Régression linéaire

I. Ajustement à un modèle affine (*linear fit*)

Position du problème

On dispose d'une série de n mesures sur un couple de variables x et y liées par des mécanismes sujets d'étude :

$$\{x_i, y_i\}_{i=1,..,n}\,.$$

DÉFINITION : *Approximation* ou *Ajustement* = détermination d'une fonction mathématique, désignée plus loin par le mot *modèle*, passant « au mieux » à proximité de la série de points donnée.

Une approximation est utile lorsqu'une loi théorique est recherchée à partir de points de mesure nombreux, mais entachés d'incertitude. On parle alors aussi d'*identification*, ou d'*ajustement*. L'intérêt est multiple :

- **valider le choix des mécanismes** suspectés d'être à l'œuvre dans le phénomène concerné ;
- **déterminer la valeur** de certains paramètres du modèle ;
- permettre la **prévision** ou l'**extrapolation** à partir du modèle obtenu.

Il est toujours possible de faire un ajustement quel qu'il soit, à l'aide d'une fonction simple ou compliquée, mais celui-ci n'a pas forcément de signification physique (ou chimique, biologique, médicale, économique, sociale... quel que soit le domaine). En particulier l'utilisation des polynômes permet toujours d'approximer les données, d'autant mieux que le degré du polynôme est élevé.

Plus un modèle est simple, moins il contient de paramètres, et donc plus il est facile à interpréter (lui donner du sens) et donc à comprendre.
Au contraire, plus un modèle est complexe, plus il contient de paramètres, et plus on a de chance de trouver comment le faire épouser au mieux les données, mais plus il est difficile d'en tirer une perception claire des mécanismes à l'œuvre.
Ainsi, il ne suffit pas qu'un modèle rende compte de très près de la variabilité des données observées, mais il faut aussi qu'il ait du sens.

C'est la raison pour laquelle on privilégie souvent la *régression linéaire*, c'est-à-dire **la recherche d'un modèle affine**, dépend seulement de 2 paramètres [1], quitte à reformuler la loi théorique recherchée pour lui faire prendre la forme d'une relation affine.

Exemple : Au cours de la décharge d'un condensateur dans un circuit $R-C$ série, on enregistre une série de valeurs de la tension u en fonction du temps t : $\{t_i, u_i\}_{i=1,..,n}$. Puis on souhaite vérifier la loi théorique

$$u(t) = E\,e^{-\frac{t}{\tau}}\,,$$

et déterminer expérimentalement les 2 paramètres E et τ optimaux. On reformule :

$$u(t) = E\,e^{-\frac{t}{\tau}} \quad \Leftrightarrow \quad \ln(u) = \ln(E) - \frac{t}{\tau}\,.$$

En posant $y = \ln(u)$ et $x = t$, on peut ainsi identifier une relation affine

$$y = f(x) = ax + b \quad \text{avec} \quad a = -\frac{1}{\tau} \quad \text{et} \quad b = \ln(E)\,.$$

Ainsi on trace le graphe de y en fonction de x, c'est-à-dire le nuage des points (x_i, y_i), puis on utilise une méthode objective pour choisir de manière optimale les paramètres a et b.

Dans la suite, on aborde la question de cette détermination de a et b, mais pas celle de leur incertitude, qui sera abordée plus tard.

II. Approximation par moindres carrés

Le modèle choisi est affine [2] : $y = f(x) = ax + b$. Les valeurs de y calculées à partir du modèle sont notées

$$\hat{y}_i = f(x_i) = ax_i + b\,.$$

Pour choisir de façon rationnelle et optimale les paramètres a et b, on doit se fixer un critère d'optimisation.

DÉFINITION : **Moyenne** (arithmétique « empirique », *mean*) : $\boxed{\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i}$

1. Ordonnée à l'origine et coefficient directeur.
2. Le mot « linéaire » remplace souvent « affine » en physique par abus de langage, ou par anglicisme.

DÉFINITION : **Variance**[3] des observations (*Sum of Squares Total, SST*) :

$$\text{var}(y) = SST = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \overline{y}^2$$

Méthode des moindres carrés

Le choix des paramètres a et b est fait de façon à maximiser la variance reproduite (ou expliquée) par le modèle. On note $\hat{a}$ et $\hat{b}$ ce choix.

