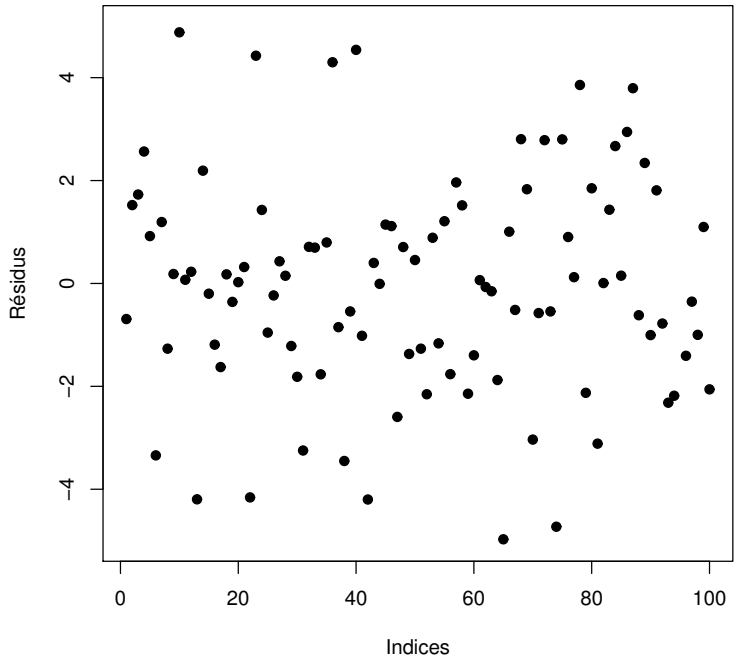
Autres exemples de droites de régression



