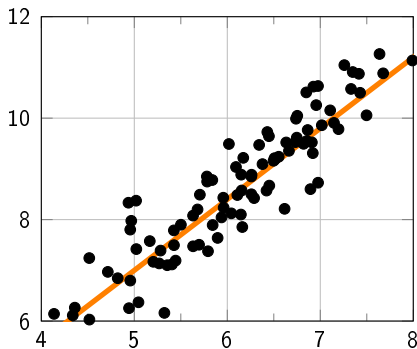




Statistiques à deux variables

I. Mise en situation

I. 1) Deux variables

On étudie **deux** espèces de grenouilles. On relève la longueur des pattes avant et la longueur des pattes arrières de 100 individus pour chaque espèce.

Deux variables :

À **chaque individu** on associe **deux caractères** (deux nombres)

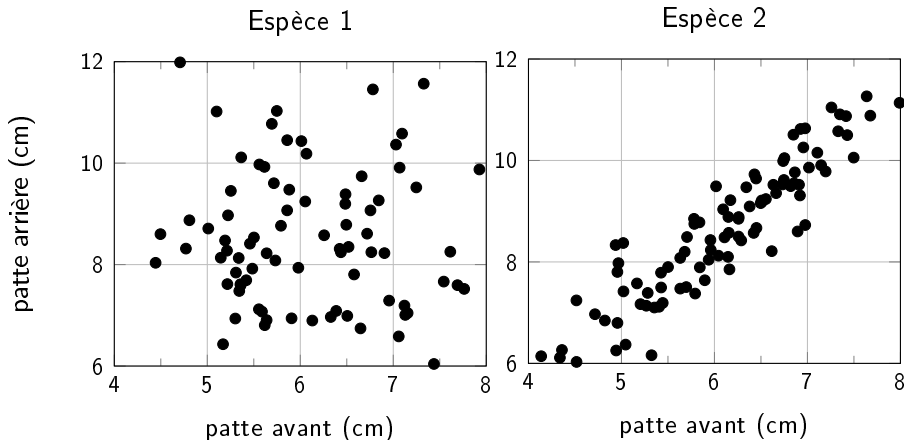
Exemple

Première grenouille : patte avant = 7,3cm, patte arrière = 9,1cm

Deuxième grenouille : ...

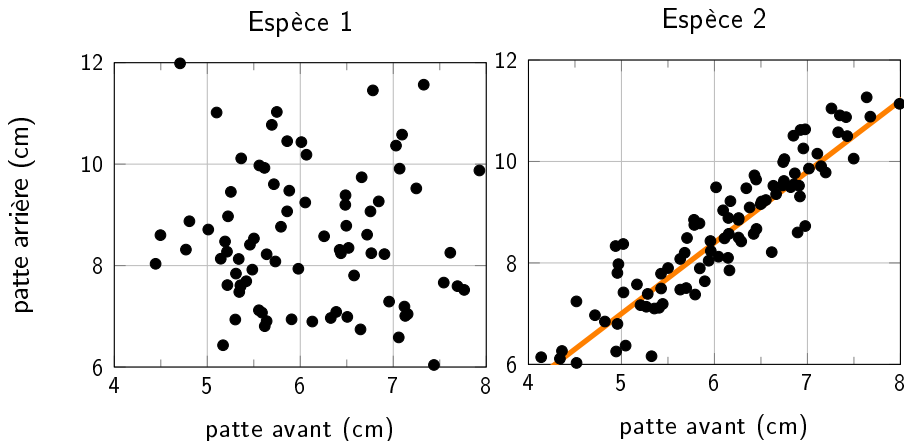
I. 2) Nuage de points

Pour chaque individu, on place un point dans un graphique. x est la longueur de la patte avant et y est la longueur de la patte arrière.



I. 2) Nuage de points

Dans le premier cas, il n'y a aucun lien apparent entre x et y . Dans le deuxième cas, les points sont **à peu près répartis autour d'une droite**.



I. 3) Modèle

On se demande s'il y a un lien entre x et y .

- On peut le supposer pour des raisons théoriques. On a produit un raisonnement qui nous amène à poser ce lien.
- On peut simplement constater que, d'après les expériences, une formule semble relier x et y sans qu'on sache le justifier théoriquement.

