Interpolation & extrapolation de données

IUT SGM

Y. Morel

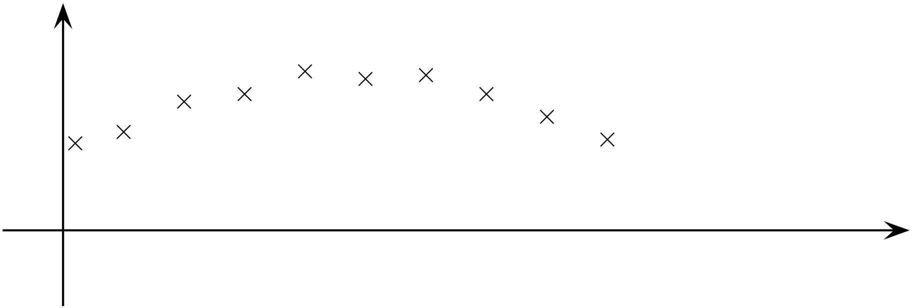
2020/2021

https://xymaths.fr/

Objectif :

Modéliser un ensemble de données, par exemples des résultat expérimentaux, c'est-à-dire formuler une loi permettant de rendre compte de ces résultats.

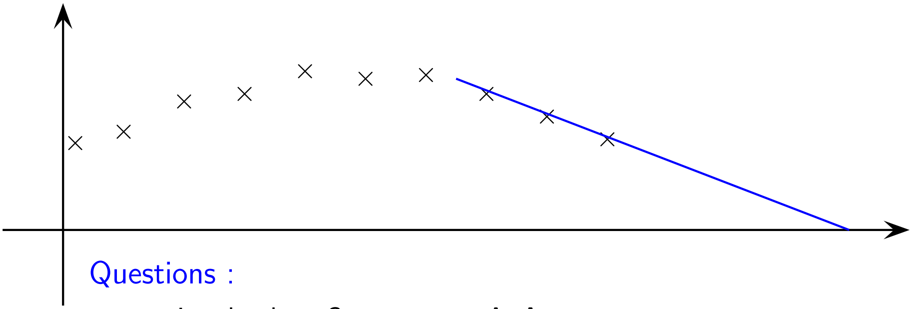
Exemple : Trajectoire d'un objet, positions mesurées :



Objectif :

Modéliser un ensemble de données, par exemples des résultat expérimentaux, c'est-à-dire formuler une loi permettant de rendre compte de ces résultats.

Exemple : Trajectoire d'un objet, positions mesurées :



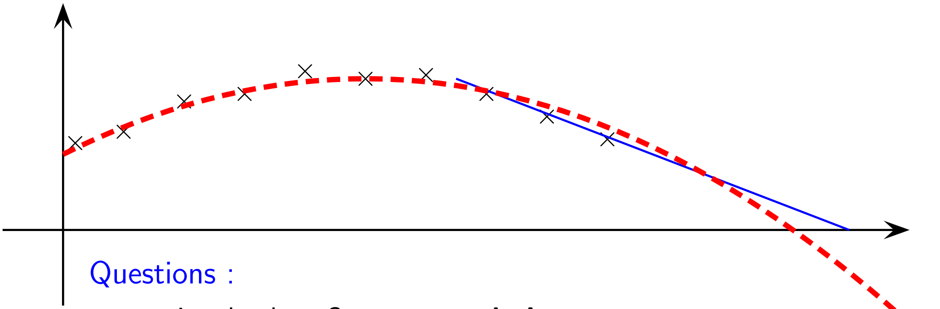
Questions :

- point de chute ? $\rightarrow$ **extrapolation**

Objectif :

Modéliser un ensemble de données, par exemples des résultat expérimentaux, c'est-à-dire formuler une loi permettant de rendre compte de ces résultats.

Exemple : Trajectoire d'un objet, positions mesurées :



Questions :

- point de chute ? $\rightarrow$ **extrapolation**
- trajectoire ? équation ? $\rightarrow$ **modélisation**

Deux principales méthodes :

- Interpolation : on impose à la fonction recherchée de passer "exactement" par tous les points.

  Autant de paramètres que de points, par exemple fonction polynomiale de degré n pour $n+1$ points.