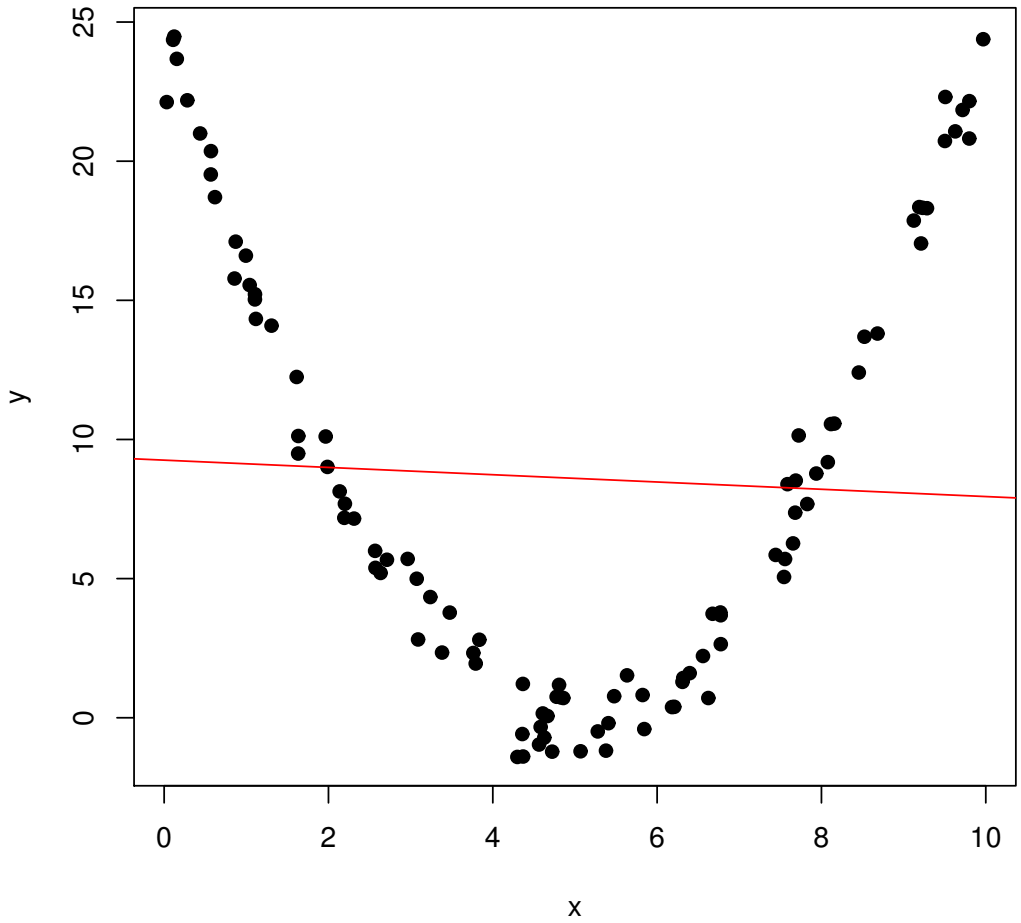


Graphiques des résidus correspondants

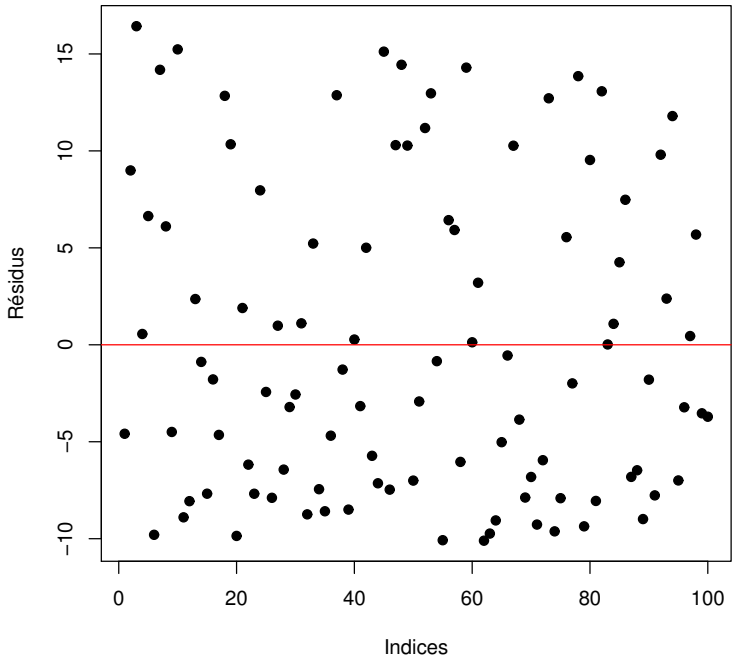


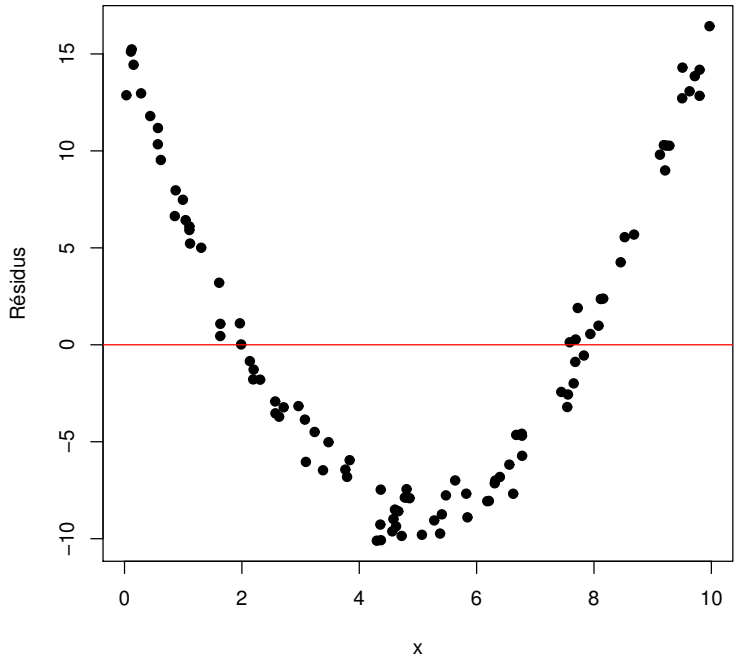


Exemple “quadratique”



Exemple “quadratique” : graphique des résidus





Moyenne et variance des séries des résidus et des valeurs ajustées

Proposition

La moyenne de la série des résidus est nulle: $\bar{\delta} = 0$
La variance (appelée variance résiduelle de y par rapport à x), vaut $s_\delta^2 = s_y^2(1 - r^2)$ *où* $r = r_{xy}$.

Proposition

La moyenne de la série des valeurs ajustées est égale à la moyenne des valeurs observées: $\bar{y} = \bar{\hat{y}}$
La variance de la série des valeurs ajustées est égale à $s_{\hat{y}}^2 = r^2 s_y^2$.

Démonstration:
Outils: exploitation des définitions de la moyenne et de la variance et des formules donnant les paramètres $\hat{a}$ et $\hat{b}$.
Au tableau

Décomposition de la variance

$$s_y^2 s_y^2 = 1 \times s_y^2 = (1 - r^2 + r^2) \times s_y^2 = (1 - r^2) s_y^2 (1 - r^2) s_y^2 + r^2 s_y^2 r^2 s_y^2$$