Ce lien peut prendre la forme $y = f(x)$. C'est un **modèle**. La fonction f peut être ajustée pour s'adapter aux résultats de mesure. Par exemple :

$$y = a \cdot x + b \quad \text{ou} \quad y = b \cdot \exp(-a \cdot x)$$

Ces deux exemples ont des paramètres a et b dont il faut choisir la valeur pour coller au mieux aux résultats observés.



Un modèle ne doit pas avoir trop de réglages. Fermi disait :
« *Avec quatre paramètres, je peux dessiner un éléphant, et avec cinq je lui fais agiter sa trompe* »

Cela veut dire que si un modèle a trop de paramètres de réglage, on arrivera toujours à l'ajuster à n'importe quel nuage de points et donc ce modèle, trop souple, n'a aucune signification.

Le modèle le plus simple (et ayant un intérêt...) est l'**ajustement affine** :

$$y = a \cdot x + b$$

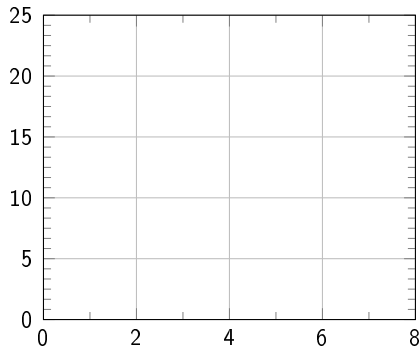
II. Ajustement affine par la méthode des moindres carrés

II. 1) Nuage de points

On étudie la série suivante :

x	1	3	4	6	4.5
y	8	9.5	15	24	12

On place les points correspondant dans un repère.



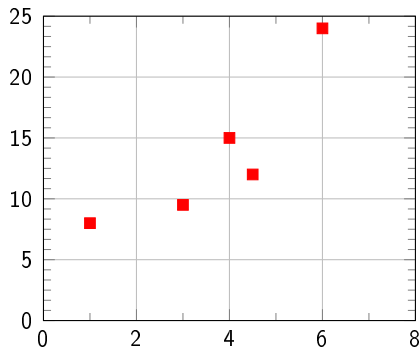
II. 1) Nuage de points

On étudie la série suivante :

x	1	3	4	6	4.5
y	8	9.5	15	24	12

On place les points correspondant dans un repère.

C'est le nuage de points :
L'ensemble des $M_i (x_i; y_i)$.



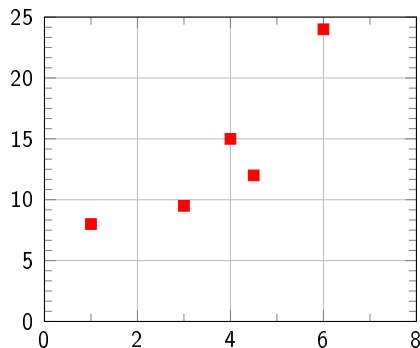
Le i dans les notations sert à numérotter les individus :

Pour $i = 1$, c'est l'individu numéro 1 dont les valeurs sont $(x_1; y_1)$.

II. 2) Point moyen

Moyenne : $\bar{x} = \frac{1}{n} \sum_i x_i$

Remarques :

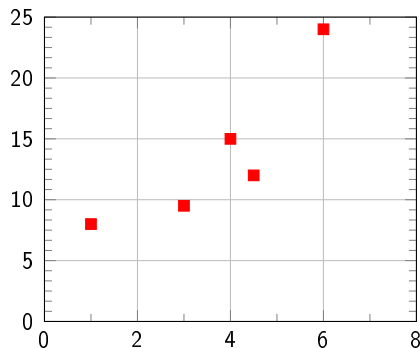


II. 2) Point moyen

Moyenne : $\bar{x} = \frac{1}{n} \sum_i x_i$

Remarques :

- n est l'effectif total. Ici $n = 5$.

