Statistiques sur R

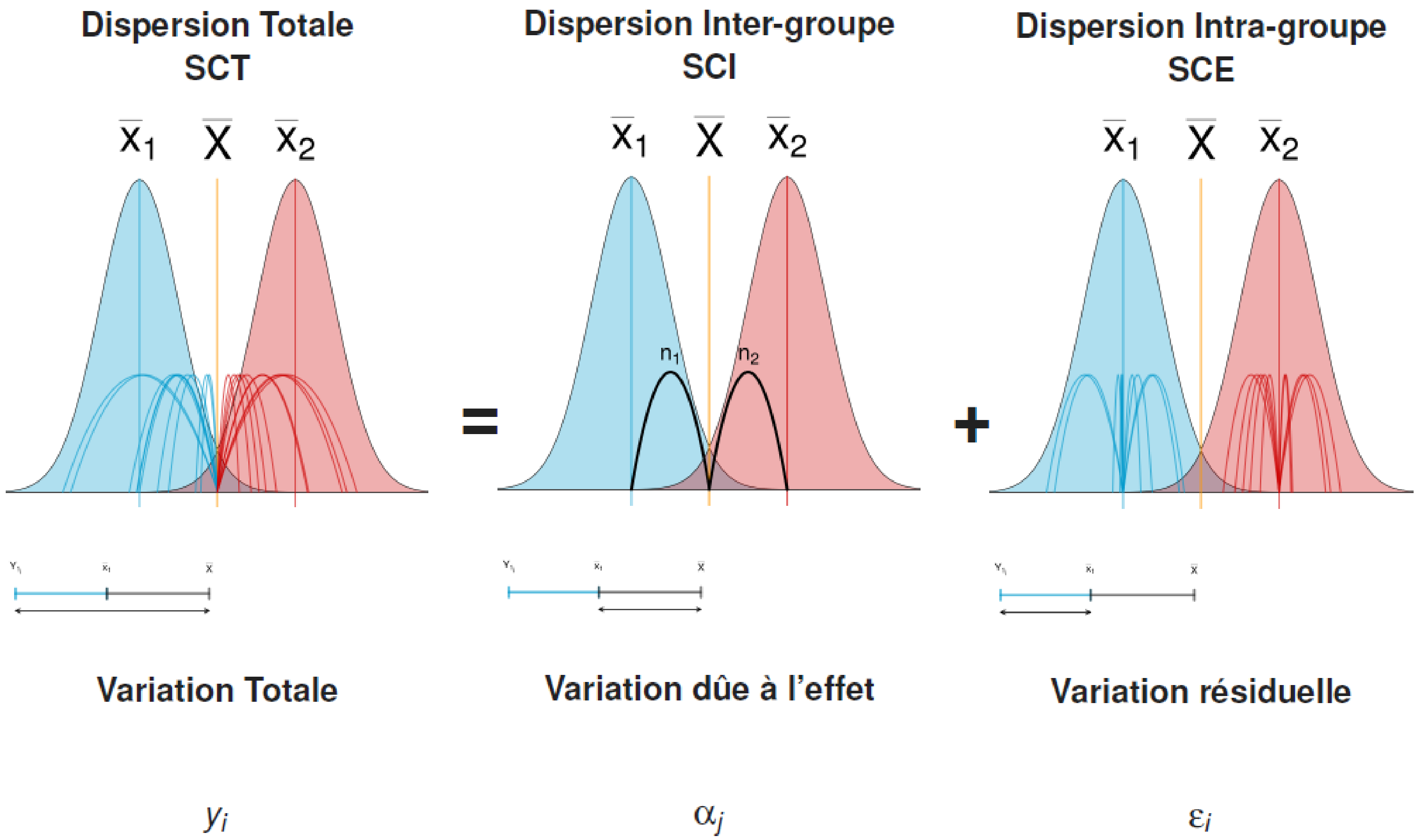
6. corrélation et régression

Xavier Bouteiller

bouteiller.xavier@gmail.com

Data University
Institut de science des données

MPDATA


Dispersion Totale
SCT
Dispersion Inter-groupe
SCI
Dispersion Intra-groupe
SCE
$\bar{x}_1$ $\bar{X}$ $\bar{x}_2$
n_1
n_2
=
+
Variation Totale
Variation dûe à l'effet
Variation résiduelle
y_i
α_j
ε_i

La corrélation et la régression sont deux notions voisines et inter-reliées.