Variance résiduelle (non expliquée par la variable X): s_δ^2

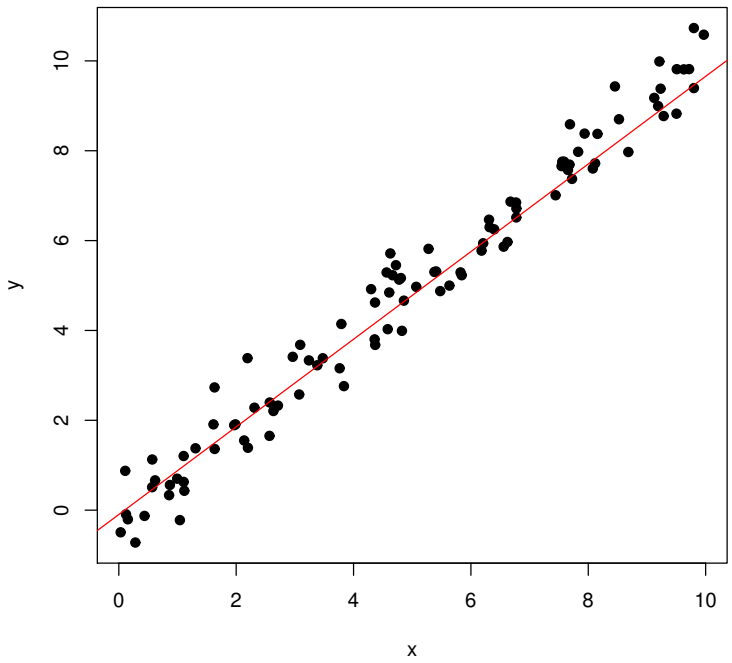
Variance expliquée par la régression linéaire (part de la variabilité totale expliquée par la variable X): $s_{\hat{y}}^2$

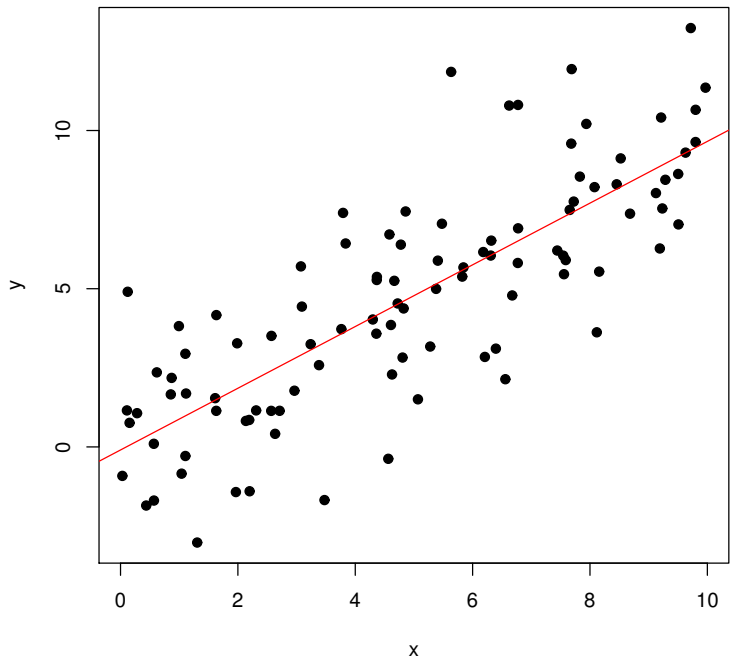
On appelle *coefficient de détermination* le paramètre

$$R^2 = \frac{s_{\hat{y}}^2}{s_y^2} = \frac{\text{variance expliquée}}{\text{variance totale}}$$

