

Statistique bivariée

Statistique bivariée

*Lorsqu'on effectue sur un échantillon simultanément, deux mesures, on obtient une série double de mesures ; par exemple; on parlera de la série double de mesures sur la taille et le poids des étudiants .. Ces mesures sont en fait des réalisations de deux variables quantitatives et la question qui se pose est celle de l'existence d'une **liaison entre ces deux variables**.*

Statistique bivariée

L'ajustement linéaire consiste à tracer une *droite* qui passe

au plus près des observations d'un *nuage de points*.

Cette droite est ensuite utilisée pour faire des *prévisions*.

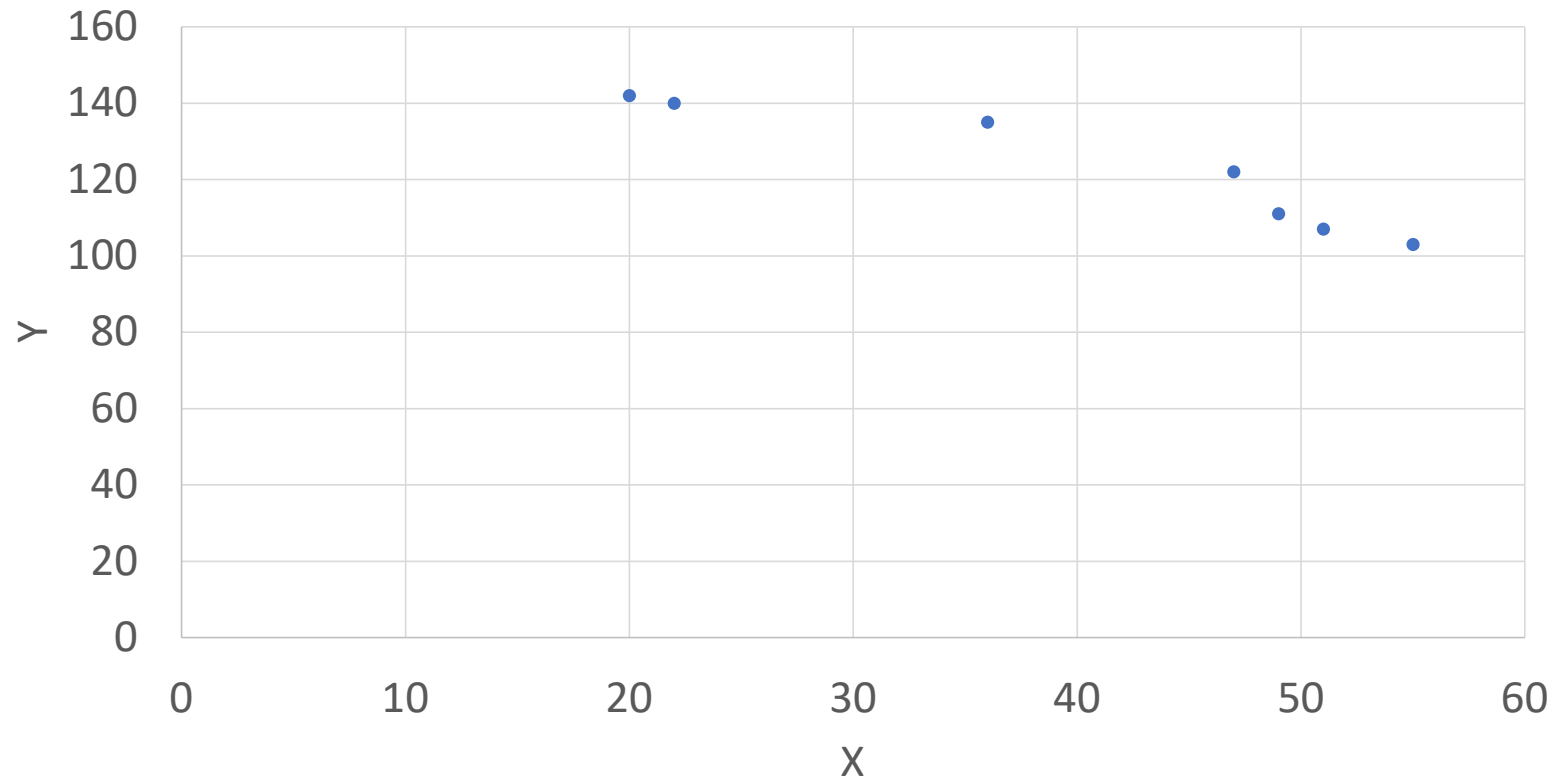
Exemple 1 introdutif

X	55	51	49	47	36	22	20
Y	103	107	111	122	135	140	142

Nuage de points Figure 2.1

Des points dispersés entre les deux axes.

l'axe des abscisse et l'axe des ordonnées



Interprétation de nuage des points

*Les deux variable se varient dans deux sens opposés,
quand X augmente Y déminue*

Qu'est-ce la régression linéaire ?

Lorsque le nuage de points associé à une série statistique double a une forme "allongée " c'est-à-dire lorsque les points sont sensiblement alignés, on peut tracer des droites passant « au plus près de ces points ».

On dit alors que chacune de ces droites réalise un **ajustement affine** du nuage de points.

X	Y
Variable explicative	Variable expliquée
Variable contrôlée	Variable réponse
Variable indépendante	Variable dépendante