- La corrélation mesure le degré de liaison entre les variables. Elle renseigne sur la force du lien entre les variables
- La régression offre une technique pour prédire la valeur d'une variable à partir de la valeur d'une autre
- la régression permet de mettre en évidence des relations fonctionnelles

Corrélation : Quel est le degré de relation ou de dépendance entre

- la pression artérielle et le taux de cholestérol chez des sujets
- la concentration en Sélénium des plans d'eau et la diversité du microbenthos
- la quantité de réserve lipidique des mammifères hibernant et la durée de l'hibernation

Régression : Après avoir déterminé l'équation de la droite de régression :

- quel est le nombre d'oeufs produit en fonction de l'âge des brochets
- est-ce que l'abondance des proies détermine l'abondance des prédateurs

Le coefficient de corrélation est égal à:

$$\rho_{x,y} = \frac{\sigma_{x,y}}{\sigma_x . \sigma_y}$$

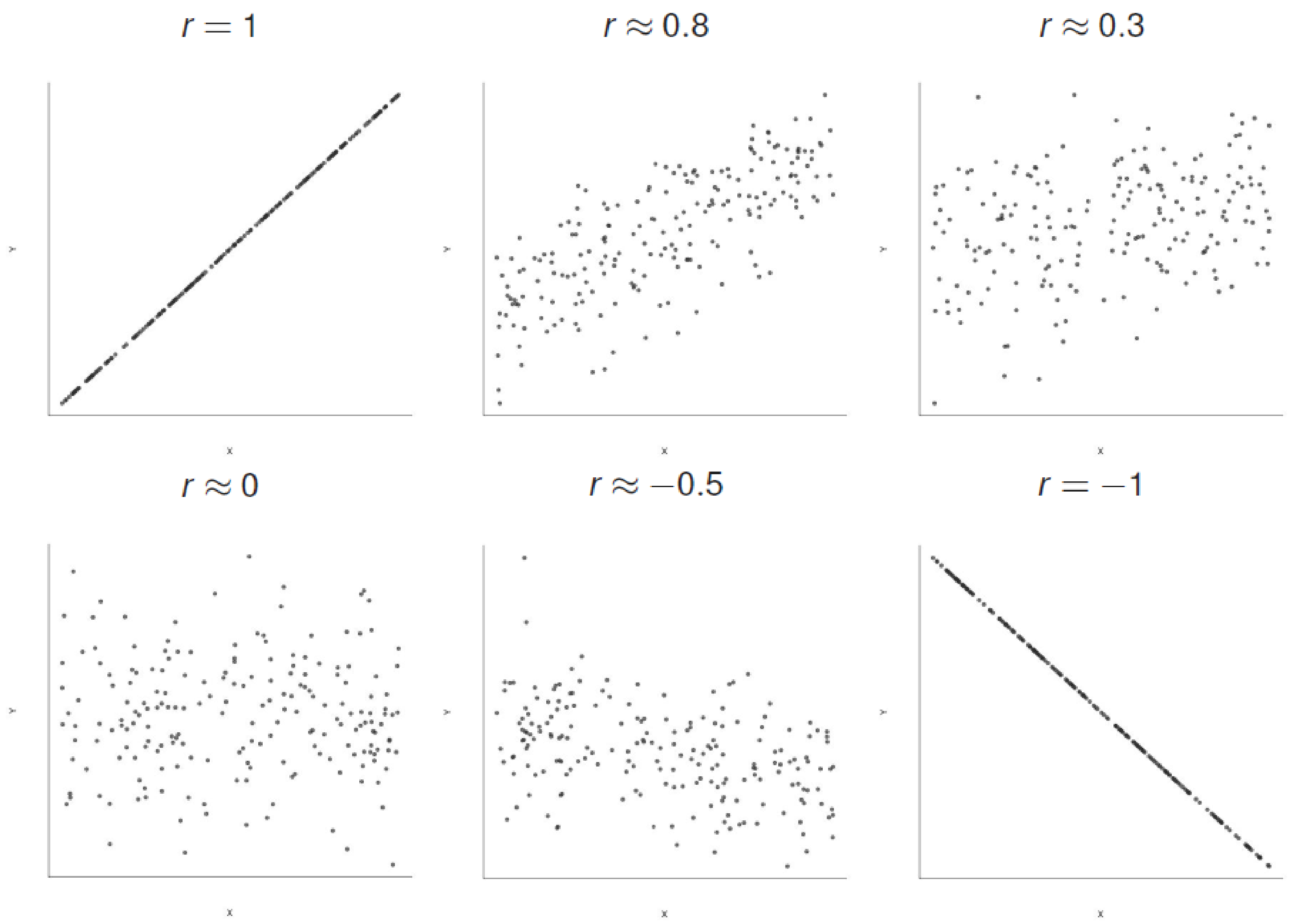
avec $\sigma_{x,y}$ correspondant à la covariance entre x et y.