LIÈGE université

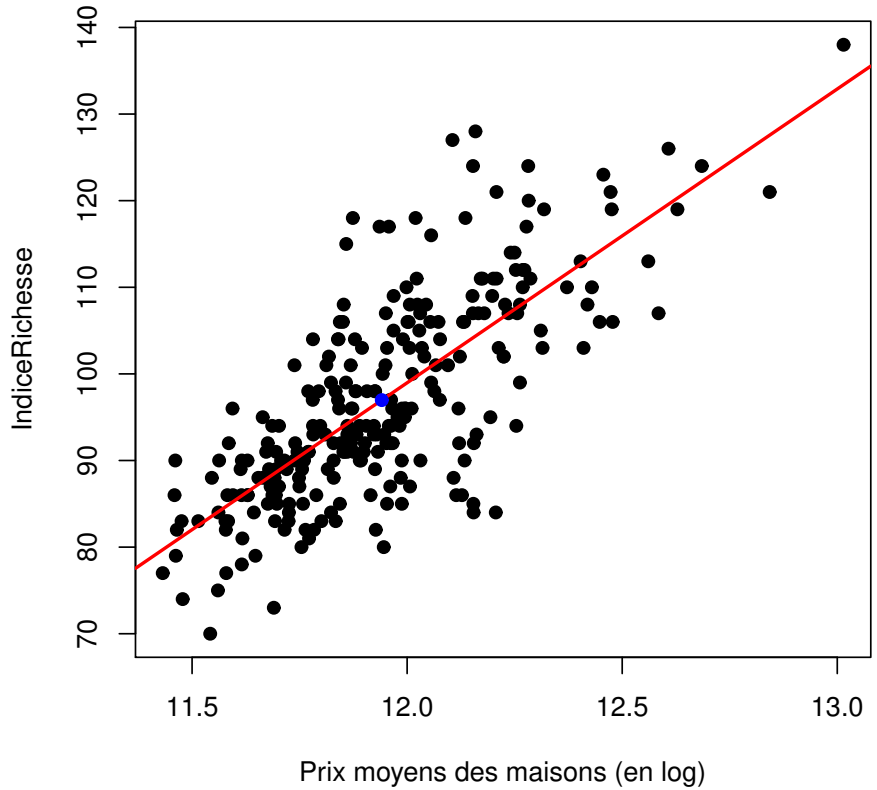
Exemples





Pour le premier graphique, $R^2 = 0.97$ et pour le second, $R^2 = 0.64$

Relation Indice de Richesse – Prix des Maisons (en log)



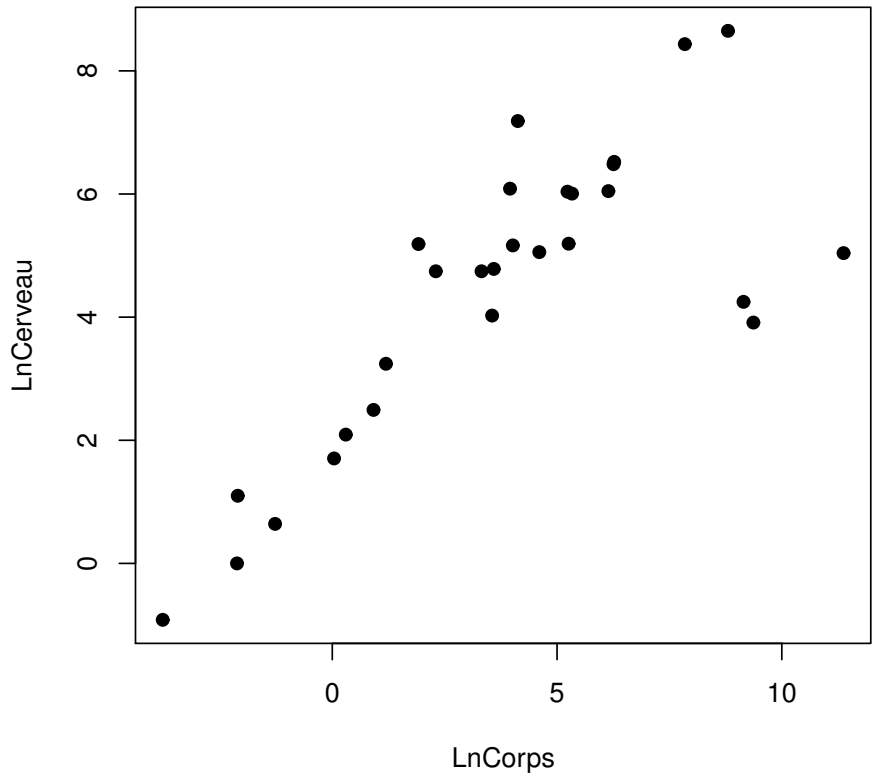
$R^2 = 0.56$

Relation Poids du cerveau – Poids du corps

i	Nom	Poids du corps (Kg)	Poids du cerveau (g)
1	Castor	1.35	8.1
2	Vache	465	423
3	Loup gris	36.33	119.5
4	Chèvre	27.66	115
5	Cobaye	1.04	5.5
6	Diplodocus	11700	50
7	Éléphant d'Asie	2547	4603
8	Âne	187.1	419
9	Cheval	521	655
10	Babouin	10	115
11	Chat	3.3	25.6
12	Girafe	529	680
13	Gorille	207	406
14	Humain	62	1320
15	Éléphant d'Afrique	6654	5712
16	Triceratops	9400	70
17	Singe Rhesus	6.8	179
18	Kangourou	35	56
19	Hamster	0.12	1
20	Souris	0.023	0.4
21	Lapin	2.5	12.1
22	Mouton	55.5	175
23	Jaguar	100	157
24	Chimpanzé	52.16	440
25	Brachiosaurus	87000	154.5
26	Rat	0.280	1.9
27	Taupe	0.122	3
28	Cochon	192	180