DÉFINITION : Variance du modèle (*Sum of Squares Explained, SSE*) :

$$SSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$$

DÉFINITION : Variance résiduelle, ou Résidus (*Sum of Squares Residuals, SSR*) :

$$SSR = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

THÉORÈME : La variance totale est la somme de la variance expliquée et de la variance résiduelle.

$$SST = SSE + SSR$$

Ainsi, on peut aussi dire que **les paramètres a et b sont choisis de façon à minimiser la variance résiduelle** :

$$SSR = \mathcal{S}(a,b)$$

THÉORÈME : Une fonction de deux variables $\mathcal{S}(a,b)$ est minimale en $(\hat{a},\hat{b})$ si :

- d'une part elle est minimale en $(\hat{a},\hat{b})$ vis à vis de a à b fixé ($b = \hat{b}$) ;
- d'autre part elle est minimale en $(\hat{a},\hat{b})$ vis à vis de b à a fixé ($a = \hat{a}$).

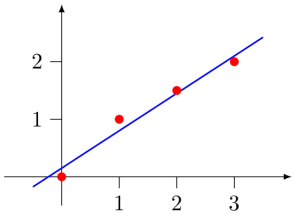
Cela se traduit (si la fonction est dérivable) par un changement de signe de chaque *dérivée partielle* au minimum :

$$\frac{\partial \text{S}}{\partial a}(\hat{a},\hat{b}) = 0 \quad \text{et} \quad \frac{\partial \text{S}}{\partial b}(\hat{a},\hat{b}) = 0\,,$$

ce qui fourni un système linéaire[4] de deux équations à deux inconnues.

Exemple : On cherche à réaliser une régression linéaire sur les 4 points {(0,0), (1,1), (2,1.5), (3,2)} avec la fonction d'équation $y = ax + b$.

- Écrire la fonctionnelle à minimiser et montrer qu'elle se met sous la forme : $\chi^2(a,b) = 14a^2 - 20a + 12ab - 9b + 4b^2 + 7,25$
- Écrire la différenciation de cette fonctionnelle en dérivant par rapport à chaque paramètre.
- En déduire le système à résoudre.
- Déterminer les valeurs de a et b.



DÉFINITION : **Covariance** des observations :

$$\text{cov}(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}).(y_i - \overline{y}) = \frac{1}{n}\sum_{i=1}^{n} x_i.y_i - \overline{x}.\overline{y}$$

THÉORÈME : Les estimateurs $\hat{a}$ et $\hat{b}$ des paramètres a et b obtenus par la méthode des moindres carrés sont [a]

$$\hat{a} = \frac{\text{cov}(x,y)}{\text{var}(x)} \quad \text{et} \quad \hat{b} = \overline{y} - \hat{a}\,\overline{x}\,.$$

a. Démonstration en annexe.

III. Pertinence de l'approximation obtenue

Selon le contexte, le modèle affine obtenu peut s'avérer une bonne approximation ou au contraire une piètre. Cela dépend de la dispersion des mesures (écart-type), mais aussi de la pertinence du choix du modèle par rapport aux grandeurs observées.

Un moyen simple de quantifier la performance de l'ajustement, notamment pour le comparer à un autre modèle, est d'indiquer la part de variance expliquée par le modèle par rapport à la variance totale :

3. L'*écart-type*, qui s'obtient simplement via $\sigma_y = \sqrt{\text{var}(y)}$, est une estimation de la distance typique entre les mesures et la moyenne.

4. car la fonction S est quadratique.

DÉFINITION : **Coefficient de détermination R^2 :**

$$\boxed{R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}} \quad \in [0,1]\,.$$

On utilise aussi couramment le *coefficient de corrélation*, c'est-à-dire la covariance normalisée entre les données $\{x_i\}$ et $\{y_i\}$, qui nous indique si les deux variables « varient bien ensemble ».

DÉFINITION : **Coefficient de corrélation r :**

$$r = \mathrm{corr}(x,y) = \frac{\mathrm{cov}(x,y)}{\sqrt{\mathrm{var}(x).\mathrm{var}(y)}} \quad \in [0,1]\,.$$

THÉORÈME : Dans le cas du modèle de régression affine par moindres carrés, le coefficient de détermination et le coefficient de corrélation sont liés par

$$\boxed{R^2 = r^2}$$

On utilise donc soit l'un soit l'autre, et **plus ce coefficient est proche de 1 plus le modèle peut être considéré pertinent.**

Toutefois on ne dispose pas de valeur universelle de r (ou R^2 à partir de laquelle considérer que le modèle est valide. Dans un domaine où la mesure est précise et les phénomènes stables, on pourra estimer que les données sont fortement corrélées si $|r| \geq 0,95$. Dans des domaines où la mesure est moins précise (notamment dans les sciences humaines), on se contentera parfois de $|r| \geq 0,75$ (soit $R^2 \geq 0,56$!).