On distingue trois cas

Cas 1: Tableau de deux lignes

$X=x_i$	x_1	x_2	x_{k-1}	x_N
$Y=y_i$	y_1	y_2		y_{k-1}	y_N

On distingue trois cas

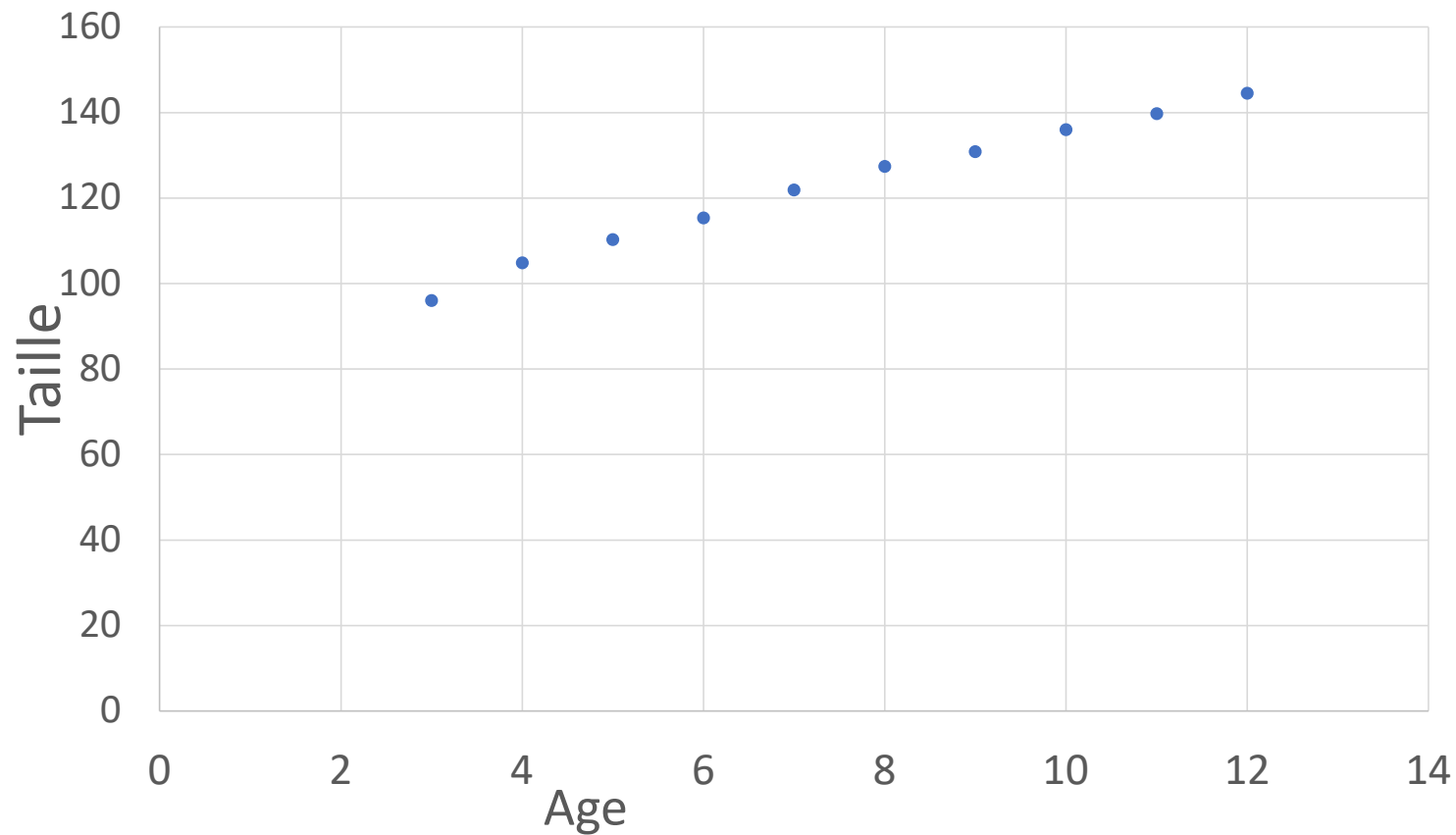
Cas 1: Tableau de deux lignes

Exemple 2:

<i>Age</i>	<i>3</i>	<i>4</i>	<i>5</i>	<i>6</i>	<i>7</i>	<i>8</i>	<i>9</i>	<i>10</i>	<i>11</i>	<i>12</i>
<i>taille</i>	<i>96</i>	<i>104,8</i>	<i>110,3</i>	<i>115,3</i>	<i>121,9</i>	<i>127,4</i>	<i>130</i>	<i>136,9</i>	<i>139,7</i>	<i>144,5</i>

Interprétation de nuage des points

*Les deux variable évoluent dans le même sens
quand X augmente Y augmente*



Ajustement graphique

Vous choisissez deux couples (5;110,3) et (10;136,9)

$$\begin{cases} 110,3 = a5 + b \dots\dots (1) \\ 136,9 = a10 + b \dots\dots (2) \end{cases}$$