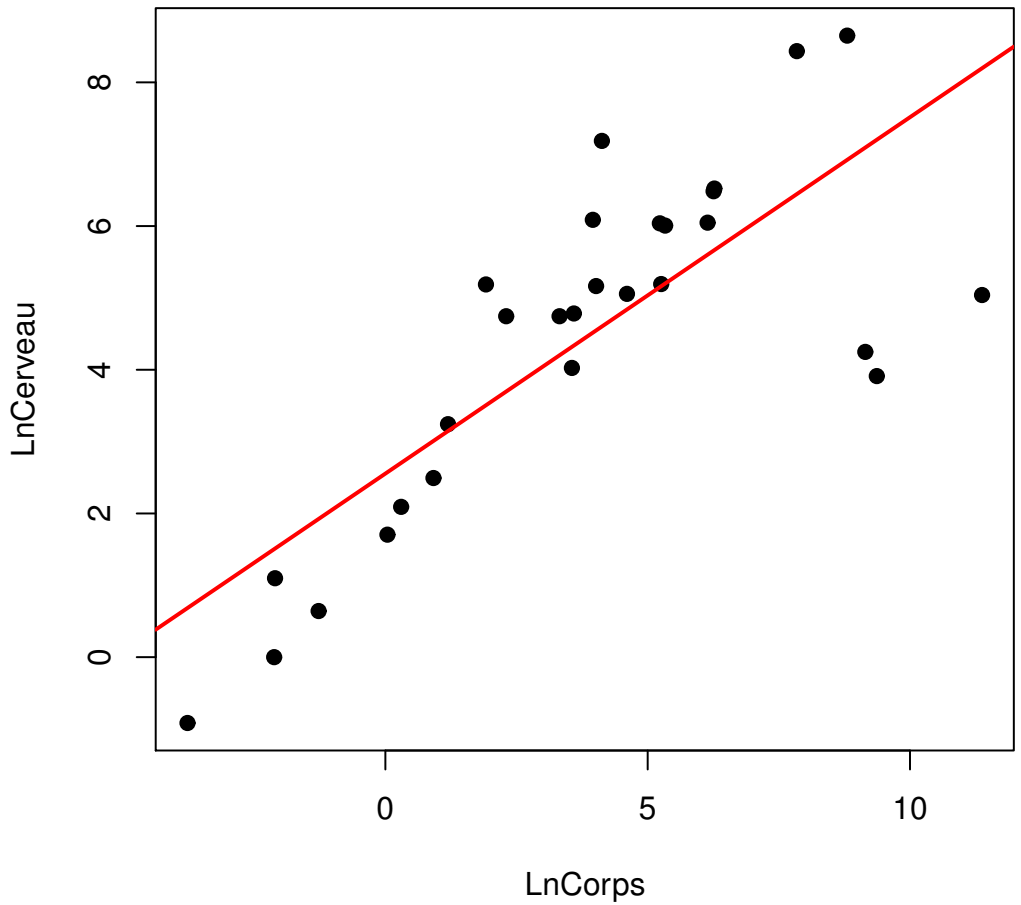
LIÈGE université

Transformation logarithmique



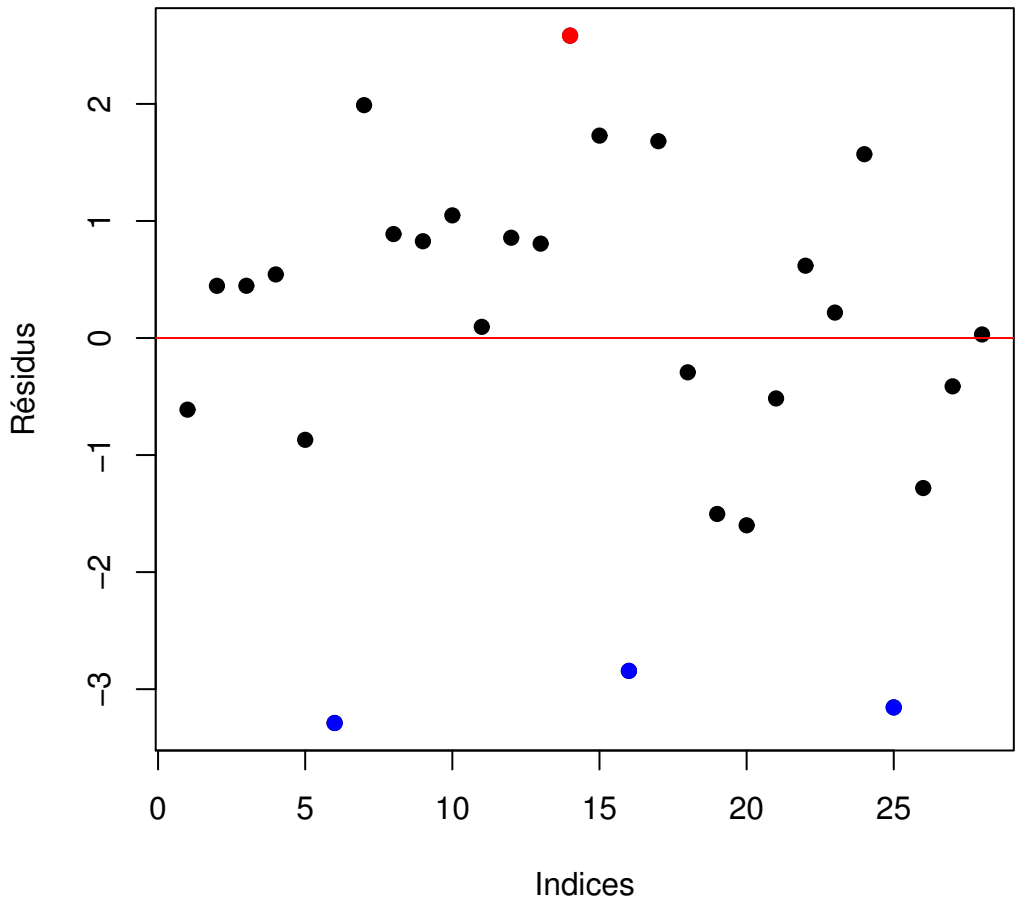
Une relation du type $\ln \texttt{Cerveau} = a \ln \texttt{Corps} + b$ parait adéquate (corrélation: 0.77).