Son estimateur (à partir d'échantillon) :

$$r_{x,y} = \frac{s_{x,y}}{s_x . s_y}$$

avec :

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


r = 1
r ≈ 0.8
r ≈ 0.3
r ≈ 0
r ≈ −0.5
r = −1

A la condition que :

- x et y sont deux variables quantitatives continues
- la distribution jointe de x et y est binormale

Si $H_0 : \rho = 0$

Alors

$$t_r = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Suit une loi de Student à $n-2$ ddl.

Le test peut être bi-latéral ou uni-latéral :
$H_1 : \rho \neq 0$ (bi-latéral)
$H_1 : \rho > 0$ (uni-latéral)
$H_1 : \rho < 0$ (uni-latéral)

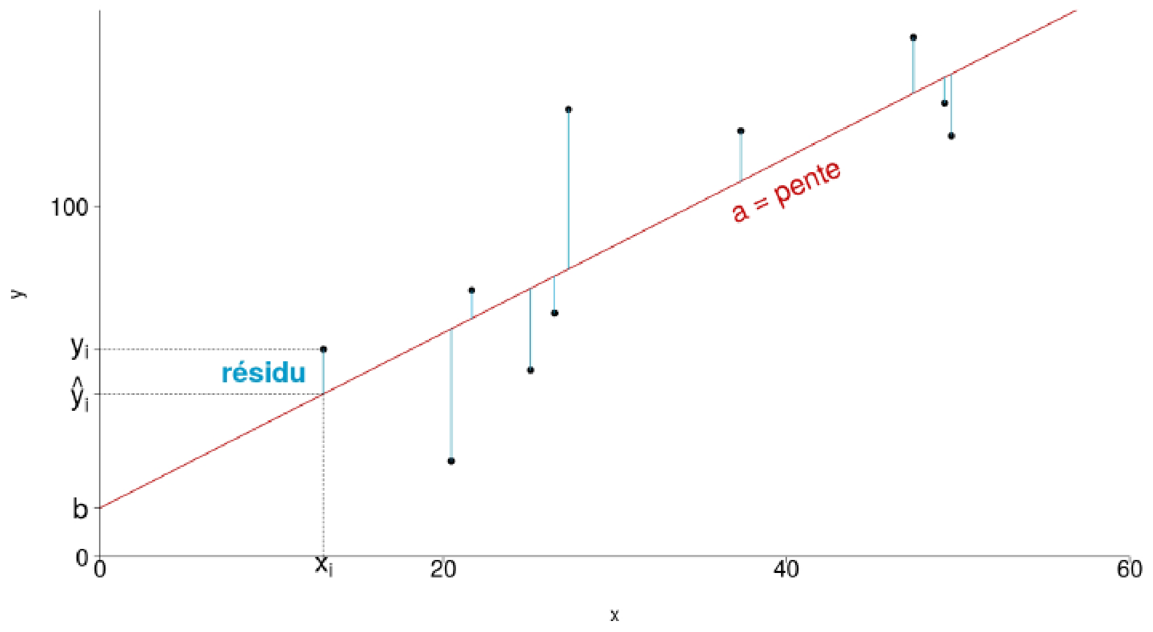
La régression linéaire simple consiste à calculer une fonction du premier degré liant les variables x et y.

La fonction linéaire est de la forme :

$$y = ax + b$$

$\rightarrow$ ligne droite qui traverse au mieux le nuage de points.

