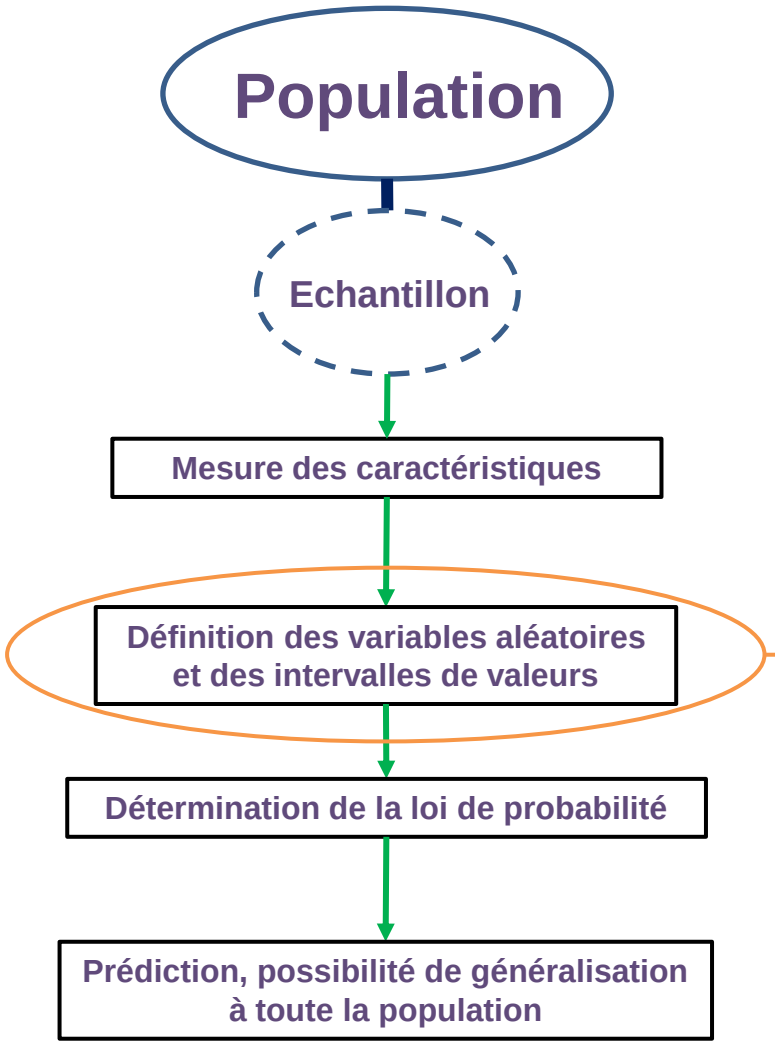
B4 – Mathématiques : La corrélation

Amine ILMANE

Le module B4 – Mathématiques

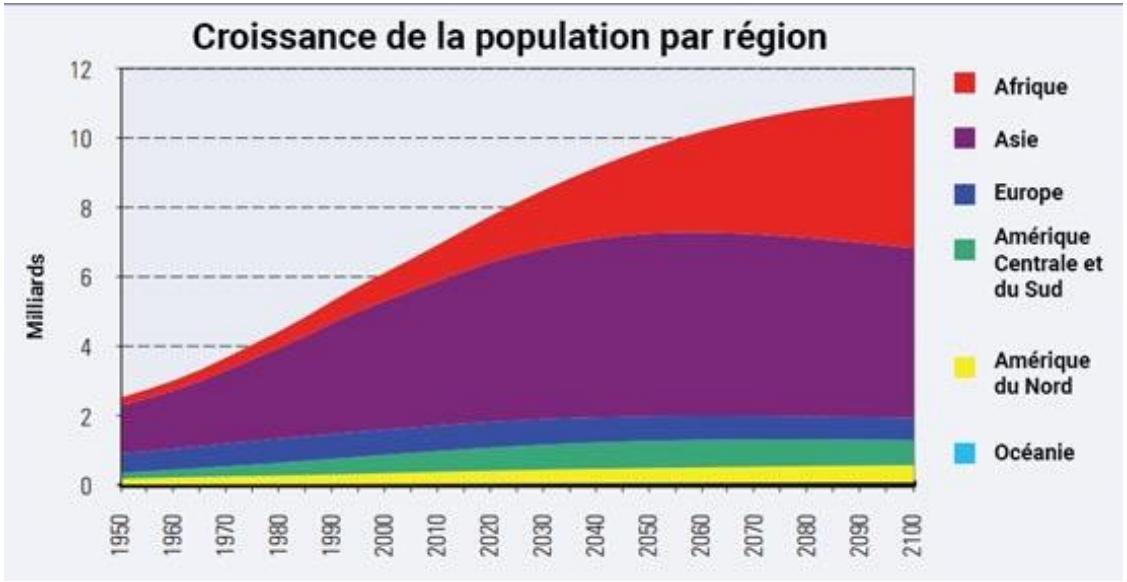


Dans ce chapitre nous allons voir comment étudier les éventuelles influences entre les variables aléatoires.

207demography

Population and regression

Est-ce qu'il nous est possible de **trouver** une **expression** qui **relie** la **population** au **temps** ?