- Optimisation : on cherche une fonction qui passe "au mieux" par les points :
  $\hookrightarrow$ méthode des moindres carrés

Interpolation polynomiale

On a $N+1$ données $A_i\,(x_i;y_i)$ par lesquelles on cherche à "faire passer" un polynôme :

$$P(x) = a_N x^N + a_{N-1}x^{N-1} + \cdots + a_1 x + a_0$$

Les coefficients a_i vérifient le système :

$$\begin{cases} P\left(x_0\right) = a_N x_0^N + a_{N-1}x_0^{N-1} + \cdots + a_1 x_0 + a_0 = y_0 \\ P\left(x_1\right) = a_N x_1^N + a_{N-1}x_1^{N-1} + \cdots + a_1 x_1 + a_1 = y_1 \\ \dots \\ P\left(x_N\right) = a_N x_N^N + a_{N-1}x_N^{N-1} + \cdots + a_N x_N + a_1 = y_N \end{cases}$$

C'est un système linéaire qui s'écrit sous la forme matricielle $MU = B$

<u>Exercice :</u> On considère les trois points $A_0(0;2)$, $A_1(1;4)$ et $A_2(2,2)$.

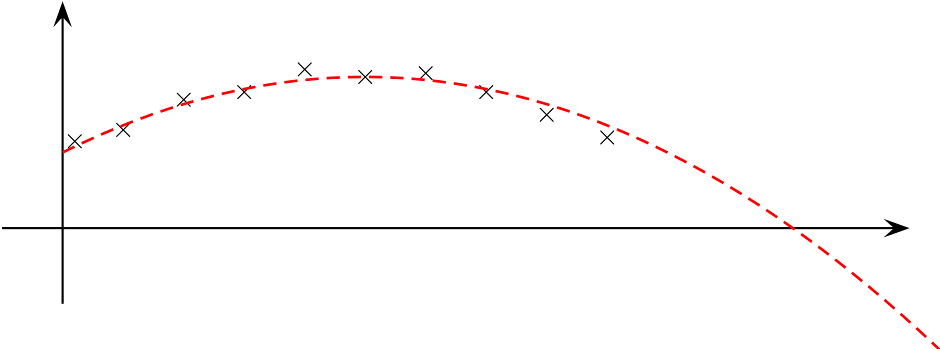
On note $P(x) = ax^2 + bx + c$ le polynôme d'interpolation de degré 2.

Écrire sous forme matricielle $AX = B$ le système vérifié par les coefficients a_0, a_1 et a_2, en précisant les matrices A, X et B.

- En présence d'un grand nombre de points, la méthode d'interpolation par un polynôme de degré élevé peut être instable.
- Les données sont parfois (souvent ?) imprécises, et les contraintes $P(x_i) = y_i$ sont inutilement trop fortes.
- La méthode n'est pas robuste : si un point est "erroné", il influe de manière significative sur tous les coefficients et le polynôme peut être grandement différent.

On souhaite modéliser le nuage de points $A_i\,(x_i; y_i)$ par une fonction qui passe "au mieux" par ces points.

Dans l'exemple du début, d'après la physique sous-jacente : la chute d'un corps, on sait que la trajectoire est parabolique, donc suit la courbe d'un polynôme de degré 2 (et pas plus !)



On cherche donc le polynôme $P(x) = ax^2 + bx + c$.

- En écrivant le système d'équations : $P(x_i) = y_i$, on obtient un système surdéterminé de $N+1$ équations à 3 inconnues, qui n'admet en général pas de solution.

- On cherche alors plutôt que la distance entre les points

  $A_i(x_i; y_i)$ donnés

  et $A'_i(x_i; P(x_i))$ modélisés

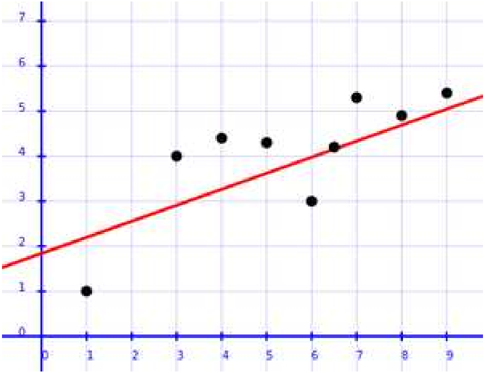
  soit minimale : on cherche a, b et c tels que

$$\begin{aligned} d(a;b;c) &= (P(x_0) - y_0)^2 + (P(x_1) - y_1)^2 + \cdots + (P(x_N) - y_N)^2 \\ &= \sum_{i=1}^{N} (P(x_i) - y_i)^2 \end{aligned}$$

  soit minimal.