Estimation par moindres carrés



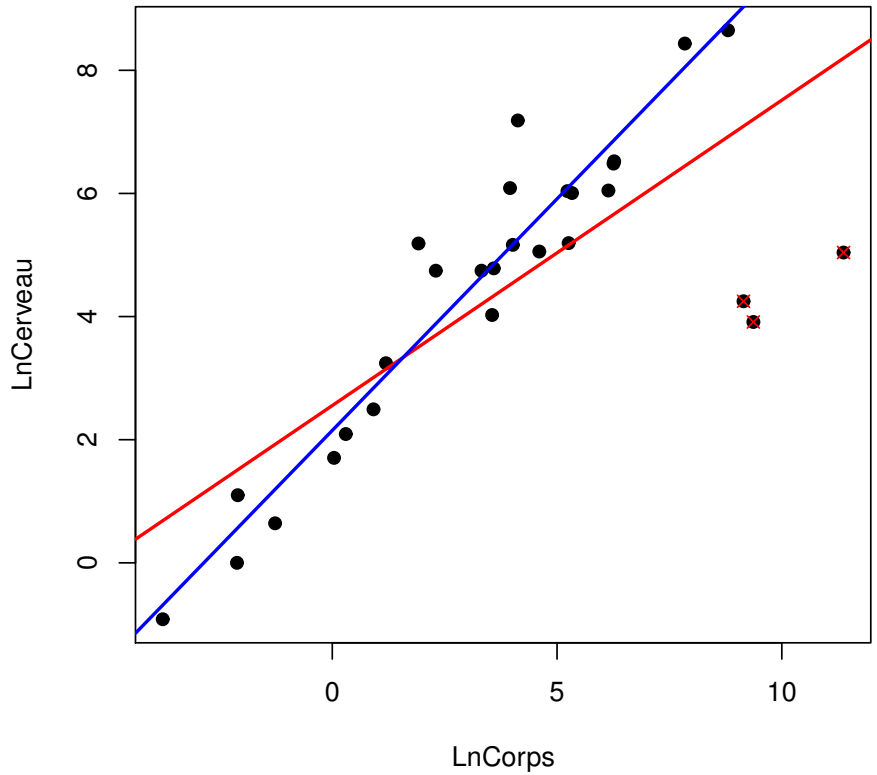
Variance expliquée: 0.61

Analyse des résidus



L'humain est presqu'aussi atypique que les dinosaures!

En supprimant les dinosaures...



Clairement, les dinosaures ne respectent pas la tendance générale. Si on les supprime, la corrélation passe à 0.96!
La droite s'ajuste mieux aux "bonnes" données (et $R^2 = 0.92$)