Notation stat

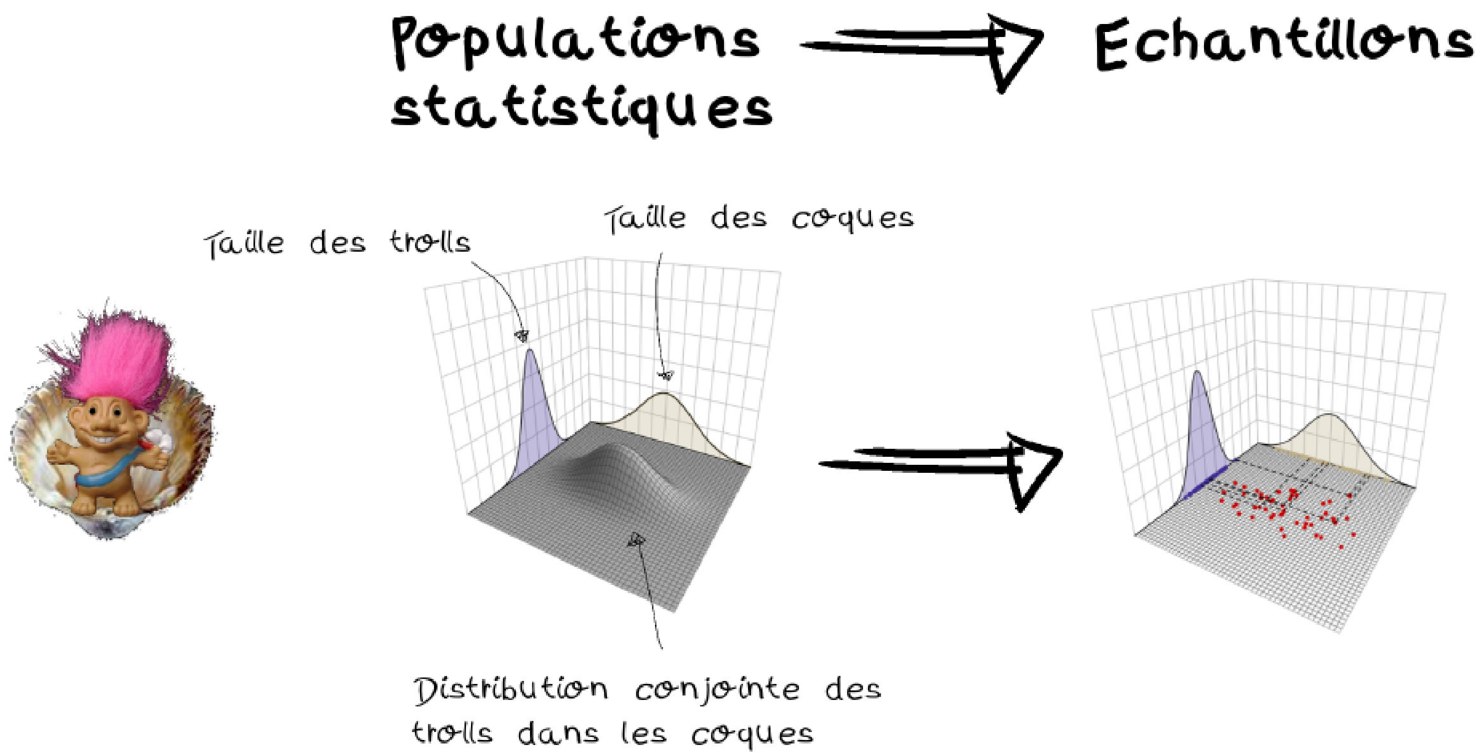


L'équation de la droite s'obtient en minimiant le carré des résidus, c.a.d pour les valeurs de a et b suivante :