Une utilisation courante est l'approximation, ou modélisation, par une fonction affine (polynôme de degré 1) :

$$P(x) = ax + b$$



Voir [Tracer et calcul de la droite des moindres carrés]

Les données $(x_i; y_i)$ sont approchées par le modèle affine : $(x_i; \widetilde{y_i})$ avec
$\widetilde{y_i} = P(x_i) = ax_i + b$
tel que l'erreur quadratique :

$$\begin{aligned} d(a,b) &= \sum_{i=1}^{N} \left[\widetilde{y_i} - y_i\right]^2 \\ &= \sum_{i=1}^{N} \left[(ax_i + b) - y_i\right]^2 \end{aligned}$$

soit minimale.

La droite d'équation alors trouvée est la droite dite des moindres carrés, ou de régression linéaire, ou encore d'ajustement affine.

Plusieurs approches, ou point de vu :

- Optimisation : $d(a,b)$ est minimal lorsque

$$\overrightarrow{\nabla} d = \overrightarrow{0} \iff \frac{\partial\, d(a,b)}{\partial a} = \frac{\partial\, d(a,b)}{\partial b} = 0$$

- Équations normales : si le sytème (surdéterminé) est $MU = B$, avec $U = \left(\begin{array}{c} a \\ b \end{array}\right)$, alors $d(a,b)$ est minimal pour U solution de

$$M^T M U = M^T B$$

- Statistique : $$a = \frac{\text{Cov}\,(X,Y)}{\text{Var}(X)} = \frac{\overline{xy} - \overline{x} \times \overline{y}}{\overline{x^2} - \overline{x}^2}$$

et $\quad b = \overline{y} - a\overline{x}$

Gauss, au tout début du 19ème siècle, a développé cette méthode pour répondre à la question :

le modèle (ici affine) est-il adapté aux données

En effet, on peut toujours calculé la droite des moindres carrés, mais est-elle pertinente ?

Le coefficient de corrélation (ou de détermination dans certain logiciels) est l'indicateur qui permet de quantifier cette pertinence :

$$r = \frac{\mathrm{cov(X,Y)}}{\sigma(X)\sigma(Y)}$$

- Si $|R| \simeq 1$, $(|R| > 0,9)$ le modèle affine est pertinent,
- sinon, $|R| < 0,9$, il vaut mieux essayer de trouver un autre modèle.

Exercice : Durée de vie et maintenance d'équipements.

Les pourcentages $R(t_i)$ des appareils mécaniques encore en service après un nombre t_i d'heures de fonctionnement ont été relevés et notés dans le tableau suivant :

t_i	100	300	500	1000	1500
$R(t_i)$	0,80	0,52	0,32	0,12	0,04

1. Placer les points sur un graphique. Un ajustement affine est-il pertinent ?
2. On pose $y_i = \ln R(t_i)$.
   Peut-on envisager un ajustement affine du nuage de points $B_i\,(t_i; y_i)$?
   Donner l'équation de la droite de régression et en déduire une expression de la forme $R(t) = ke^{-\lambda t}$, avec k et λ des constantes.
3. Déterminer à l'aide du modèle précédent, le nombre d'équipements encore en service au bout de 900 heures de fonctionnement.

Attention : corréler n'est pas expliquer.

De nombreux phénomènes peuvent être mis en corrélation, c'est-à-dire, en termes maintenant plus précis, le modèle (affine par exemple) reliant les grandeurs observées pour ces deux phénomènes a un bon coefficient de corrélation, ce n'est pas une explication pour autant.

1. Position du problème

2. Interpolation

3. Optimisation - Moindres carrés
   - Droite des moindres carrés
   - Modélisation, corrélation, causalité
   - Corrélation entre phénomènes - Quelques exemples !