Projet 206 : demography

Terminal

```
~/B-MAT-100> ./105demography -h
USAGE
    ./105demography [code]+

DESCRIPTION
    code     country code
```

Country Name	Country Code	1960	1961	1962	...	2015	2016	2017
Afghanistan	AFG	8996351	9166764	9345868		33736494	34656032	35530081
Albania	ALB	1608800	1659800	1711319		2880703	2876101	2873457
Algeria	DZA	11124888	11404859	11690153		39871528	40606052	41318142
American Samoa	ASM	20013	20486	21117		55537	55599	55641
Andorra	AND	13411	14375	15370		78014	77281	76965
Angola	AGO	5643182	5753024	5866061		27859305	28813463	29784193
Antigua and Barbuda	ATG	55339	56144	57144		99923	100963	102012
Argentina	ARG	20619075	20953077	21287682		43417765	43847430	44271041
Armenia	ARM	1874120	1941491	2009526		2916950	2924816	2930450
Aruba	ABW	54211	55438	56225		104341	104822	105264
Australia	AUS	10276477	10483000	10742000		23815995	24190907	24601860
Austria	AUT	7047539	7086299	7129864		8642699	8736668	8797566

Projet 206 : demography

In the following, Y is the population (in million people) and X the year. With one or several country codes as inputs, your program will print:

1. the a_X and b_X coefficients of the linear fit $Y = a_X X + b_X$,
2. the root-mean-square deviation of this fit,
3. the population prediction in 2050 according to this fit,
4. the a_Y and b_Y coefficients of the linear fit $X = a_Y Y + b_Y$,
5. the root-mean-square deviation of this fit,
6. the population prediction in 2050, according to this fit,
7. the correlation coefficient between X and Y.

```
Terminal
~/B-MAT-400> ./207demography EUU
Country: European Union
Fit1
    Y = 1.62 X - 2749.67
    Root-mean-square deviation: 5.22
    Population in 2050: 570.85
Fit2
    X = 0.60 Y + 1707.97
    Root-mean-square deviation: 5.32
    Population in 2050: 574.54
Correlation: 0.9820
```

Projet 206 : demography

In the following, Y is the population (in million people) and X the year. With one or several country codes as inputs, your program will print:

1. the a_X and b_X coefficients of the linear fit $Y = a_X X + b_X$,
2. the root-mean-square deviation of this fit,
3. the population prediction in 2050 according to this fit,
4. the a_Y and b_Y coefficients of the linear fit $X = a_Y Y + b_Y$,
5. the root-mean-square deviation of this fit,
6. the population prediction in 2050, according to this fit,
7. the correlation coefficient between X and Y.

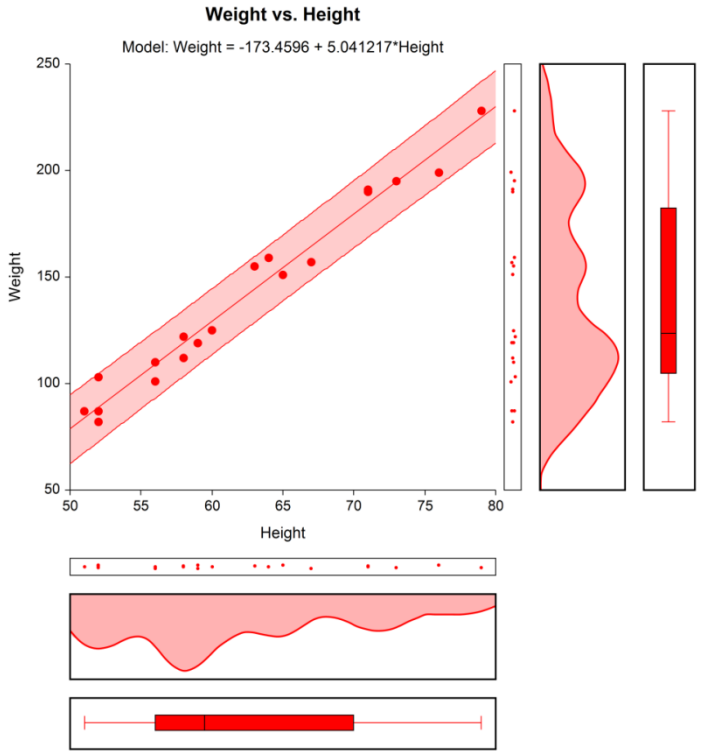
Additionner les populations

Terminal

```
~/B-MAT-400> ./207demography BRA BOL PER
Country: Bolivia, Brazil, Peru
Fit1
    Y = 3.06 X - 5906.34
    Root-mean-square deviation: 2.22
    Population in 2050: 359.35
Fit2
    X = 0.33 Y + 1932.53
    Root-mean-square deviation: 2.22
    Population in 2050: 359.70
Correlation: 0.9991
```

Chapitre 7 : la corrélation

- Relation entre variable
- Méthodes d'ajustement
- Théorie de la corrélation

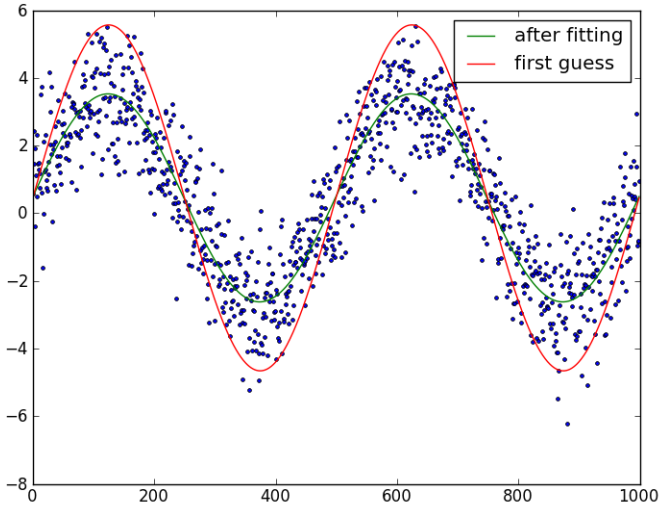


Traitement de données

Nous avons vu que pour une ‘’population’’ donnée (au sens large), il possible de définir plusieurs **caractéristiques** (que l’on représente par des **variables aléatoires**).