$$(2)-(1) \quad 5a=26,6 \quad a = \frac{26,6}{5} = 5,32$$

$$\text{De (2) } b = 136,9 - 10(5,32) = 83,7$$

*Alors la droite **d'ajustement graphique** est de*

$$y = 5,32x + 83,7$$

Ajustement graphique

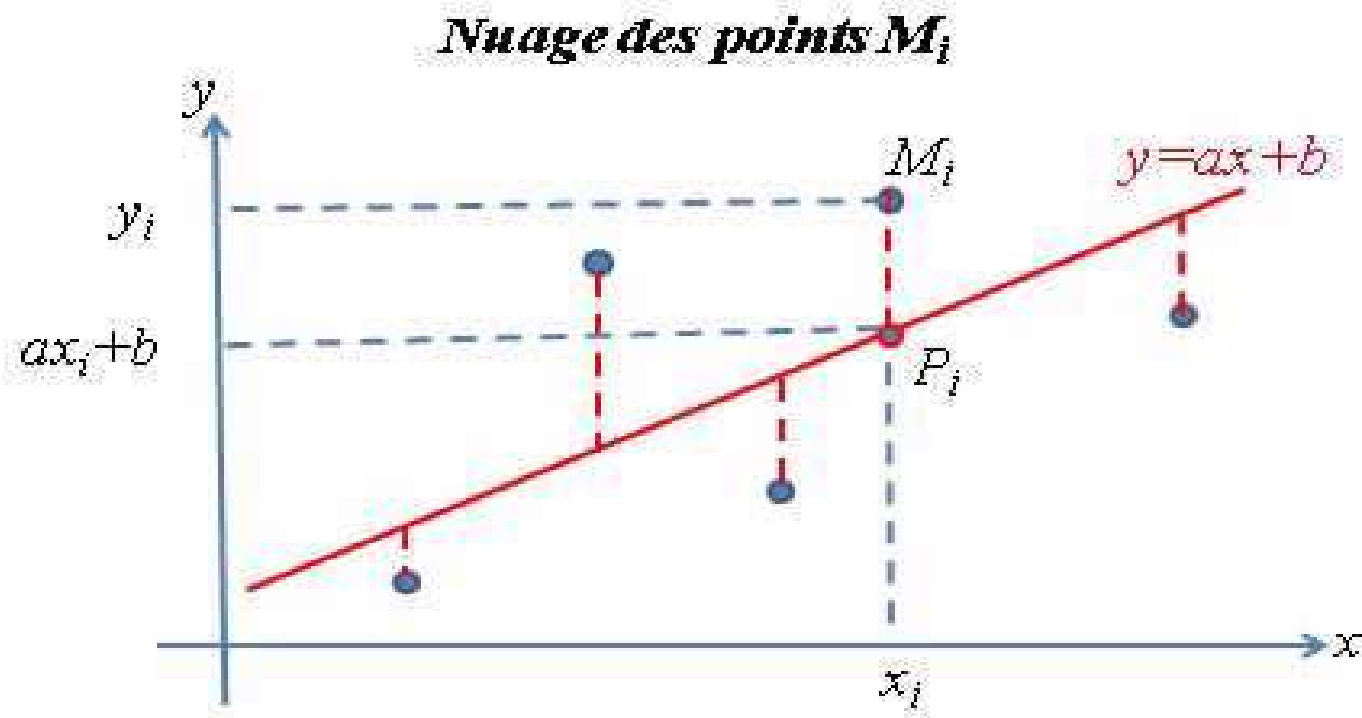
Les inconvénients de l'ajustement graphique, c'est que ce dernier est une méthode subjective , chaque étudiant trouve une droite et finalement vous n'avez pas des critères objectifs pour choisir entre ces droites .

Problématique ?

Comment faire pour avoir une droite unique?

Droite de La méthode des moindres carrés (MOC)

Cette méthode consiste à minimiser la somme des carrés des écarts



Démonstration

Minimiser les écarts, soit la fonction

$$F(a, b) = \sum_{i=1}^N (y_i - (ax_i + b))^2$$

Après une chaîne d'opérations mathématiques, on obtient:

Méthode des moindres carrées

$$\begin{cases} a = \frac{\sum_{i=1}^N x_i y_i - \bar{Y} \sum_{i=1}^N x_i}{\sum_{i=1}^N x_i^2 - \bar{X} \sum_{i=1}^N x_i} \\ b = \bar{Y} - a\bar{X} \end{cases}$$

➤ La 2^{ème} équation signifie que la droite de régression passe la

point moyen $G(\bar{X}; \bar{Y})$

Droite de la méthode des moindres carrés

Droite de Y en fonction de X: Notée par

$$***D_Y(x): y = ax + b***$$

Tels que:

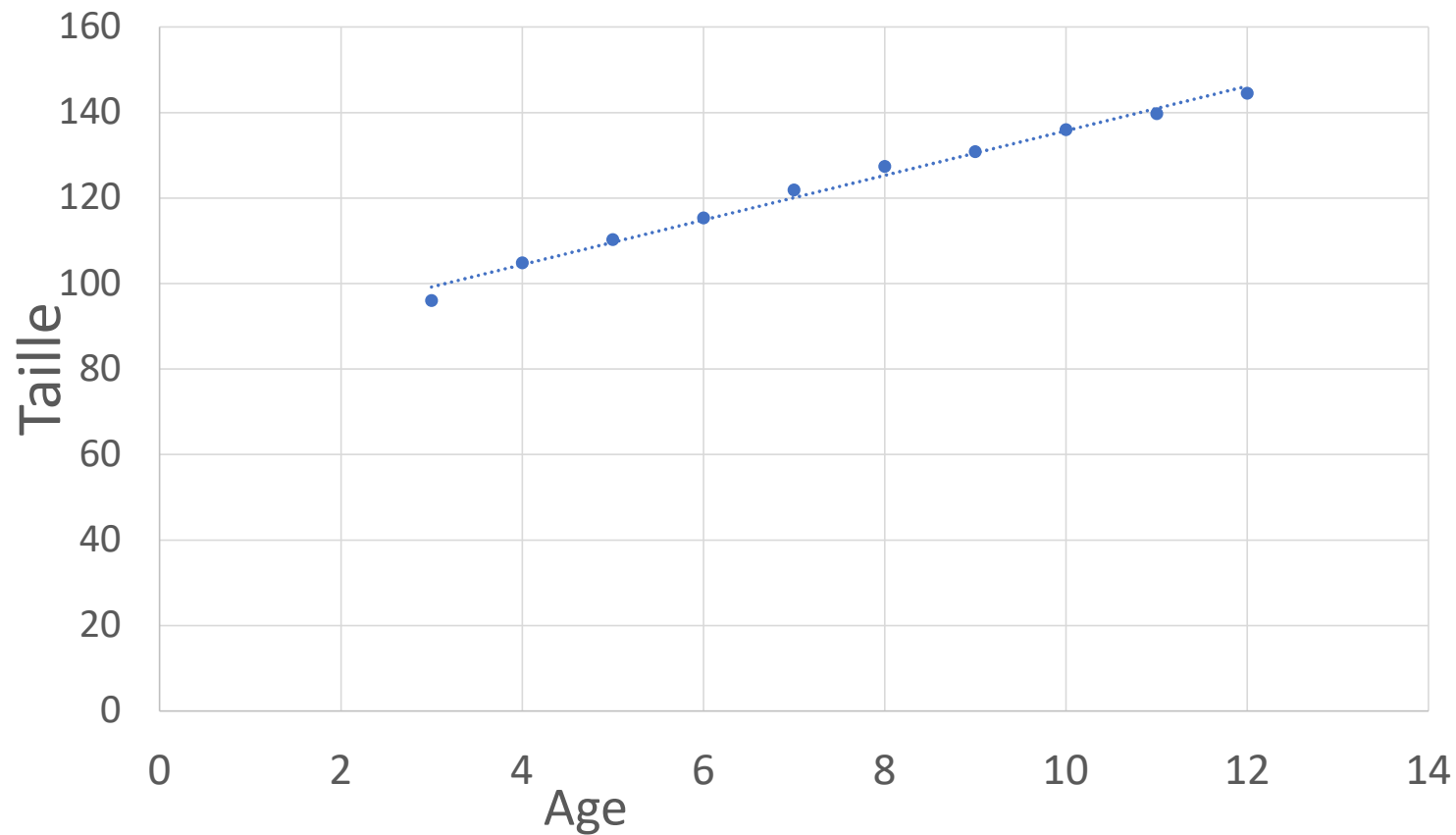
$$\left\{ \begin{array}{l} a = \frac{\sum_{i=1}^N x_i y_i - \bar{Y} \sum_{i=1}^N x_i}{\sum_{i=1}^N x_i^2 - \bar{X} \sum_{i=1}^N x_i} \\ b = \bar{Y} - a\bar{X} \end{array} \right.$$