En conclusion

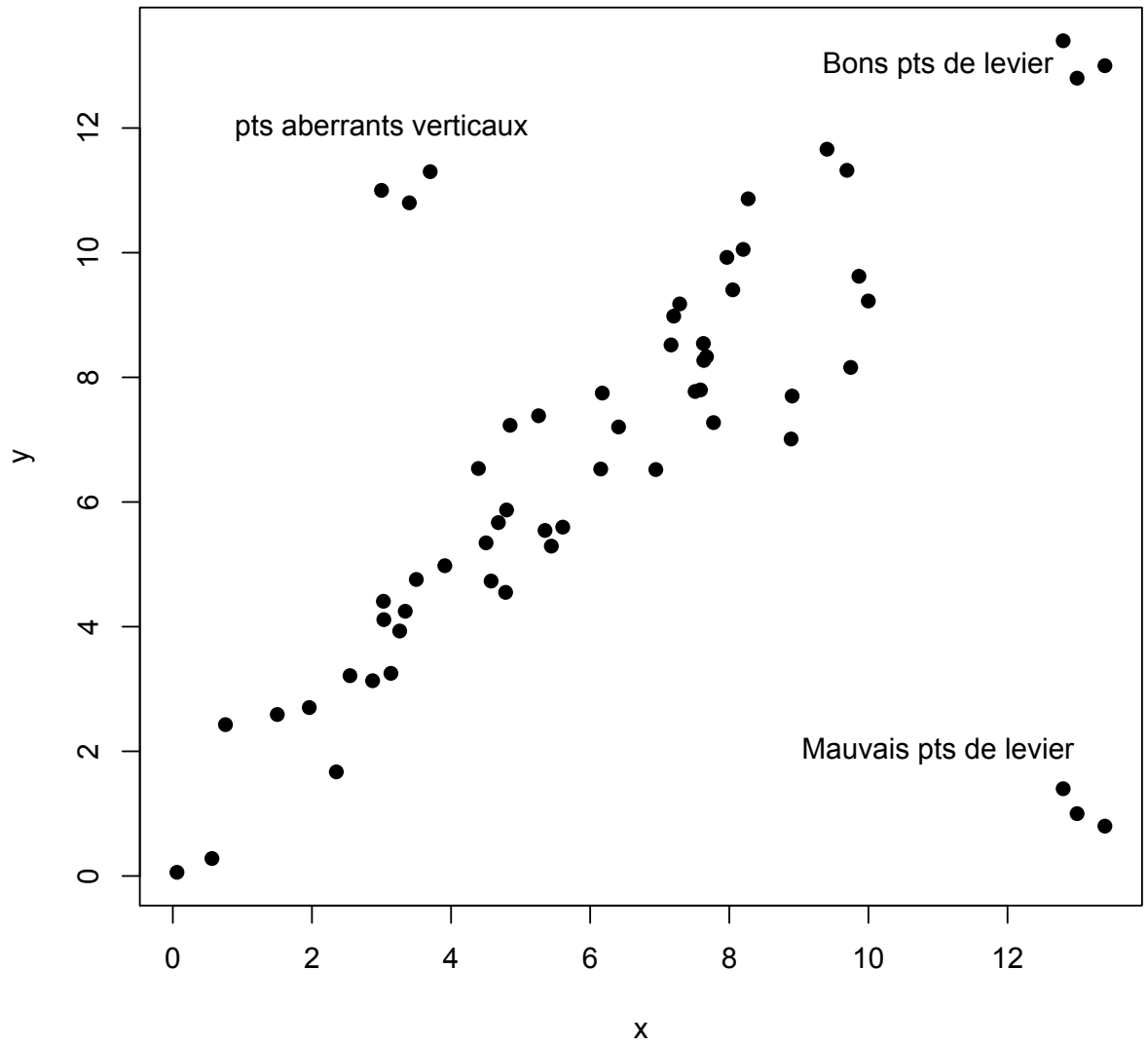
La technique des moindres carrés

✓ correspond à des estimations de a et b très simples à calculer;
✓ se généralise sans problème à la dimension p;

MAIS

✗ est très sensible à la présence d'observations atypiques...

Observations atypiques en régression linéaire



Raisons de la "sensibilité" des moindres carrés

Le critère des moindres carrés (ou critère L_2) correspond au problème d'optimisation suivant:

$$\min_{a,b \in \boldsymbol{R}} \sum_{i=1}^{n} r_i^2(a,b)$$

où $r_i(a,b) = y_i - \hat{y}_i = y_i - ax_i - b$.

Le manque de robustesse est dû essentiellement aux deux points suivants:

- Les résidus sont élevés au carrés;
- Tous les résidus sont exploités.

"Robustifications" de la technique des moindres carrés (L_2)

Deux approches principales:

1. Modification de la fonction des résidus:

$$\min_{a,b \in \boldsymbol{R}} \sum_{i=1}^{n} r_i^2(a,b) \Leftrightarrow \min_{a,b \in \boldsymbol{R}} \sum_{i=1}^{n} \rho(r_i(a,b))$$