Dans ce chapitre, nous allons voir comment étudier les liens que pourraient avoir ces variables aléatoires.

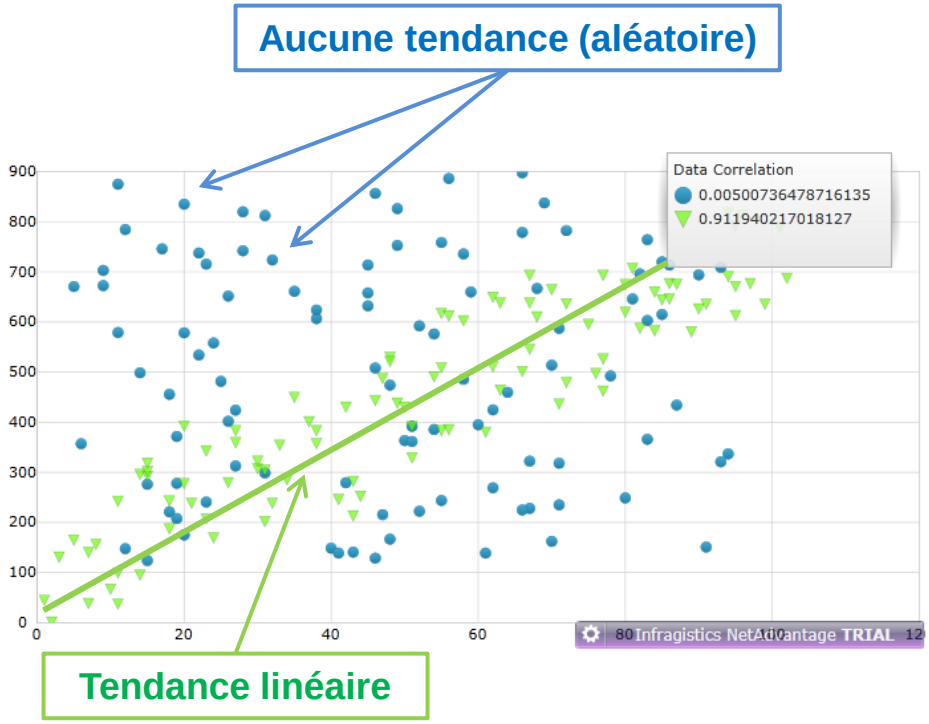
Age	Months on Job	Annual Salary	Weight	Years of Education	How Many People You Supervise	Internet Usage	Age at Birth of First Child
18	12	18	116	12	2	20	16
18	4	20	120	13	0	19	16
18	1	14	117	12	0	18	17
19	12	29	116	13	5	18	18
19	6	22	120	14	1	17	17
20	22	25	130	14	10	16	20
20	8	27	132	14	2	15	19
21	4	30	140	15	6	14	20
21	2	35	150	15	0	8	17
22	53	38	130	15	20	14	20
25	65	48	148	16	30	13	22
25	2	50	130	18	0	12	25
27	7	40	135	18	1	11	21
28	8	48	136	18	1	10	20
29	45	45	137	19	25	9	25
30	2	50	137	19	0	8	28
30	120	65	138	20	40	8	30
34	75	67	139	22	0	0	30
35	26	150	149	22	300	17	35
40	3	66	160	20	0	5	38
40	12	80	149	20	45	4	30
41	23	95	152	21	98	3	31
42	264	80	130	21	56	2	32
45	16	150	170	22	267	0	41
45	270	70	160	22	75	0	40



Traitement de données

Pour vérifier si un lien existe entre les variables, il suffit juste de visualiser le graphe : G = { (X_i , Y_i) }

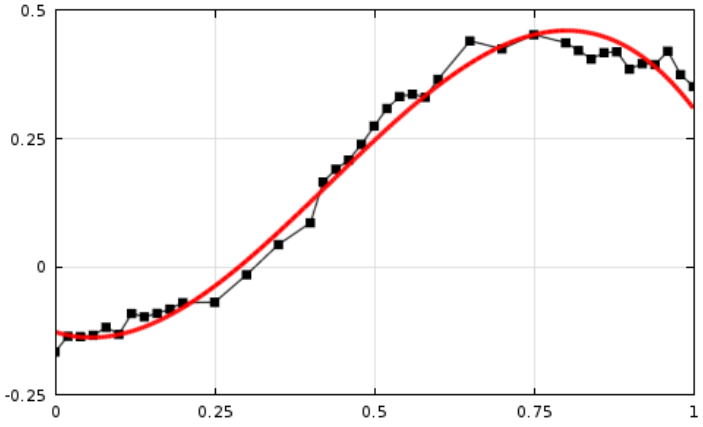
Age	Months on Job	Annual Salary	Weight	Years of Education	How Many People You Supervise	Internet Usage	Age at Birth of First Child
18	12	18	116	12	2	20	16
18	4	20	120	13	0	19	16
18	1	14	117	12	0	18	17
19	12	29	116	13	5	18	18
19	6	22	120	14	1	17	17
20	22	25	130	14	10	16	20
20	8	27	132	14	2	15	19
21	4	30	140	15	6	14	20
21	2	35	150	15	0	8	17
22	53	38	130	15	20	14	20
25	65	48	148	16	30	13	22
25	2	50	130	18	0	12	25
27	7	40	135	18	1	11	21
28	8	48	136	18	1	10	20
29	45	45	137	19	25	9	25
30	2	50	137	19	0	8	28
30	120	65	138	20	40	8	30
34	75	67	139	22	0	0	30
35	26	150	149	22	300	17	35
40	3	66	160	20	0	5	38
40	12	80	149	20	45	4	30
41	23	95	152	21	98	3	31
42	264	80	130	21	56	2	32
45	16	150	170	22	267	0	41
45	270	70	160	22	75	0	40