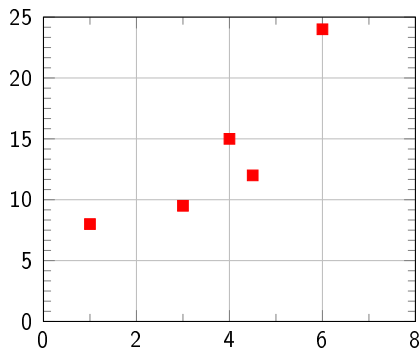


II. 2) Point moyen

Moyenne : $\bar{x} = \frac{1}{n} \sum_i x_i$

Remarques :

- n est l'effectif total. Ici $n = 5$.
- Chaque point a un effectif de 1, ce qui simplifie la formule ($n_i = 1$)

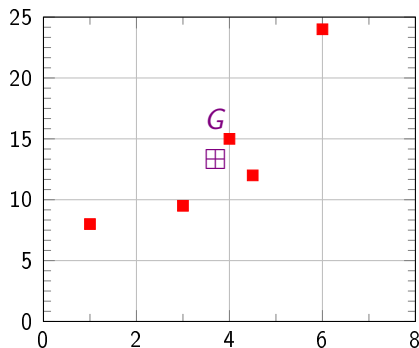


II. 2) Point moyen

Moyenne : $\bar{x} = \frac{1}{n} \sum_i x_i$

Remarques :

- n est l'effectif total. Ici $n = 5$.
- Chaque point a un effectif de 1, ce qui simplifie la formule ($n_i = 1$)
- On calculera $\bar{x}$ et $\bar{y}$ directement avec la **calculatrice**.



Point moyen : $G(\bar{x} ; \bar{y})$

II. 3) Droite

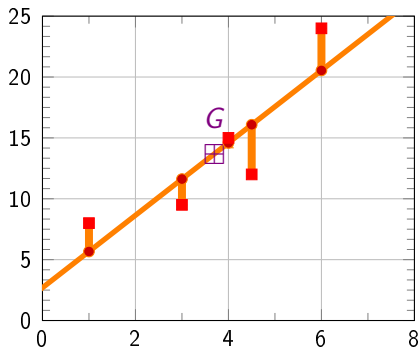
On obtient directement avec la calculatrice :

$$y \approx 2,97 \cdot x + 2,69$$

C'est la droite des moindres carrés.

Cette droite :

- rend minimum le carré des erreurs (*traits verticaux sur la figure*)
- passe par *G*



II. 4) Interpoler - Extrapoler

- Quelle valeur aura y si $x = 2$?
- Quelle valeur aura y si $x = 8$?

II. 4) Interpoler - Extrapoler

- Quelle valeur aura y si $x = 2$?

Avec la droite d'ajustement, $y \approx 2,97 \times 2 + 2,69 = 8,63$.

C'est une **interpolation**.

- Quelle valeur aura y si $x = 8$?

II. 4) Interpoler - Extrapoler

- Quelle valeur aura y si $x = 2$?

Avec la droite d'ajustement, $y \approx 2,97 \times 2 + 2,69 = 8,63$.

C'est une **interpolation**.

- Quelle valeur aura y si $x = 8$?

Avec la droite d'ajustement, $y \approx 2,97 \times 8 + 2,69 = 26,45$.

C'est une **extrapolation**.

III. Calculs

Je donne dans cette partie quelques éléments de calcul et de démonstration. Cela semblera trop technique à certains. Cette partie n'est pas indispensable. La machine fait tous les calculs. Ce qui est indispensable, c'est d'avoir des idées claires sur la signification de la droite d'ajustement (c'est à dire avoir bien compris la partie d'avant)

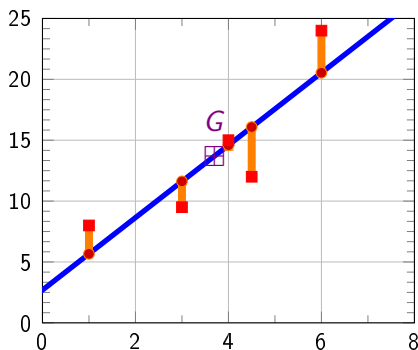
L'équation de la droite en bleu est $y = a \cdot x + b$. Choisissons a et b quelconque puis calculons les erreurs.