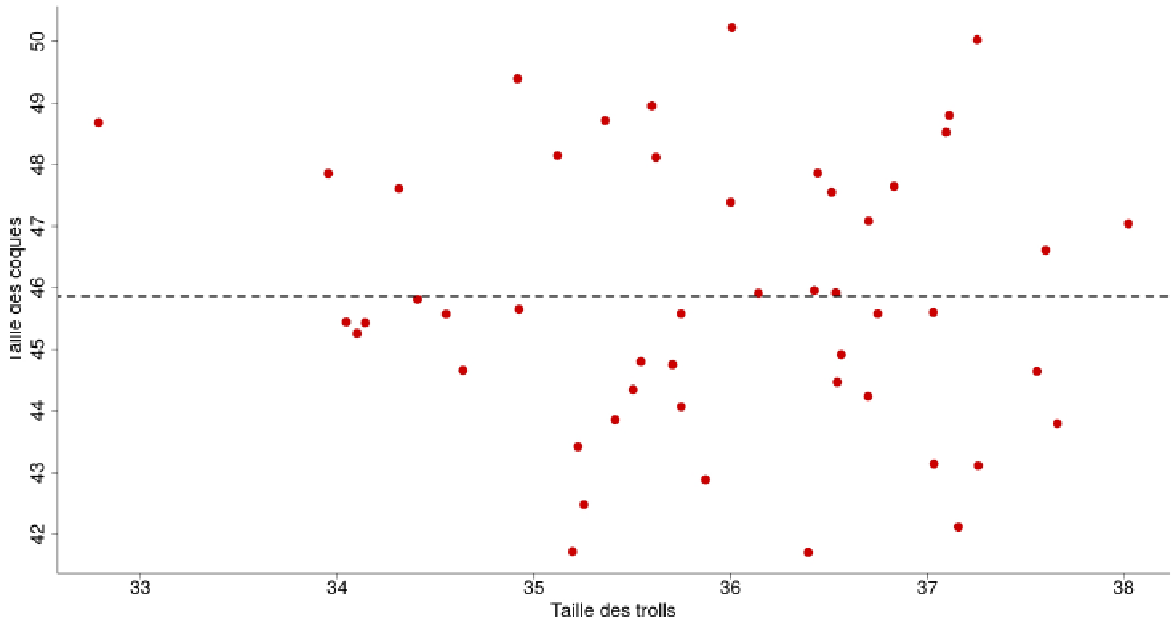
$$a = \frac{s_{xy}}{s_x^2}$$

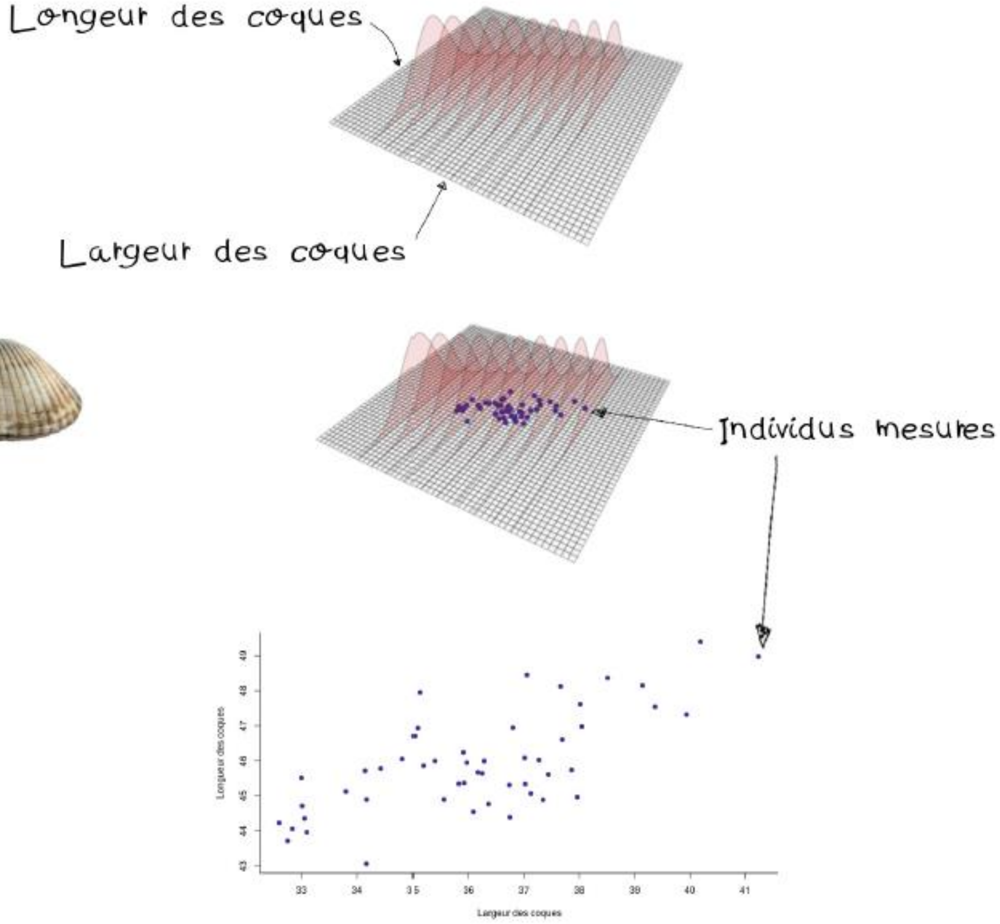
$$b = \bar{y} - a\bar{x}$$

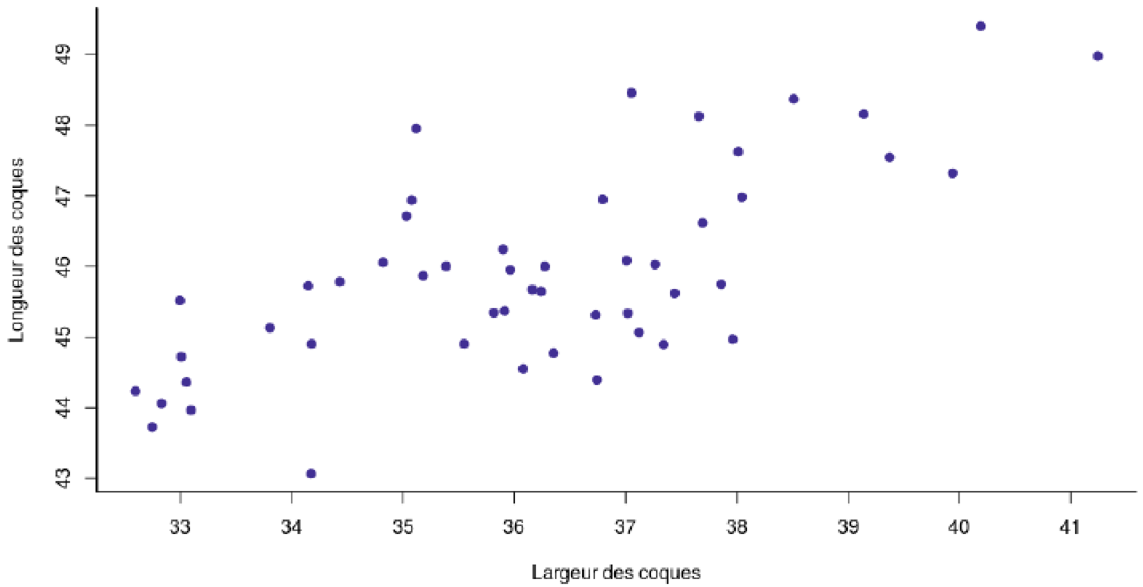
Que se passe-t-il lorsque x et y sont deux variables indépendantes ?