- Une étude a montré que la fréquence des maladies des personnes habitant à proximité de lignes à haute tension est plus élevée que pour le reste de la population.
  Plus précisément, il y a une corrélation siginificative entre la distance du logement à la ligne haute tension et la fréquence des maladies.

- Une étude a montré que la fréquence des maladies des personnes habitant à proximité de lignes à haute tension est plus élevée que pour le reste de la population.
  Plus précisément, il y a une corrélation siginificative entre la distance du logement à la ligne haute tension et la fréquence des maladies.

  Donc : l'influence de la haute tension est néfaste !

- Une étude a montré que la fréquence des maladies des personnes habitant à proximité de lignes à haute tension est plus élevée que pour le reste de la population.
  Plus précisément, il y a une corrélation siginificative entre la distance du logement à la ligne haute tension et la fréquence des maladies.

  Donc : l'influence de la haute tension est néfaste !

- Il y a une corrélation significative entre la probabilité de mourir et le nombre de jours passés à l'hopital :

- Une étude a montré que la fréquence des maladies des personnes habitant à proximité de lignes à haute tension est plus élevée que pour le reste de la population.
  Plus précisément, il y a une corrélation siginificative entre la distance du logement à la ligne haute tension et la fréquence des maladies.

  Donc : l'influence de la haute tension est néfaste !

- Il y a une corrélation significative entre la probabilité de mourir et le nombre de jours passés à l'hopital :

  Donc : dès entré à l'hôpital, partez en le plus vite possible si vous voulez augmenter vos chances de survie !

- Une étude a montré que la fréquence des maladies des personnes habitant à proximité de lignes à haute tension est plus élevée que pour le reste de la population.
  Plus précisément, il y a une corrélation siginificative entre la distance du logement à la ligne haute tension et la fréquence des maladies.

  Donc : l'influence de la haute tension est néfaste !

- Il y a une corrélation significative entre la probabilité de mourir et le nombre de jours passés à l'hopital :

  Donc : dès entré à l'hôpital, partez en le plus vite possible si vous voulez augmenter vos chances de survie !

"Quand on est malade, il ne faut surtout pas aller à l'hôpital : la probabilité de mourir dans un lit d'hôpital est 10 fois plus grande que dans son lit à la maison" Coluche

- Une étude a montré que la fréquence des maladies des personnes habitant à proximité de lignes à haute tension est plus élevée que pour le reste de la population.
  Plus précisément, il y a une corrélation siginificative entre la distance du logement à la ligne haute tension et la fréquence des maladies.

  Donc : l'influence de la haute tension est néfaste !

- Il y a une corrélation significative entre la probabilité de mourir et le nombre de jours passés à l'hopital :

  Donc : dès entré à l'hôpital, partez en le plus vite possible si vous voulez augmenter vos chances de survie !

"Quand on est malade, il ne faut surtout pas aller à l'hôpital : la probabilité de mourir dans un lit d'hôpital est 10 fois plus grande que dans son lit à la maison" Coluche

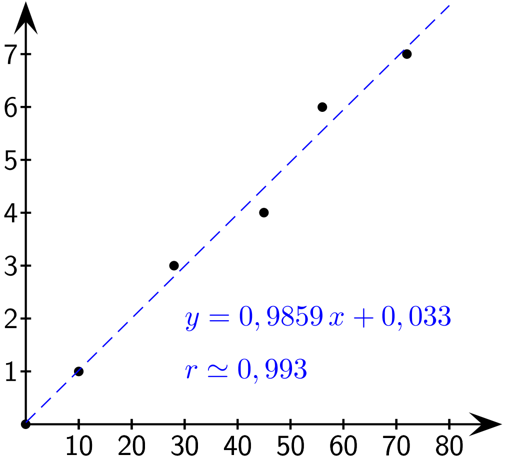
- La majorité des accidents arrivent pour des trajets de moins de 30 km

  Donc : habitez plus loin, ou faites des détours pour aller travailler !