Le point numéroté i a les coordonnées $(x_i; y_i)$. Il n'est pas forcément sur la droite. L'écart vertical est :

$$\varepsilon_i = |y - y_i| = |a \cdot x_i + b - y_i|$$

On choisit de considérer le cumul des carrés de ces écarts :

$$TOTAL = \sum_i \varepsilon_i^2 = \sum_i (a \cdot x_i + b - y_i)^2$$



On a sommé les carrés des erreurs verticales. C'est un choix. On aurait pu en faire d'autres ce qui aurait donné d'autres droites.

$$TOTAL(a, b) = \sum_i \varepsilon_i^2 = \sum_i (a \cdot x_i + b - y_i)^2$$

Comme vous le constatez, *TOTAL* est fonction de *a* et *b*. On cherche le choix de *a* et *b* rendant *TOTAL* minimum.

On va chercher $\frac{\partial TOTAL}{\partial a} = 0$ et $\frac{\partial TOTAL}{\partial b} = 0$

Tous calculs faits on trouve :

$$a = \frac{C_{xy}}{V_x} \text{ et } b = \bar{y} - a \cdot \bar{x}$$

$$C_{xy} = cov(x, y) = \frac{1}{n} \sum_i x_i \cdot y_i - \bar{x} \cdot \bar{y} \text{ la } \mathbf{covariance} \text{ entre } x \text{ et } y$$

$$V_x = C_{xx} = \sigma_x^2 \text{ la variance de } x.$$

Remarquez que les unités sont bien homogènes.

Pas du tout demandé en BTS

$$\frac{\partial TOTAL}{\partial a} = \sum_i 2 \cdot x_i \cdot (a \cdot x_i + b - y_i) = 2 \cdot a \cdot \sum_i x_i^2 + 2b \sum_i x_i - 2 \sum_i x_i \cdot y_i$$

$$\frac{\partial TOTAL}{\partial b} = \sum_i 2 \cdot (a \cdot x_i + b - y_i) = 2 \cdot a \cdot \sum_i x_i + 2b \sum_i 1 - 2 \sum_i y_i$$

On sait que $\sum_i 1 = n$, $\sum_i x_i = n \cdot \bar{x}$, $\sum_i y_i = n \cdot \bar{y}$,

$\sum_i x_i^2 = n \cdot (V_x + \bar{x}^2)$ et enfin $\sum_i x_i \cdot y_i = n \cdot (C_{xy} + \bar{x} \cdot \bar{y})$

Les calculs pour ceux que ça intéresse... (suite)

On pose $\frac{\partial TOTAL}{\partial a} = 0$ et $\frac{\partial TOTAL}{\partial b} = 0$, on remplace les $\sum_i$, on divise les deux équations par $2n$ et on obtient :

$$\begin{cases} a \cdot (V_x + \bar{x}^2) + b \cdot \bar{x} - (C_{xy} + \bar{x} \cdot \bar{y}) = 0 \\ a \cdot \bar{x} + b - \bar{y} = 0 \end{cases}$$

La 2e équation donne $b = \bar{y} - a \cdot \bar{x}$ et en remplaçant dans la première on obtient :

$$a \cdot (V_x + \bar{x}^2) + (\bar{y} - a \cdot \bar{x}) \cdot \bar{x} - (C_{xy} + \bar{x} \cdot \bar{y}) = 0$$

Soit en simplifiant :

$$a \cdot V_x - C_{xy} = 0 \Rightarrow a = \frac{C_{xy}}{V_x}$$

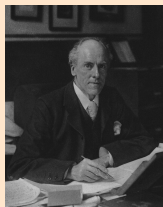
IV. Coefficient r^2

Quand on calcule a et b avec une machine, on reçoit un coefficient r et r^2 .
Que signifie-t-il ?

r est le coefficient de corrélation linéaire.

$$r = \frac{C_{xy}}{\sigma_x \cdot \sigma_y}$$

r est sans unité et $-1 \leq r \leq 1$.



Karl Pearson



r^2 est le coefficient de détermination linéaire de Pearson. Il représente le pourcentage de la variation de y expliquée par le modèle $y = a \cdot x + b$. On sait que $0 \% \leq r^2 \leq 100 \%$. Plus r^2 proche de 100% , meilleur c'est. En général, on veut au moins 99% mais attention, ce n'est pas une règle absolue !