Propriétés de la droite de régression $D_Y(x)$

- 1. C'est une droite **unique***
- 2. Elle passe toujours par le points moyen $G(\bar{X}; \bar{Y})$*

Interprétation de nuage des points

*Les deux variable se varient dans le même sens
quand X augmente Y augmente*



Notations:

1) *La covariance* $Cov(X; Y) = \sigma_{XY} = \frac{1}{N} \sum_{i=1}^N x_i y_i - \bar{X} \bar{Y}$

2) *La variance de X* $\sigma_X^2 = \frac{1}{N} \sum_{i=1}^N x_i^2 - \bar{X}^2$

3) *La variance de Y* $\sigma_Y^2 = \frac{1}{N} \sum_{i=1}^N y_i^2 - \bar{Y}^2$

Alors $a = \frac{\sigma_{XY}}{\sigma_X^2}$

Coefficient de corrélation linéaire

$$\text{Symbolisé par } r = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

Le coefficient de corrélation est une valeur sans unité comprise toujours entre (-1) et (+1)

$$-1 \leq r \leq +1$$

Coefficient de corrélation linéaire

- *Les valeurs positives de r indiquent une corrélation positive*

Lorsque les valeurs des deux variables tendent augmenter ensemble

- *Les valeurs négatives de r indiquent une corrélation négative lorsque les valeurs d'une variable tend à augmenter et que les valeurs de l'autre variable diminuent*

Signification de r

- $r = 0$ liaison nulle
- $r = +1$ forte liaison positive
- $r = -1$ forte liaison négative

On distingue trois cas

Cas 1: Tableau de deux lignes

Exemple 2:

<i>Age</i>	<i>3</i>	<i>4</i>	<i>5</i>	<i>6</i>	<i>7</i>	<i>8</i>	<i>9</i>	<i>10</i>	<i>11</i>	<i>12</i>
<i>taille</i>	<i>96</i>	<i>104,8</i>	<i>110,3</i>	<i>115,3</i>	<i>121,9</i>	<i>127,4</i>	<i>130</i>	<i>136,9</i>	<i>139,7</i>	<i>144,5</i>

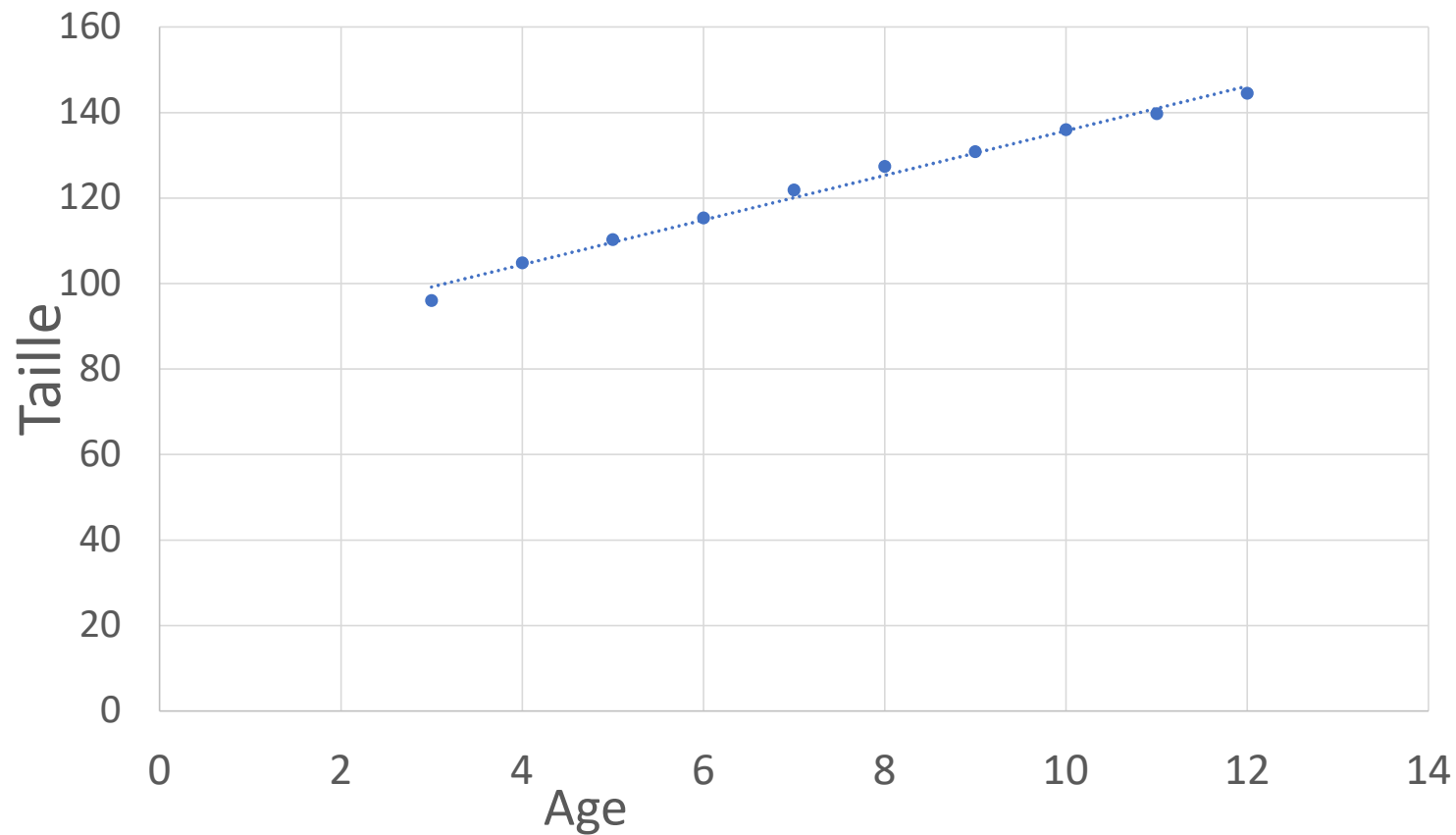
Droite de la méthode des moindres carrés