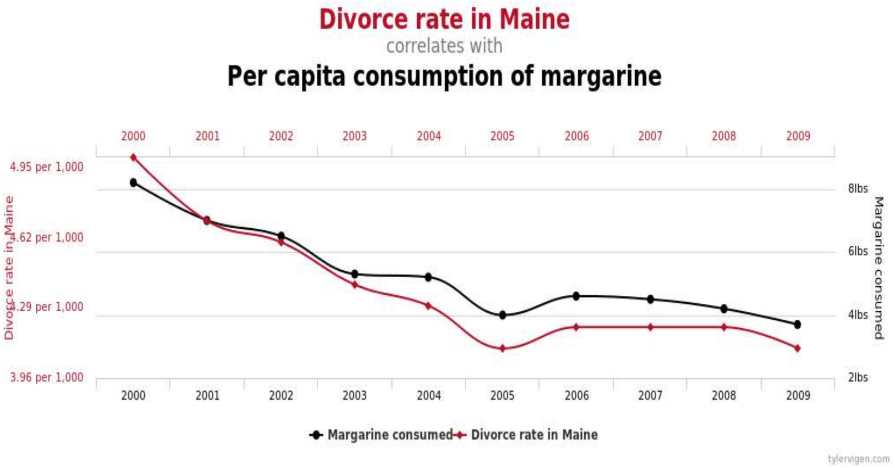
Un exemple détaillé

Nombre de morts par noyade dans une ville de la méditerranée en fonction du nombre de climatiseurs vendus dans la zone commerciale de la ville :

Nombre de clims vendues	Nombre de noyades
0	0
10	1
28	3
45	4
56	6
72	7

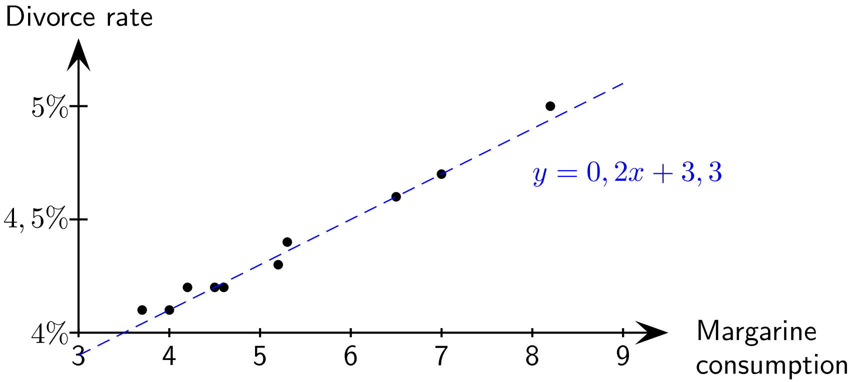


Autre exemple ...



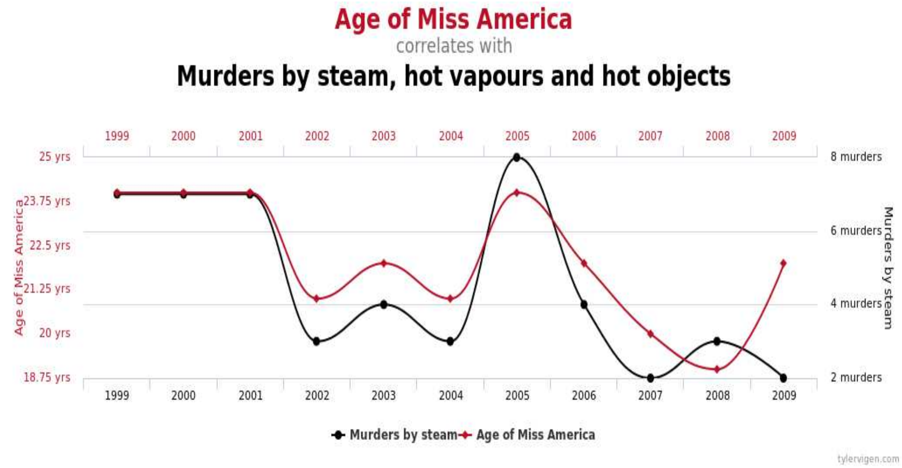
Source : http://tylervigen.com/spurious-correlations

Divorce vs. margarine



Corrélation : $r \simeq 0,993$ ...

Autre exemple, bis, ...



Source : http://tylervigen.com/spurious-correlations