Quelques explications

On raisonne en termes de variances :

- la série y_i a une variance $V_y = \frac{1}{n} \sum_i (y_i - \bar{y})^2$,
- si on note $\hat{y}_i = a \cdot x_i + b$ la valeur prédite par le modèle, la série $\hat{y}_i$ a elle aussi une variance $V_{\hat{y}} = \frac{1}{n} \sum_i (\hat{y}_i - \bar{y})^2$
 $V_{\hat{y}}$ est donc la variance prévue par le modèle $y = a \cdot x + b$.

Remarque : y_i et $\hat{y}_i$ ont la même moyenne $\bar{y}$.

On prouve que :

$$r^2 = \frac{V_{\hat{y}}}{V_y}$$

La part de la variance totale expliquée par le modèle est donc r^2 .

Le calcul, toujours pour ceux que ça intéresse...

D'abord $\frac{1}{n} \sum_i \hat{y}_i = \frac{1}{n} \sum_i (a \cdot x_i + b) = a \cdot \bar{x} + b = \bar{y}$ donc les $\hat{y}_i$ et mes y_i ont bien la même moyenne $\bar{y}$.

$$V_{\hat{y}} = \frac{1}{n} \sum_i (\hat{y}_i - \bar{y})^2 = \frac{1}{n} \sum_i (a \cdot x_i + b - \bar{y})^2$$

Mais on sait que $b = \bar{y} - a \cdot \bar{x}$ donc :

$$V_{\hat{y}} = \frac{1}{n} \sum_i (a \cdot x_i + \bar{y} - a \cdot \bar{x} - \bar{y})^2 = a^2 \cdot \frac{1}{n} \sum_i (x_i - \bar{x})^2 = a^2 \cdot V_x$$

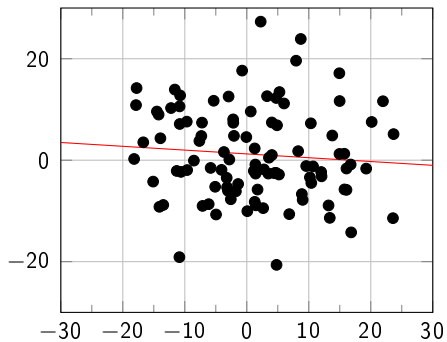
$$\frac{V_{\hat{y}}}{V_y} = a^2 \cdot \frac{V_x}{V_y} = \frac{C_{xy}^2}{V_x^2} \cdot \frac{V_x}{V_y} = \frac{C_{xy}^2}{V_x \cdot V_y} = \left(\frac{C_{xy}}{\sigma_x \cdot \sigma_y} \right)^2 = r^2$$

Quelle doit être une bonne valeur de r^2 ?

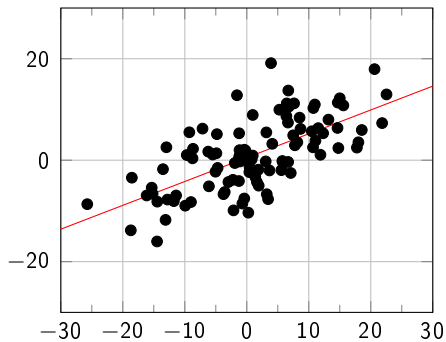
Il n'y a pas de réponse unique à la question. Voici quelques éléments :

- Cela dépend du domaine. En électricité, on attendra quelque chose de plus précis qu'en biologie. En effet, les lois de l'électricité sont mieux maîtrisées et la biologie est beaucoup plus complexe. En physique, on peut exiger un r^2 à 98 % ou même 99 %. En sociologie, on peut se satisfaire d'un r^2 à 70 %.
- Si on attend une loi prédictive comme $U = RI$, on veut un r^2 plutôt élevé.
- Parfois, a et b sont secondaires et c'est le r^2 qui contient l'information.
Par exemple : Je fais une étude liant la consommation de tabac avec l'espérance de vie. Je sais bien sûr que l'espérance de vie n'est pas fixée par la consommation de tabac. Mais peut-être y a-il un lien, une **corrélation**. Si j'obtiens par exemple $r^2 = 30\%$, on peut considérer cette valeur très grande car cela veut dire que la consommation de tabac a une très grande influence sur l'espérance de vie.