Droite de X en fonction de Y

$$***D_Y(x): y = ax + b***$$

Interprétation de nuage des points

*Les deux variable se varient dans le même sens
quand X augmente Y augmente*



$$**D_Y(x): y = ax + b**$$

$$B \rightarrow a = 5,23$$

$$A \rightarrow b = 83,43$$

$$r = 0,9945$$

$$**D_Y(x): y = 5,23x + 83,43**$$

2^{ème} droite de la méthode des moindres carrés

Droite de X en fonction de Y: Notée par

$$D_X(Y): x = \hat{a}y + \hat{b}$$

Tels que:

$$\left\{ \begin{array}{l} \hat{a} = \frac{\sum_{i=1}^N x_i y_i - \bar{Y} \sum_{i=1}^N x_i}{\sum_{i=1}^N y_i^2 - \bar{Y} \sum_{i=1}^N y_i} \\ \hat{b} = \bar{X} - \hat{a} \bar{Y} \end{array} \right.$$

On reprend l'exemple précédant:

Droite de X en fonction de Y

$$D_X(Y): x = \hat{a}y + \hat{b}$$

$$B \rightarrow \hat{a} = 0,189$$

$$A \rightarrow \hat{b} = -15,69$$

$$r = 0,9945$$

$$D_X(Y): x = 0,189y - 15,69$$

Remarques

1. a , \hat{a} et r ayant le même signe

Les trois sont positives ou les trois sont négatives

2. $r^2 = a\hat{a}$

Cas 2: Tableau de trois lignes

$X=x_i$	x_1	x_2	x_{k-1}	x_k
$Y=y_i$	y_1	y_2	y_{k-1}	y_k
n_i	n_1	n_2	n_{k-1}	n_k

En adaptant les formules pour le cas 2

1) La covariance $Cov(X; Y) = \sigma_{XY} = \frac{1}{N} \sum_{i=1}^k n_i x_i y_i - \bar{X} \bar{Y}$

2) La variance de X $\sigma_X^2 = \frac{1}{N} \sum_{i=1}^k n_i x_i^2 - \bar{X}^2$

3) La variance de Y $\sigma_Y^2 = \frac{1}{N} \sum_{i=1}^k n_i y_i^2 - \bar{Y}^2$

Alors $a = \frac{\sigma_{XY}}{\sigma_X^2}$

Cas 2: Tableau de trois lignes

Exemple

$X=x_i$	1	2	3	4	5
$Y=y_i$	10	15	17	3	18
n_i	1	3	5	7	4

$$D_Y(x): y = ax + b$$

$$B \rightarrow a = -0,66$$

$$A \rightarrow b = 13,96$$

$$r = -0,1122$$

(Corrélation faible négative)

$$D_Y(x): y = -0,66 + 13,96$$

Cas 3

Tableau de contingence

$X=x_i$ $Y=y_i$	y_1	\dots	y_j	\dots	y_l	<i>Totaux</i>
x_1	n_{11}	\dots	n_{1j}	\dots	n_{1l}	$n_{1.}$
x_i	n_{i1}		n_{ij}		n_{il}	$n_{i.}$
x_k	n_{k1}		n_{kj}		n_{kl}	$n_{k.}$
<i>Totaux</i>	$n_{.1}$		$n_{.j}$		$n_{.l}$	$n_{..} = N$

En adaptant les formules pour le cas d'un tableau de contingence

1) La covariance $Cov(X; Y) \stackrel{!}{=} \sigma_{XY}$

$$= \frac{1}{N} \sum_{i=1}^k \sum_{j=1}^l n_{ij} x_i y_j - \bar{X} \bar{Y}$$

1) La variance de X $\sigma_X^2 = \frac{1}{N} \sum_{i=1}^k n_i x_i^2 - \bar{X}^2$

2) La variance de Y $\sigma_Y^2 = \frac{1}{N} \sum_{j=1}^l n_j y_j^2 - \bar{Y}^2$

Alors $a = \frac{\sigma_{XY}}{\sigma_X^2}$

$Y=y_i$ $X=x_i$	1	4	5	Totaux
2	1	10	3	14
3	2	4	1	7
7	3	1	4	8
8	7	2	2	11
<i>Totaux</i>	13	17	10	40

Exemple: La saisie dans la calculatrice

$k = 4$ lignes et $l = 3$ colonnes, donc $k \times l = 4 \times 3 = 12$ valeurs

X	Y	ni
2	1	1
2	4	10
2	5	3
3	1	2
3	4	4
3	5	1
7	1	3
7	4	1
7	5	4
8	1	7
8	4	2
8	5	2

$$D_Y(x): y = ax + b$$

$$B \rightarrow a = -0,04539$$

$$A \rightarrow b = 3,3816$$

$$r = -0,77$$

(Corrélation moyenne négative)

$$D_Y(x): y = -0,04539x + 3,3816$$

Distribution marginale de X

$X=x_i$	<i>Totaux</i>
x_1	$n_{1.}$
x_i	$n_{i.}$
x_k	$n_{k.}$
<i>Totaux</i>	$n_{..} = N$

Distribution marginale de Y

$Y=y_i$	y_1	\dots	y_j	\dots	y_l	<i>Totaux</i>
<i>Totaux</i>	$n_{.1}$		$n_{.j}$		$n_{.l}$	$n_{..} = N$

Régression non linéaire

La régression non linéaire est une méthode permettant de déterminer un modèle non linéaire de relation entre deux variables

Régression non linéaire

Formes non linéaires:

1) Exponentielle

2) Logarithmique

3) Puissance

4) Inverse

Régression exponentielle

$$**Y = Be^{aX}**$$

Régression exponentielle

Forme non linéaire :

$$Y = Be^{aX}$$

Forme linéaire (transformation logarithmique) :

$$\mathit{lin}Y = \mathit{lin}Be^{aX} = \mathit{lin}B + \mathit{lin}e^{aX} = \mathit{lin}B + aX$$

1^{ère} méthode:

Linéarisation de la forme exponentielle

Avec:

$$(*) \begin{cases} y = \mathit{lin} Y \\ b = \mathit{lin} B \end{cases}$$

On obtient:

$$y = aX + b$$

C'est-à-dire

Création d'une relation linéaire entre (X) et $(\mathit{lin} Y)$

Ensuite, vous choisissez une régression linéaire (par la méthode des moindres carrés)

Les deux variables sont X et $lin Y=y$

Exemple:

X	1	2	13	19	25	31
Y	1240	1190	1160	1135	1094	1065

$$Y = Be^{aX}$$

*Trouver les valeurs des constantes
a et B*

$$Y = Be^{aX}$$

Création d'une relation linéaire entre (X) et (lin Y)

Les deux variables sont X et $y = \text{lin } Y$

X	1	2	13	19	25	31
$y = \text{lin } Y$	7,123	7,082	7,056	7,034	7,085	6,971

$$**y = aX + b**$$

$$B \rightarrow a = -4,445552 \times 10^{-3}$$

$$A \rightarrow b = 7,111334964$$

$$r = -0,974911601$$

$$**y = -4,445552 \times 10^{-3} X + 7,111334964**$$

Et voila la relation linéaire

Avec (*), on obtient:

$$y = -4,445552 \times 10^{-3} X + 7,111334964$$

En introduisant l'exponentiel des deux cotés de

l'équation, on obtient:

$$\begin{aligned} e^y &= e^{-4,445552 \times 10^{-3} X + 7,111334964} \\ &= e^{-4,445552 \times 10^{-3} X} \times e^{7,111334964} \end{aligned}$$

Alors:

$$Y = Be^{aX} = 1225,78e^{-4,45 \times 10^{-3} X}$$

2^{ème} méthode:

*Directement vous choisissez le modèle:
régression exponentielle*

$$Y = Be^{aX}$$

Avec la calculatrice :

$$Y = 1225,78238e^{-4,445555 \times 10^{-3} X}$$

Remarque:

*La valeur donnée par la calculatrice du coefficient de
corrélacion r est la valeur de la corrélacion linéaire*

C'est-à-dire:

$$*r = cor(X, lin Y) = -0,9749*$$

Régression Puissance

$$Y = BX^a$$

*C'est-à-dire X dans la **Base***

Linéarisation de la forme puissance

Forme non linéaire :

$$Y = BX^a$$

Forme linéaire (transformation logarithmique) :

$$\mathit{lin}(y) = \mathit{lin}(BX^a)$$

$$\mathit{lin}(Y) = \mathit{lin}(B) + a\mathit{lin}(X)$$

1^{ère} méthode:

Linéarisation da le forme puissance

Avec:

$$(**) \begin{cases} y = \text{lin}(Y) \\ x = \text{lin}(X) \\ b = \text{lin}(B) \end{cases}$$

On obtient:

$$y = ax + b$$

C'est-à-dire

Création d'une relation linaire entre $\text{lin}(X)$ et $\text{lin}(Y)$

Ensuite, vous choisissez une régression linéaire (par la méthode des moindres carrés)

Les deux variables sont $\ln(X)$ et $\ln(Y)$

Exemple:

X	100	200	130	190	250	310
Y	1240	1190	1160	1135	1094	1065

Les deux variables sont X et $y = \ln Y$

<i>$\ln(X)$</i> <i>$= x$</i>	2,30	5,289	4,868	5,247	5,521	5,737
<i>y</i> <i>$= \ln(Y)$</i>	7,123	7,082	7,056	7,034	7,085	6,971

$$y = aX + b$$

$$A \rightarrow b = 7,65708$$

$$B \rightarrow a = -0,11763$$

$$r = -0,882249$$

$$y = -0,11763x + 7,65708$$

Et voila la relation linéaire

Avec (**), on obtient:

$$y = -0,11763x + 7,65708$$

En introduisant l'exponentiel des deux cotés de

l'équation, on obtient: $Y = BX^a$

$$e^y = e^{-0,11763x+7,65708} = e^{-0,11763x} \times e^{7,65708}$$

$$e^{\ln(Y)} = e^{-0,11763\ln(X)} \times e^{7,65708}$$
$$= e^{\ln(X)^{-0,11763}} \times e^{7,65708} =$$

Ajustement puissance

$$Y = e^{\ln(X)^{-0,11763}} \times e^{7,65708}$$

Finalemment, on a:

$$Y = 2115,570934 \times X^{-0,11736}$$

2^{ème} méthode:

*Directement vous choisissez le modèle:
régression puissance*

$$Y = BX^a$$

Avec la calculatrice :

$$Y = 2115,58 \times X^{-0,1176}$$

Remarque:

La valeur donnée par la calculatrice du coefficient de corrélation r est la valeur de la corrélation linéaire

C'est-à-dire:

$$r = \text{cor}(\text{lin}(X), \text{lin}(Y)) = -0,882249$$

Choix du modèle de régression valable

Choix entre la régression linéaire , exponentielle et puissance

1) Vous calculez les trois coefficient de corrélation

2) Vous choisissez le coefficient le plus élevé en *valeur absolue*

$r_1 = cor(X, Y)$, $r_2 = cor(X, lin(Y))$ et

$r_3 = cor(lin(X), lin(Y))$

3) Le modèle associé à ce coefficient est le plus valable

Choix entre la régression linéaire , exponentielle et puissance

Ou bien, directement vous choisissez le coefficient de détermination R^2 le plus élevé