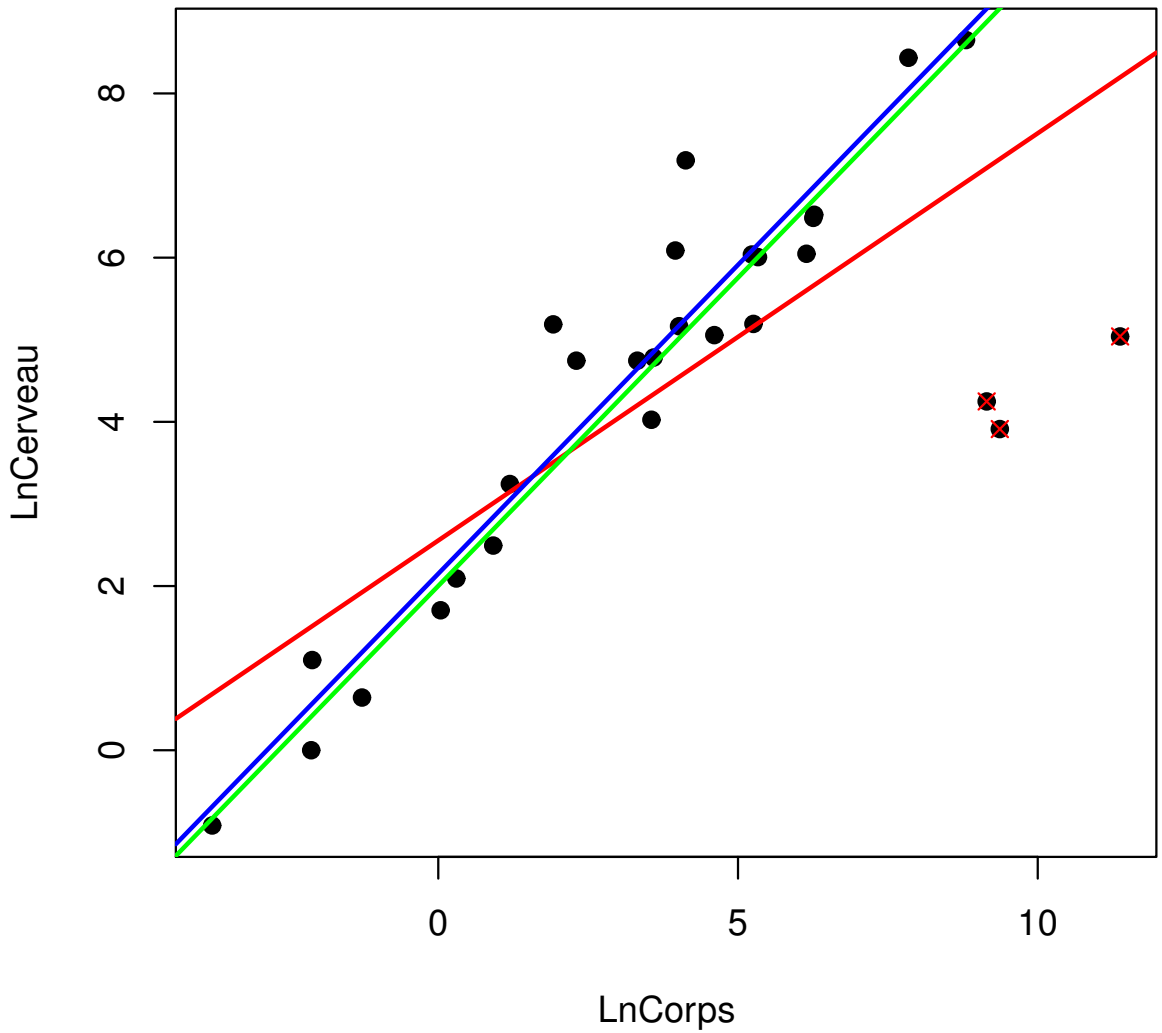
avec $\rho : t \rightarrow t^2$.
D'où, en changeant la fonction ρ, la technique peut être robustifiée. Notamment, $\rho(t) = |t|$ donne la technique L_1. Beaucoup d'autres propositions existent dans la littérature!

2. Elimination des (trop) grands résidus:

$$\min_{a,b \in \boldsymbol{R}} \sum_{i=1}^{n} r_i^2(a,b) \Rightarrow \min_{a,b \in \boldsymbol{R}} \sum_{i=1}^{h} r_{(i)}^2(a,b)$$

En privilégiant une somme tronquée, on obtient le critère LTS (Least Trimmed Squares).

Retour aux données



Ce qu'il faut retenir de ce chapitre

Compétences pratiques:

- Construction d'un tableau de contingence complet;
- Analyse (y compris résumé) des séries marginales et conditionnelles;
- Définition, interprétation et exploitation de la covariance et de la corrélation;
- Ajustement par la technique des moindres carrés (pour Y en fonction de X ou pour X en fonction de Y, ou après une transformation éventuelle);
- Analyse de la qualité du modèle ajusté;
- Exploitation de la formule de décomposition de la variance de la variable dépendante;
- Exploitation d'une technique robuste (L_1 ou LTS).

LIÈGE université

Ce qu'il faut retenir de ce chapitre

Théorie:

1. Savoir citer et démontrer les propriétés suivantes:
   - Effet d'un changement d'origine et d'échelle sur la covariance et la corrélation;
   - Formule équivalente $s_{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}$;
   - Moyennes et variances des séries des résidus et des valeurs ajustées et décomposition de la variance.
2. Savoir citer et interpréter les propriétés suivantes:
   - Borne sur la covariance;
   - Equation de la droite de régression par la technique des moindres carrés.

LIÈGE université