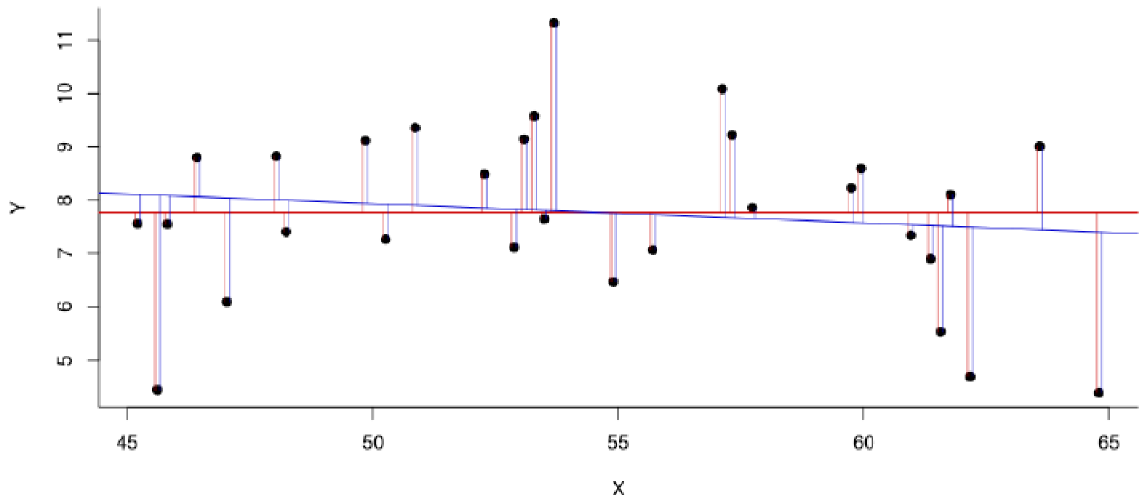


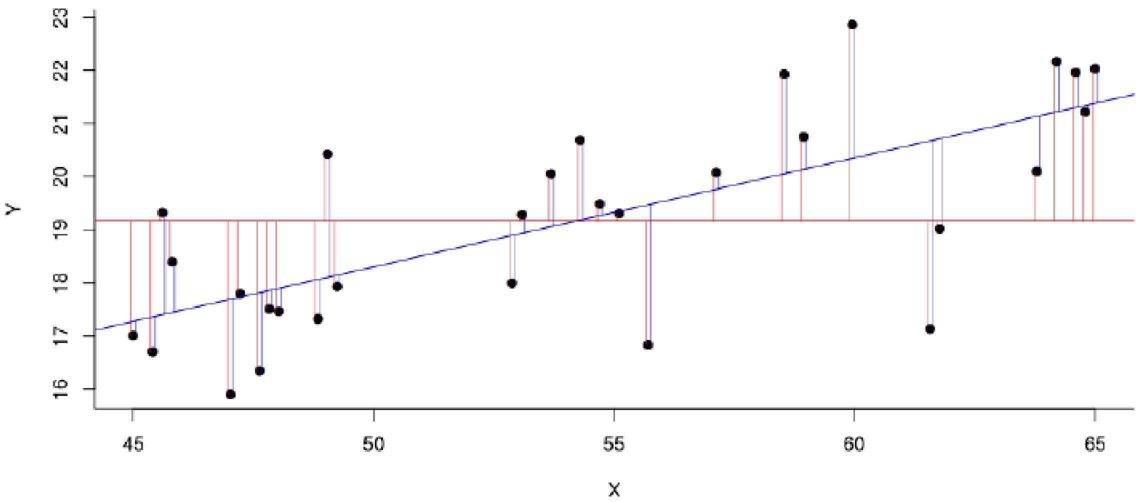
Le meilleur prédicteur est la droite de pente nulle et d'ordonnée à l'origine égale à la moyenne de la distribution