D'autre part et plus fondamentalement, il est essentiel de comprendre que le R^2 ou le r sont bien loin de résumer tout ce qui est important à savoir sur l'ajustement linéaire. En particulier **une représentation graphique est toujours indispensable pour juger de la pertinence du modèle**, et notamment déceler des éléments propres à l'invalider totalement : dérive, points aberrants, erreur systématique, non-linéarité flagrante... En témoigne la figure ci-contre qui compare 4 ensembles de données de même taille et de moyennes, variances et corrélations identiques !

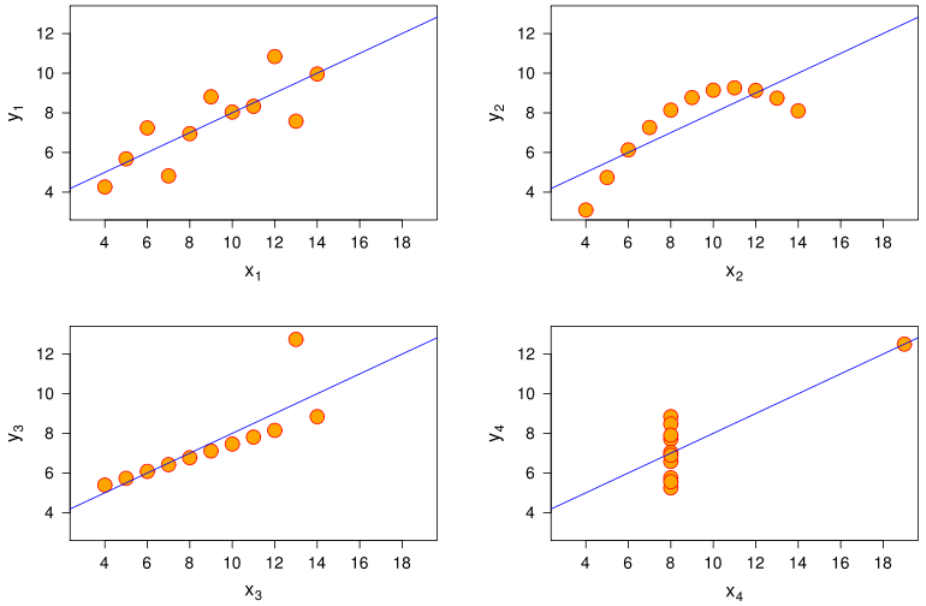


FIGURE 1 – Quartet d'Anscombe
(GPL, https://commons.wikimedia.org/w/index.php?curid=863306).

IV. Annexe - Calcul des estimateurs $\hat{a}$ et $\hat{b}$

Démonstration :

$$\mathcal{S}(a,b) = \frac{1}{n}\sum_{i=1}^{n}(y_i - ax_i - b)^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 + a^2x_i^2 + b^2 - 2ax_iy_i - 2by_i + 2abx_i$$

$$\begin{cases} \frac{\partial \mathcal{S}}{\partial a}(\hat{a},\hat{b}) &= 0 \\ \frac{\partial \mathcal{S}}{\partial b}(\hat{a},\hat{b}) &= 0 \end{cases} \Leftrightarrow \begin{cases} \left(\sum_{i=1}^{n} x_i^2\right)\hat{a} + \overline{x}\,\hat{b} &= \frac{1}{n}\sum_{i=1}^{n} x_iy_i \\ \overline{x}\,\hat{a} + \hat{b} &= \overline{y} \end{cases}$$

$$\Leftrightarrow \begin{cases} \left(\sum_{i=1}^{n} x_i^2 - \overline{x}^2\right)\hat{a} &= \frac{1}{n}\sum_{i=1}^{n} x_iy_i - \overline{x}\,\overline{y} \\ \overline{x}\,\hat{a} + \hat{b} &= \overline{y} \end{cases} \Leftrightarrow \begin{cases} \mathrm{var}(x)\,\hat{a} &= \mathrm{cov}(x,y) \\ \overline{x}\,\hat{a} + \hat{b} &= \overline{y} \end{cases}$$