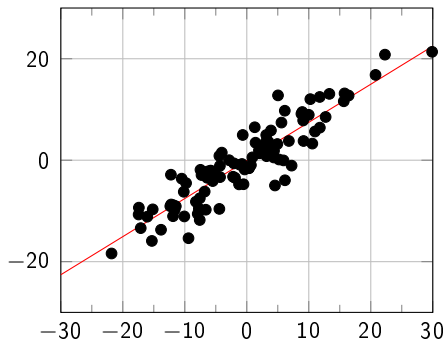
$$r^2 = 1\%$$



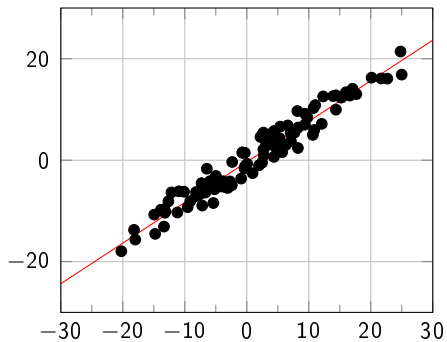
$$r^2 = 44 \%$$



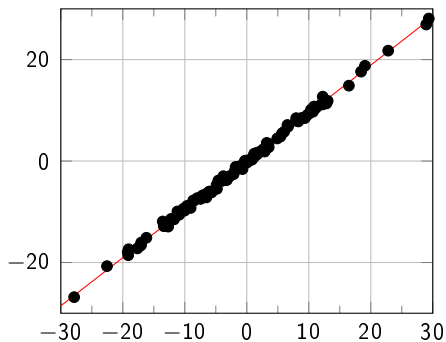
$$r^2 = 84 \%$$



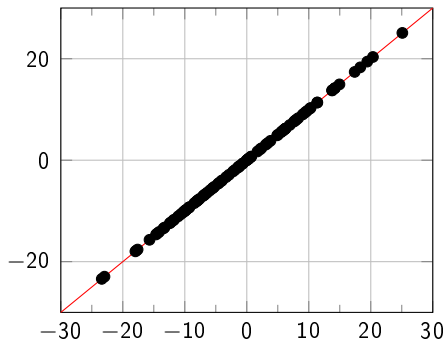
$$r^2 = 95 \%$$



$$r^2 = 99,8 \%$$

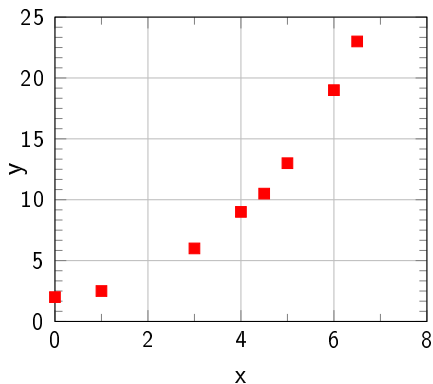


$$r^2 = 100 \%$$



V. Changement de variable

Considérons le nuage ci-contre.
Un ajustement affine ne semble pas approprié.

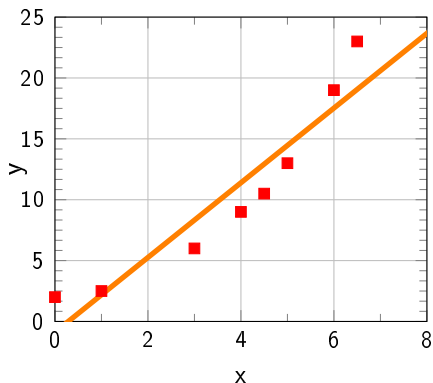


Considérons le nuage ci-contre.
Un ajustement affine ne semble pas approprié.

On peut toutefois faire cet ajustement :

$$y = 3.0676x - 0.8784 \quad r^2 \approx 0,8851$$

L'ajustement n'est pas bon.