Le modèle associé à ce coefficient de détermination

est le plus convenable

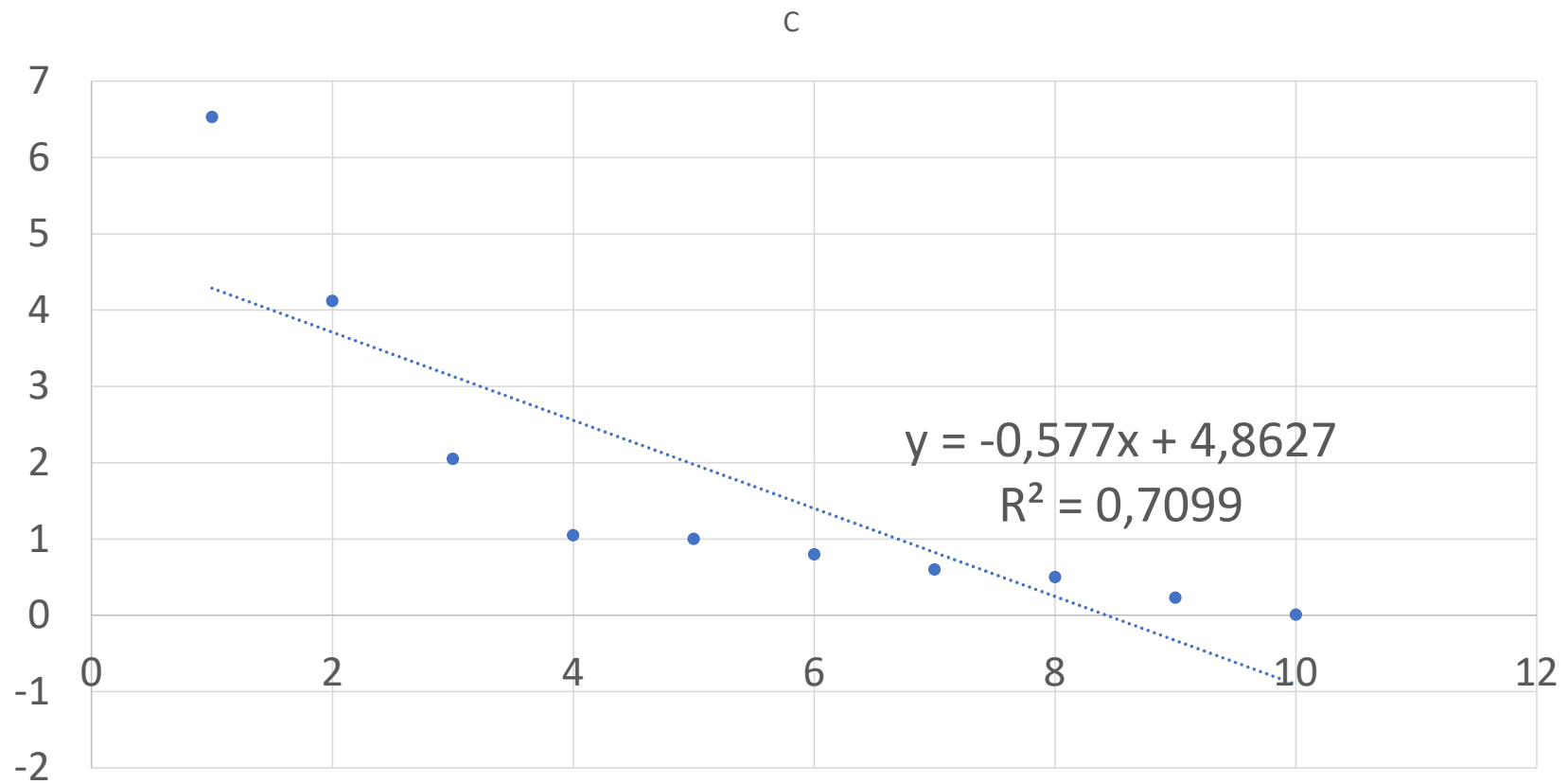
R^2 le carré du coefficient de corrélation

Exemple d'entraînement:

V mol	1	2	3	4	5	6	7	8	9	10
C (mol /l)	6,53	4,12	2,05	1,5	1	0,8	0,6	0,5	0,23	0,01

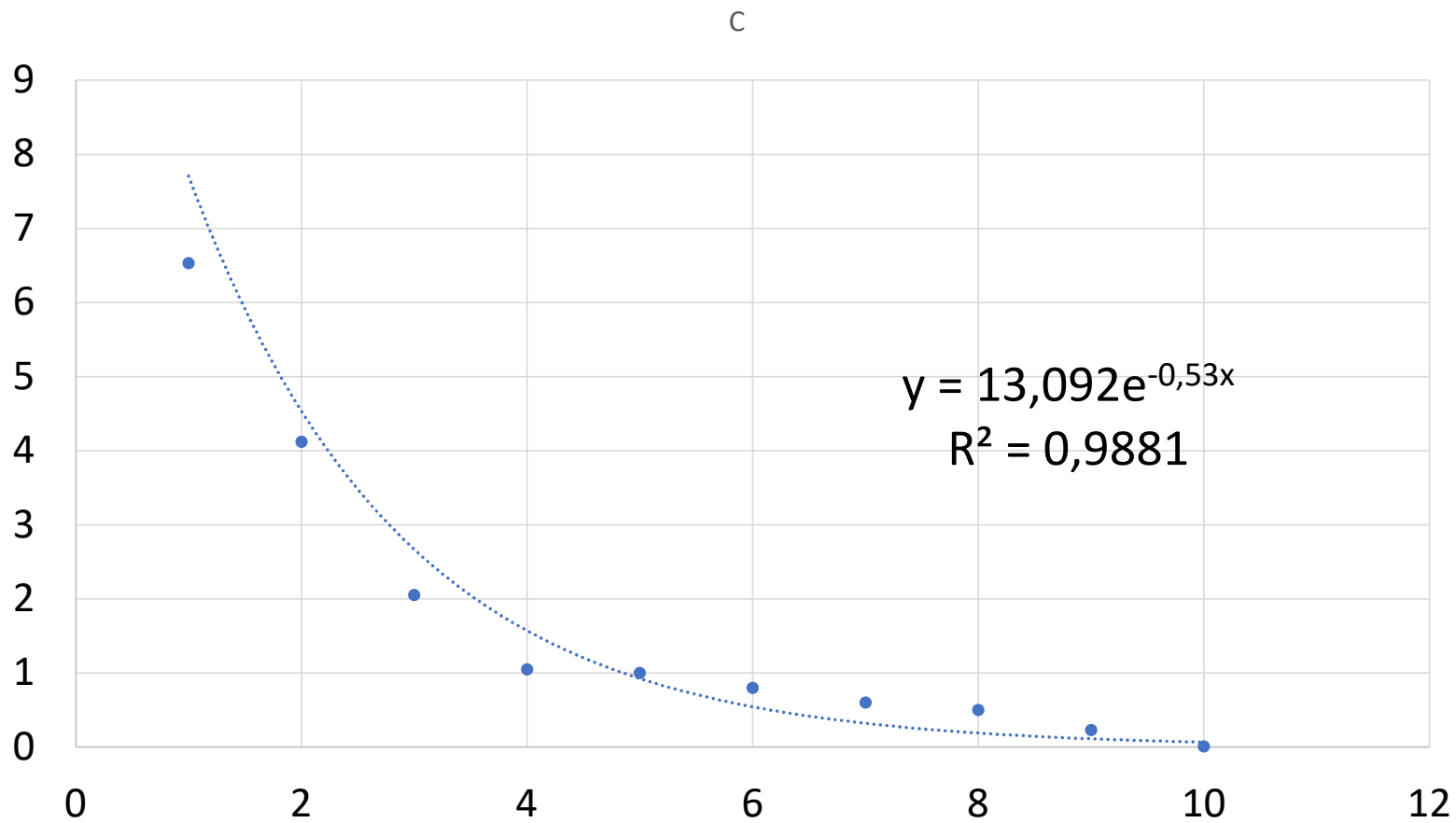
Le modèle linéaire

Nuage de points



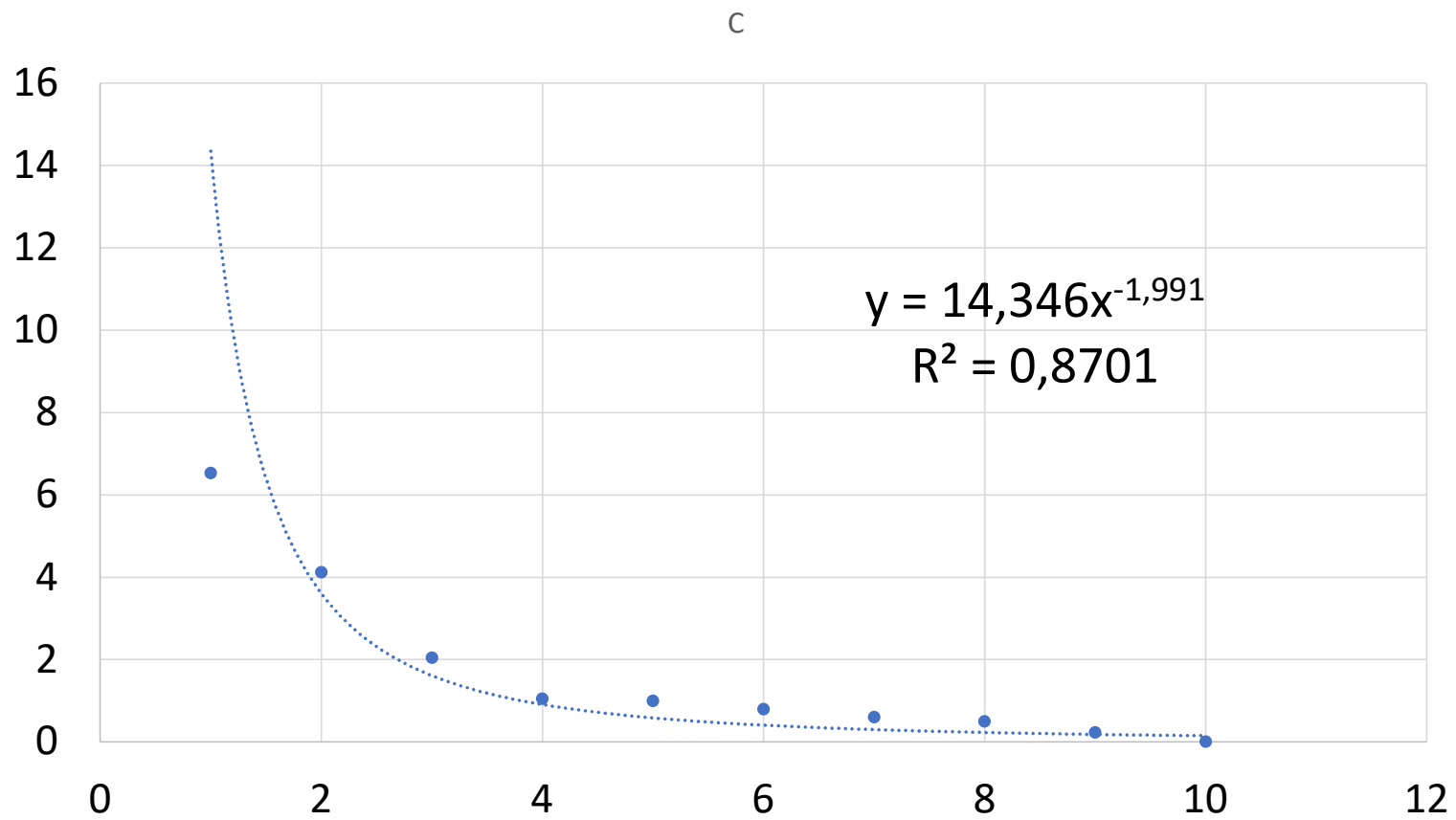
Le modèle exponentiel

Nuage de points



Le modèle puissance

Nuage de points



Choix entre la régression linéaire, exponentielle et puissance

Calcul par SPSS

Modèle	Régression	Coefficient de Détermination
Linéaire	$C = -0,577V + 4,8627$	0,7099
Exponentiel	$C = 13,092e^{-0,53V}$	0,9881
Puissance	$Ph = 14,346X^{-1,991}$	0,8701

*Le coefficient de détermination le plus élevé est
0,9881*

***Le modèle valable c'est le modèle
exponentiel***