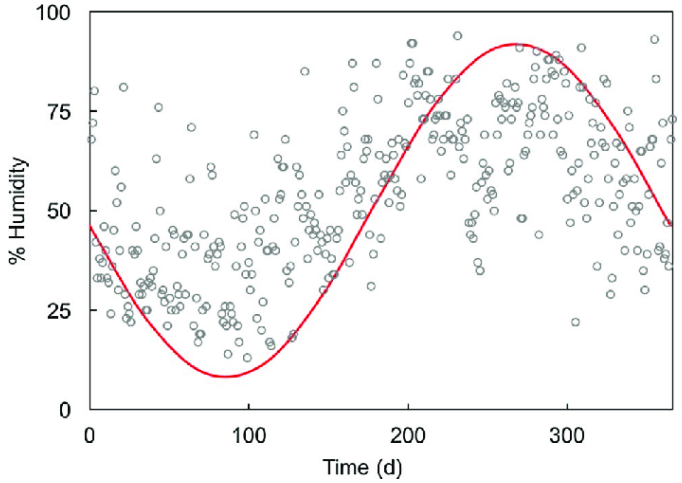


Courbe d'ajustement

Une fois le lien établi, il serait intéressant de trouver une expression du type **Y = f**(**X**). Ce qui permettra de faire des prédictions : **extrapoler**, ou, de déduire des valeurs intermédiaires : **interpoler**

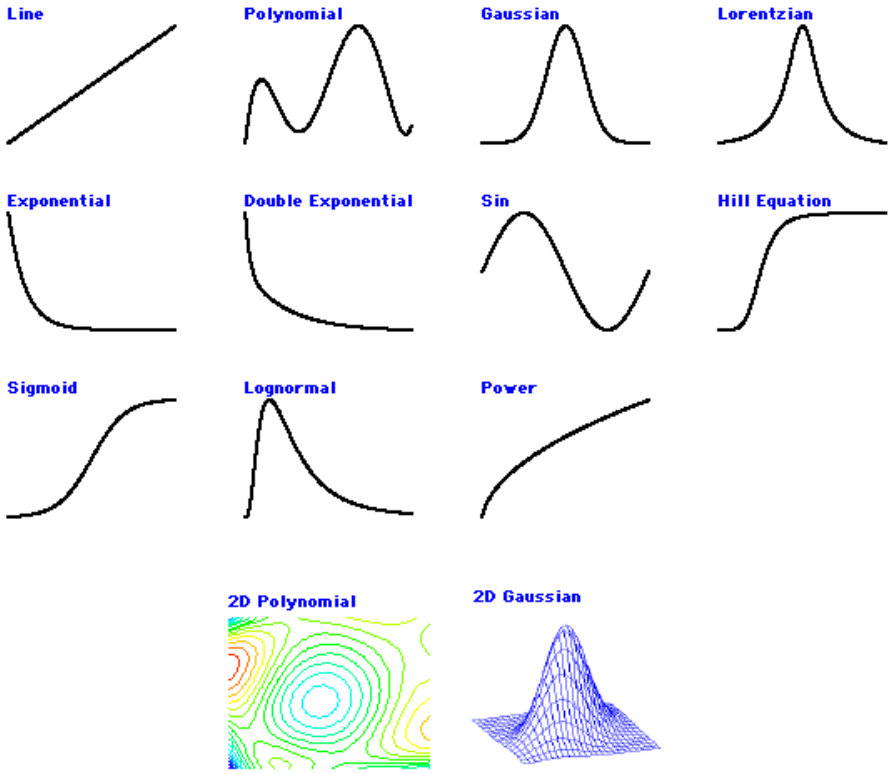
Cette étape s'appelle **ajustement de données (data fit).**

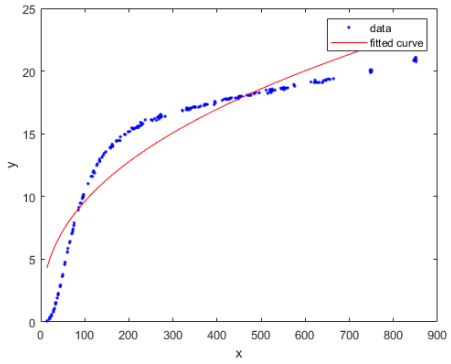


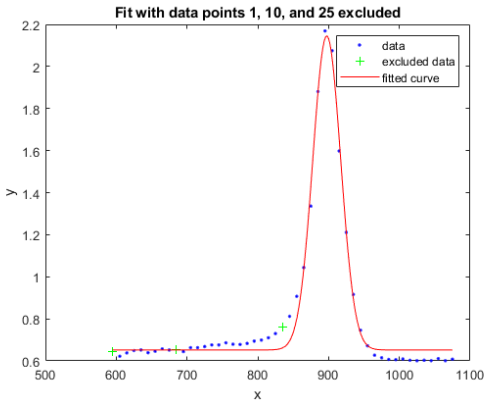


Courbe d'ajustement

Souvent il est difficile, voir impossible, de connaitre l'expression de **f**. Un solution intéressante est de choisir une fonction qui lui ressemble et dont l'expression nous ait connue.