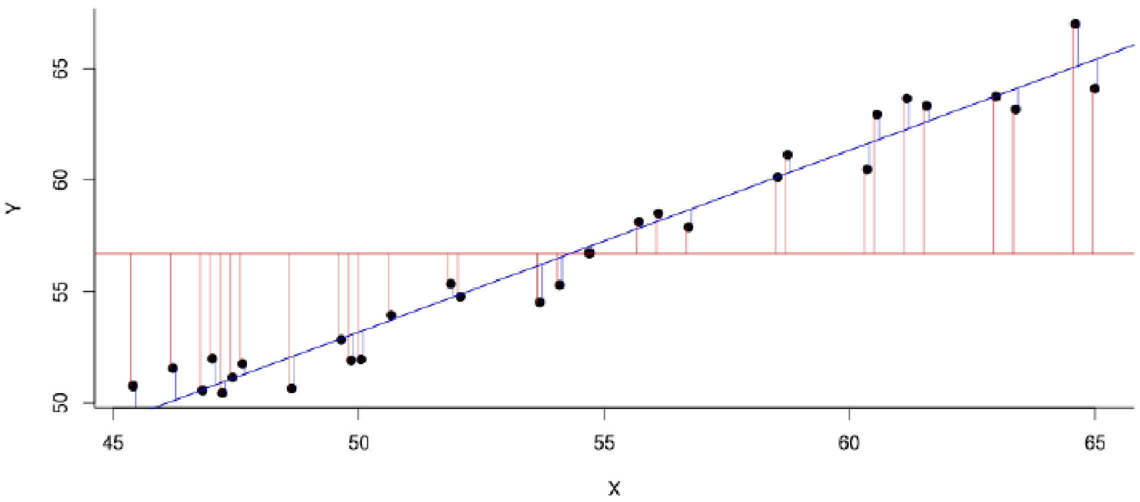

Longeur des coques
Largeur des coques
Individus mesures
Longueur des coques
Largeur des coques


Longueur des coques
43
44
45
46
47
48
49
33
34
35
36
37
38
39
40
41
Largeur des coques


Y
X
45
50
55
60
65
5
6
7
8
9
10
11



Y
X
16
17
18
19
20
21
22
23
45
50
55
60
65



Y
65
60
55
50
45
50
55
60
65
X

Un modèle de régression linéaire se paramétrise comme suit:

$$y_i = \alpha + \beta \times x_i + \varepsilon_i$$

avec

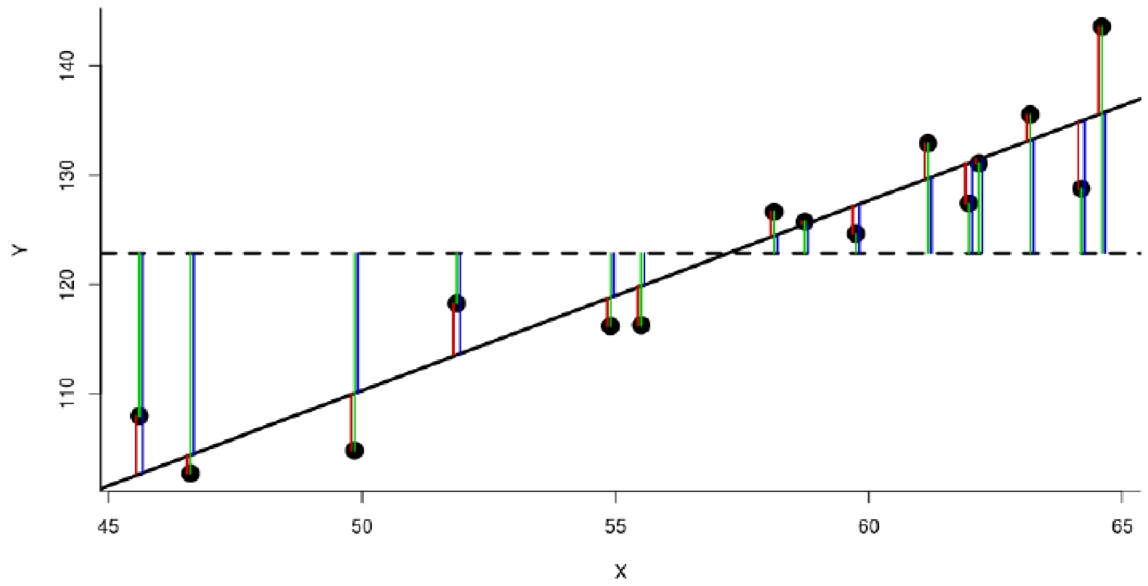
$$\varepsilon_i \sim Normal(0, \sigma^2)$$