Supposons qu'une certaine étude du problème nous a permis de deviner qu'il fallait chercher un ajustement de la forme :

$$y = \exp(a \cdot x + b)$$

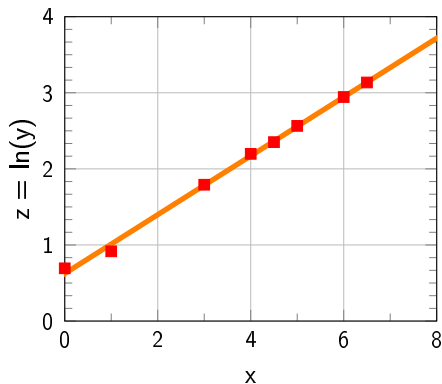
On souhaite chercher a et b . Une bonne façon de le faire est de commencer par calculer :

$$z = \ln(y) = a \cdot x + b$$

On calcule tous les $z_i = \ln(y_i)$ et on obtient un nouveau nuage $(x_i; z_i)$.

Cette fois, l'alignement semble bon. On peut faire l'ajustement de z en x :

$$z = 0,3866 \cdot x + 0,6246 \quad r^2 \approx 0,9974$$

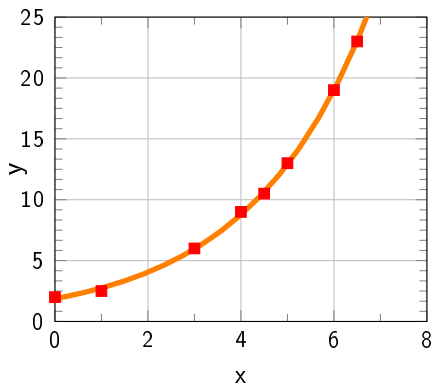


$$z = \ln(y) = 0,3866 \cdot x + 0,6246$$

On en déduit :

$$y = \exp(0,3866 \cdot x + 0,6246)$$

Le résultat est beaucoup plus satisfaisant.

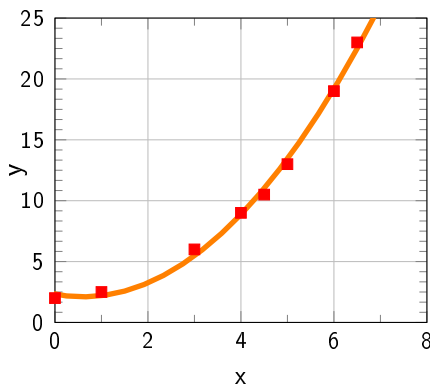


Attention

On est satisfait parce que la courbe du modèle semble passer près des points.

Mais on peut trouver beaucoup de courbes passant près des points. Par exemple :

$$y = 0.5866 \cdot x^2 - 0.7075 \cdot x + 2.3165$$

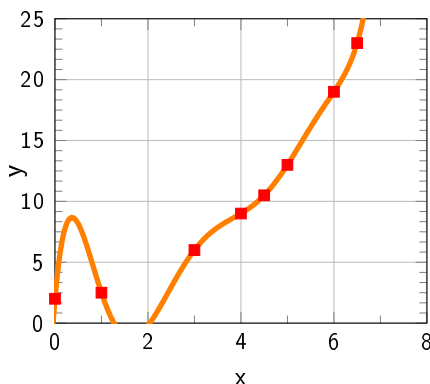


Attention

On peut même trouver un polynôme de degré 7 qui passe exactement par les 8 points !

Mais on voit bien que ce n'est pas satisfaisant.

Passer au plus près des points n'est donc pas un critère définitif. Le modèle que l'on choisit doit rester assez simple et c'est mieux s'il a une justification théorique.



Rappelez-vous, l'éléphant de Fermi : un polynôme de degré 7 a 8 coefficient ajustables. C'est beaucoup trop.

Nouveaux points

Pour bien juger de la qualité d'un modèle, il faut de nouveaux points de mesure en dehors de ceux déjà utilisés dans l'ajustement.

On voit que l'ajustement rouge (degré 7) qui était parfait sur les 8 points connus, se montre incapable de prédire correctement de nouveaux points.

Ce phénomène correspond à ce qu'en intelligence artificielle on appelle le **sur-apprentissage**.