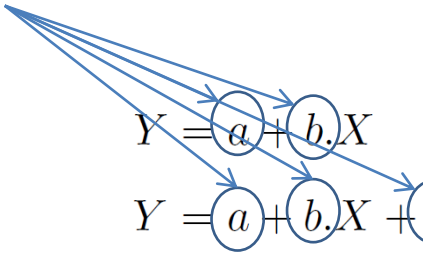






Courbe d'ajustement

Maintenant que nous avons établi le lien et déterminé une forme ou une expression, il faut à présent **déterminer le jeux paramètres** qui nous donnera la courbe la plus proche possible des points.



Droite $Y = a + b.X$

Parabole $Y = a + b.X + c.X^2$

Cubique $Y = a + b.X + c.X^2 + d.X^3$

Polynôme $Y = a_0 + a_1.X + ... + a_n.X^n$

Hyperbole $Y = \frac{1}{a + b.X}$ ou $\frac{1}{Y} = a + b.X$

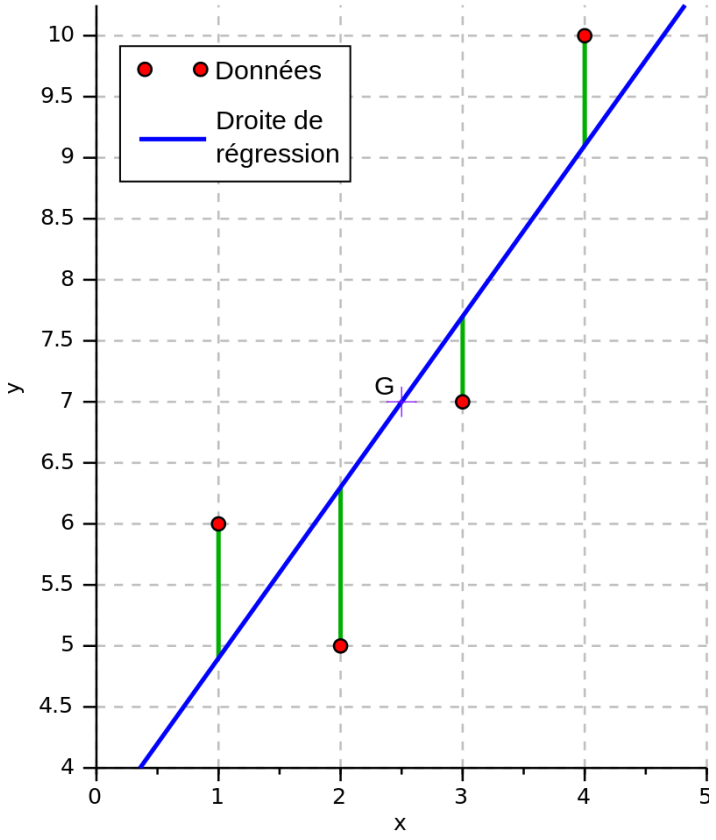
Exponentielle $Y = Y = a.X^b$ ou $log(Y) = log(a) + b.log(X)$

Géométrique $Y = Y = a.b^X$ ou $log(Y) = log(a) + log(b).X$

Erreur

$$Y_i = f(X_i) + \epsilon_i$$

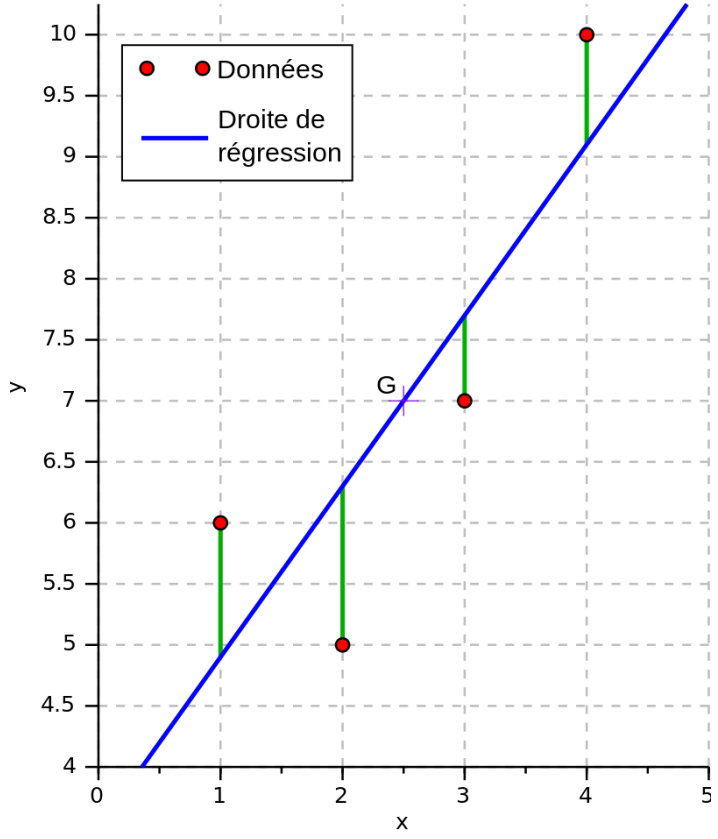
- Etant donnée que l'ajustement est une approximation une erreur est donc commise. Plus elle petite plus l'ajustement est de qualité.
- Les erreurs proviennent d'un mauvais choix de la fonction d'ajustement, ou d'une mauvaise estimation des paramètres si les données sont trop peu nombreuses.
- **Minimiser l'erreur** est un bon critère d'ajustement, il est souvent utilisée pour déterminer les paramètres.