La différence avec l'analyse de variance précédente est que la variable x ne prends pas juste 2 ou plus valeur pour indiquer l'appartenance à un groupe mais peut prendre toutes les valeurs possibles entre 2 limites.
La représentation géométrique de cette relation est :

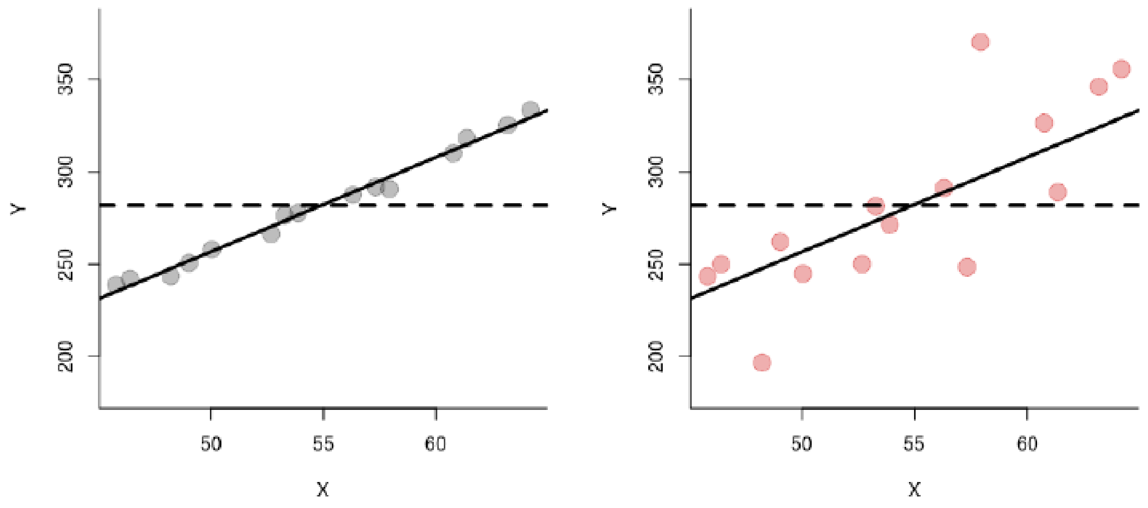
- une ordonnée à l'origine α
- une pente β

Analyse de variance de la régression

Variance Totale	=	Variance Expliquée	+	Variance Résiduelle
$\sum_{i=1}^{n}(y_i - \bar{y})^2$	=	$\sum_{i=1}^{n}(\hat{y} - \bar{y})^2$	+	$\sum_{i=1}^{n}(\hat{y} - y_i)^2$
SCET		SCER		SCEE
$n-1$ ddl		1 ddl		$n-2$ ddl



2 jeux de données avec les mêmes paramètres de pente et d'ordonnée à l'origine peuvent être très différents !



R^2 exprime le ratio entre la part de variation de y expliquée sur la part de variation totale de y (cf. page suivante). Il permet d'estimer la qualité de l'ajustement de la droite de régression aux données.

$$R^2 = \frac{SCER}{SCET}$$

Si tous les points sont alignés, les erreurs d'estimations (les résidus e_i) sont nulles et SCEE = 0.
Donc SCET = SCER et $R^2 = 1$.

Si les deux variables sont indépendantes, la pente a est nulle et SCER = 0.
Donc $R^2 = 0$.

Deux approches pour tester la signification de la régression de y en x :

- Test à partir de la distribution d'échantillonnage de pente *a*
- Principe de l'analyse de la variance

Les deux approches sont rigoureusement équivalentes

Sous H_0 : $\alpha = 0$,la vraie pente de la régression est nulle

La variable de décision :

$$t_{ac} = \frac{a}{\sqrt{\frac{s_e^2}{(n-1)s_x^2}}}$$

Suit une Loi de Student à $n-2$ degrés de liberté.

s_e^2 est l'estimation de la σ^2 et

$$s_e^2 = \frac{\sum(y_i - \hat{y}_i)^2}{n-2} = \frac{SCEE}{n-2}$$

Le test est bilatéral, H_1 : $\alpha \neq 0$

$H_0 : \rho^2 = 0$, la vraie proportion de variance expliquée par la régression est nulle

La variable de décision :

$$F_{calc} = \frac{SCER/1}{SCEE/(n-2)}$$

suit une loi de Fischer $F_{(1-\alpha;\upsilon_1=1;\upsilon_2=n-2)}$

Le test est unilatéral, $H_1 : \rho^2 > 0$

Let’s practice …