Méthodes des moindres carrées

$$\min_{(a_p)} \sum_{i=1}^{N} [Y_i - f(X_i)]^2 = \min_{(a_p)} \sum_{i=1}^{N} \epsilon_i^2 = \min_{(a_p)} [\epsilon_1^2 + \epsilon_2^2 + ... + \epsilon_N^2]$$

Le plus de cette méthode est que l'on minimise **la somme des carrées de l'erreur.**



N. B.
Chaque méthode donnera un jeu de paramètres différents.

Droite des moindres carrées

$$Y = a + b.X$$

Le critère des moindres carrées, dans le cas où la courbe d'ajustement est une droite, est équivalent au système d'équations suivant :

$$\begin{cases} \sum_{i=1}^{N} Y_i = a.N + b.\sum_{i=1}^{N} X_i \\ \sum_{i=1}^{N} X_i.Y_i = a.\sum_{i=1}^{N} X_i + b.\sum_{i=1}^{N} X_i^2 \end{cases}$$

Les paramètres **a** et **b** (solutions du système) sont alors donnés par :

$$a = \frac{(\sum Y).(\sum X^2) - (\sum X).(\sum XY)}{N.(\sum X^2) - (\sum X)^2} \qquad b = \frac{N.(\sum XY) - (\sum X).(\sum Y)}{N.(\sum X^2) - (\sum X)^2}$$

Parabole des moindres carrées

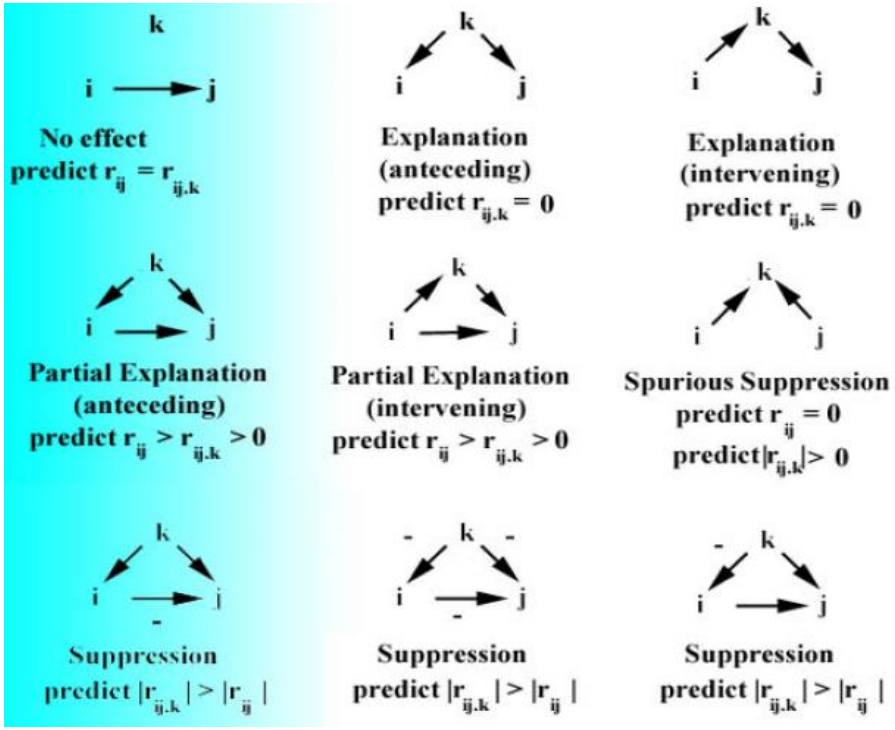
$$Y = a + b.X + c.X^2$$

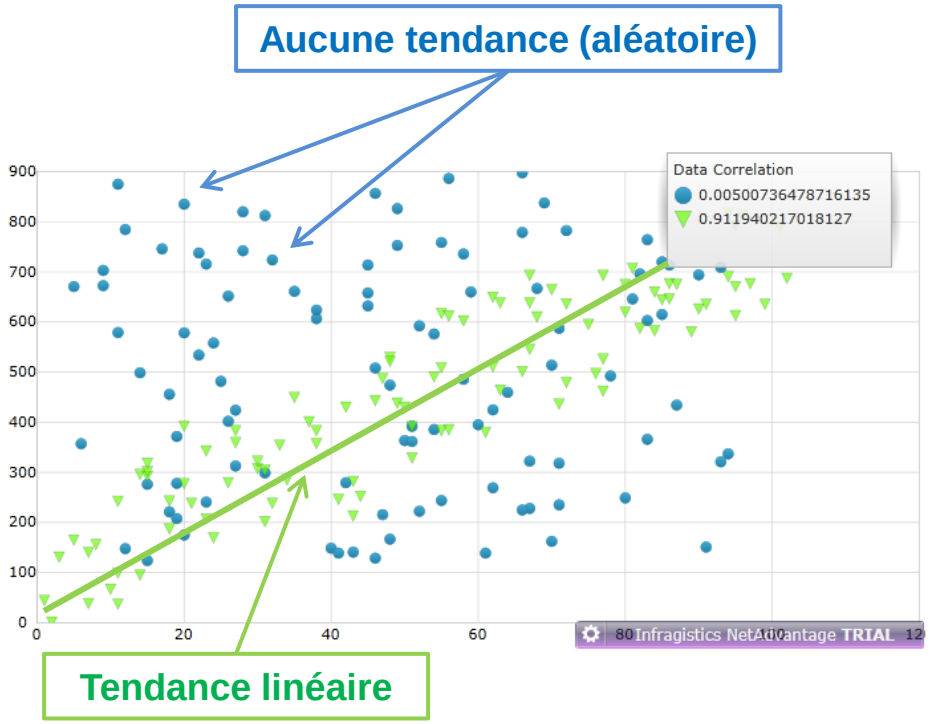
Le critère des moindres carrées, dans le cas où la courbe d'ajustement est une parabole, est équivalent au système d'équations suivant :

$$\begin{cases} \sum Y & = a.N & +b.\sum X & +c.\sum X^2 \\ \sum XY & = a.\sum X & +b.\sum X^2 & +c.\sum X^3 \\ \sum X^2Y & = a.\sum X^2 & +b.\sum X^3 & +c.\sum X^4 \end{cases}$$

La corrélation

- Les variables peuvent être liée, plus ou moins fortement ou pas du tout. IL est intéressant d'avoir un indicateur numérique qui **<u>représentera la force du lien</u>**.
- Puisque les variables s'influence mutuellement **on dit qu'elles sont corrélées**.
- Le fait que les variables soient liées n'implique en rien un rapport de causalité.





Coefficient de corrélation

Variation expliquée et inexpliquée : La variation totale de **Y** est définie comme la somme des carrées des écarts entre les valeurs de Y et sa moyenne.

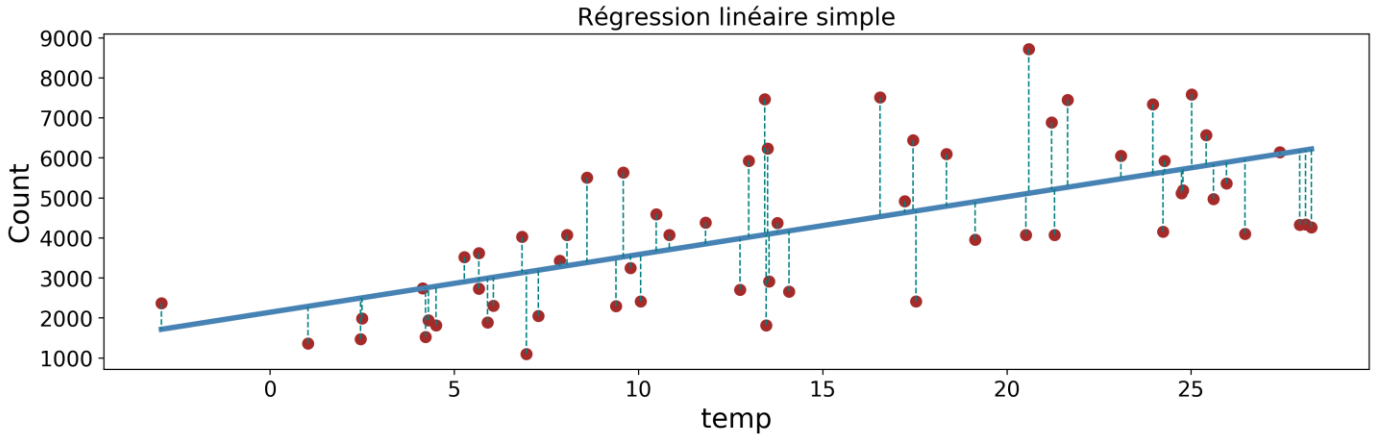
Variation inexpliquée :
Les écarts se distribue de façon aléatoire

Données obtenues par ajustement

$$\sum (Y - \overline{Y})^2 = \sum (Y - f(X))^2 + \sum (f(X) - \overline{Y})^2$$

Données d'origine

Variation expliquée :
Les écarts ont une forme définie



Coefficient de corrélation

$$r = \sqrt{\frac{\sum (f(X) - \overline{Y})^2}{\sum (Y - \overline{Y})^2}}$$

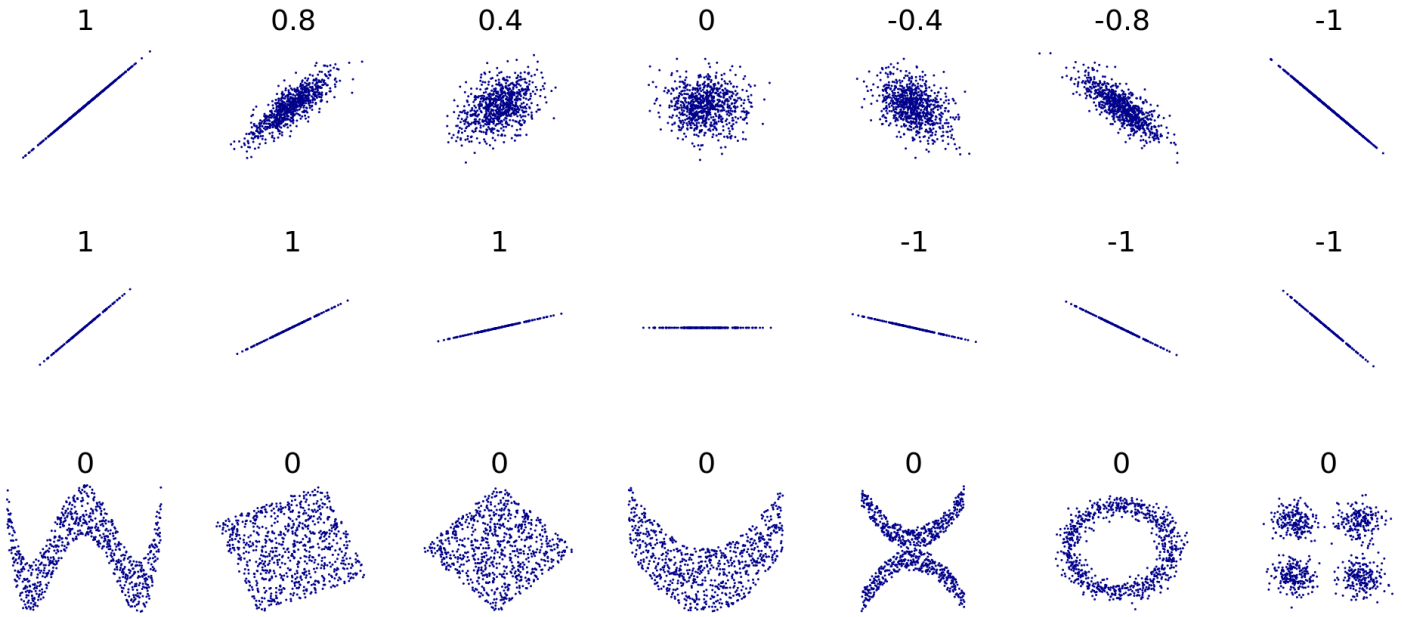
Variation expliquée

Variation totale

Coefficient de corrélation

Ce rapport est compris entre **0** et **1** :

- Si **r = 1**, la corrélation est **totale** et la variation **inexpliquée est nulle** ;
- Si **r = 0**, **aucune** corrélation et la **variation expliquée est nulle** ;



Coefficient de corrélation

$$r = \sqrt{\frac{\sum (f(X) - \overline{Y})^2}{\sum (Y - \overline{Y})^2}} = \sqrt{1 - \frac{s_{Y,X}^2}{s_Y^2}}$$

Ecart-type de Y :

$$s_Y = \sqrt{\frac{\sum (Y - \overline{Y})^2}{N}}$$

Ecart-type de l'ajustement :

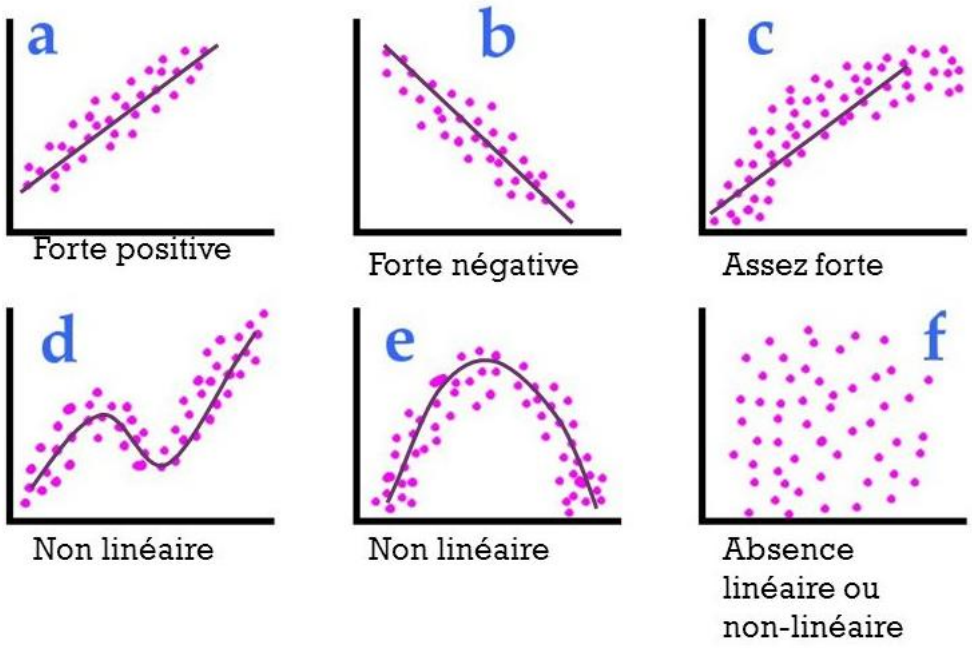
$$s_{Y,X} = \sqrt{\frac{\sum (Y - f(X))^2}{N}} = \sqrt{\frac{\sum \epsilon^2}{N}}$$

Se lit écart-type de l'estimation de **Y** sur **X**

$$Y = a + b.X$$

Coefficient de corrélation

- Le coefficient de corrélation se calcule une fois que la fonction d'ajustement **f** ait était choisie. **Il est différent pour chacun des choix de f**.
- Si l'on choisit une droite, par exemple, et si **r = 0**, tout ce que l'on peut dire, est que le lien entre les variables n'est pas linéaire.
- Si la relation est non-linéaire et monotone, on utilisera le coefficient de corrélation de Spearman.



corrélation et Covariance (Bravais-Pearson)

$$r = \frac{s_{XY}}{s_X . s_Y}$$

$$s_{XY} = \text{Cov(X,Y)} = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{N}$$

Covariance de **X** et **Y**

$$s_X = \sqrt{\frac{\sum (X - \overline{X})^2}{N}}$$

Ecart-types de **X**

$$s_Y = \sqrt{\frac{\sum (Y - \overline{Y})^2}{N}}$$

Ecart-types de **Y**