Fiches – numéro 4

Régression linéaire

Sommaire

Capacités exigibles

- ▢ Analyser les résultats obtenus à l'aide d'une procédure de validation : analyse graphique intégrant les barres d'incertitude ou analyse des écarts normalisés.
- ▢ Utiliser un logiciel de régression linéaire afin d'obtenir les valeurs des paramètres du modèle.

L'essentiel

Définitions

Erreurs communes

Points importants

Applications

I Principe

La pratique de la science passe par deux pivots inséparables : la *théorie* et la *mesure*. La première pose un cadre de travail pour aboutir à des conclusions sur le fonctionnement du monde physique sous la forme de relations mathématiques, et la seconde relève des valeurs expérimentales du monde physique pour en archiver les propriétés effectives.

L'une et l'autre se nourrissent conjointement : une théorie physique pose une loi et la mesure permet de la tester, ou un relevé expérimental particulier permet de mettre en évidence un point sombre des cadres théoriques actuels. Il reste que les sciences physiques ont pour unique objectif de rendre compte du réel ; ainsi, si l'expérience est bien contrôlée, **c'est toujours la mesure qui clôt la discussion**.

Dans ce cadre, il est commun de disposer de deux jeux de données réelles représentant des **grandeurs reliées** entre elles par une loi physique (disons x et y), et normalement compatibles avec une **relation linéaire** mathématique :

$$\boxed{y = ax + b}$$

avec a et b des coefficients propres à la relation :

- ◇ a le *coefficient directeur* ;
- ◇ b l'*ordonnée à l'origine*.

La régression linéaire[1] vise à **tester l'accord entre mesure et théorie**, et le cas échéant à **estimer les coefficients** et leurs incertitudes. Elle repose sur la détermination de la meilleure droite qui serait susceptible de représenter le nuage de points donné en deux dimensions par x et y, en l'occurrence en minimisant la somme des distances au carré entre chacun des points et la droite. Cet **écart aux points mesurés** est caractérisé par un coefficient de corrélation, noté r, ou de régression, noté r^2, et compris entre 0 (nul) et 1 (parfait).

Définition F4.1 : Régression linéaire

Une régression linéaire est une **opération mathématique** entre des points de mesure x_i et y_i qui consiste à trouver les meilleurs coefficients a et b tels que les valeurs $ax_i + b$ soient les plus proches en moyenne des points de mesures y_i.

Cette régression peut être faite à la main, mais la détermination des coefficients ne sera pas rigoureuse. Nous allons ici détailler le processus pour une régression *via* un logiciel.

II Pertinence du modèle

Une fois la régression effectuée, il faut juger de sa validité. En effet, **une régression linéaire va toujours « fonctionner »**, mais ceci n'assure en rien qu'elle est bonne.

Attention F4.1 : Validité d'une régression linéaire

La seule façon valable de conclure à la validité d'une régression linéaire est d'observer l'alignement des points avec la droite de régression.

Une erreur des plus basiques est de ne considérer que l'accord numérique *via* la valeur de r ou r^2. S'il est *nécessaire* d'avoir et d'écrire

$$|r| > 0{,}999\text{x} \quad \text{ou} \quad r^2 > 0{,}99\text{x}$$

où x représente le premier chiffre autre qu'un 9, il est loin d'être *suffisant* et c'est la vérification visuelle qui prévaut.

Important F4.1 : Ma régression est-elle valide ?

Le modèle sera validé sous deux conditions :

1) les points de mesure sont **bien alignés** entre eux (pas de courbure visible à l'œil) ;
2) la **droite** de régression passe le **plus proche de tous les points** possibles, en **incluant leurs incertitudes-type**.

Application F4.1 : Quel modèle est validé ?

On présente ci-après des données et leur régression associée. Pour chaque modèle, dire si la régression est validée ou non, et expliquer pourquoi.

1. Il faudrait en toute rigueur parler de régression *affine* et non linéaire.

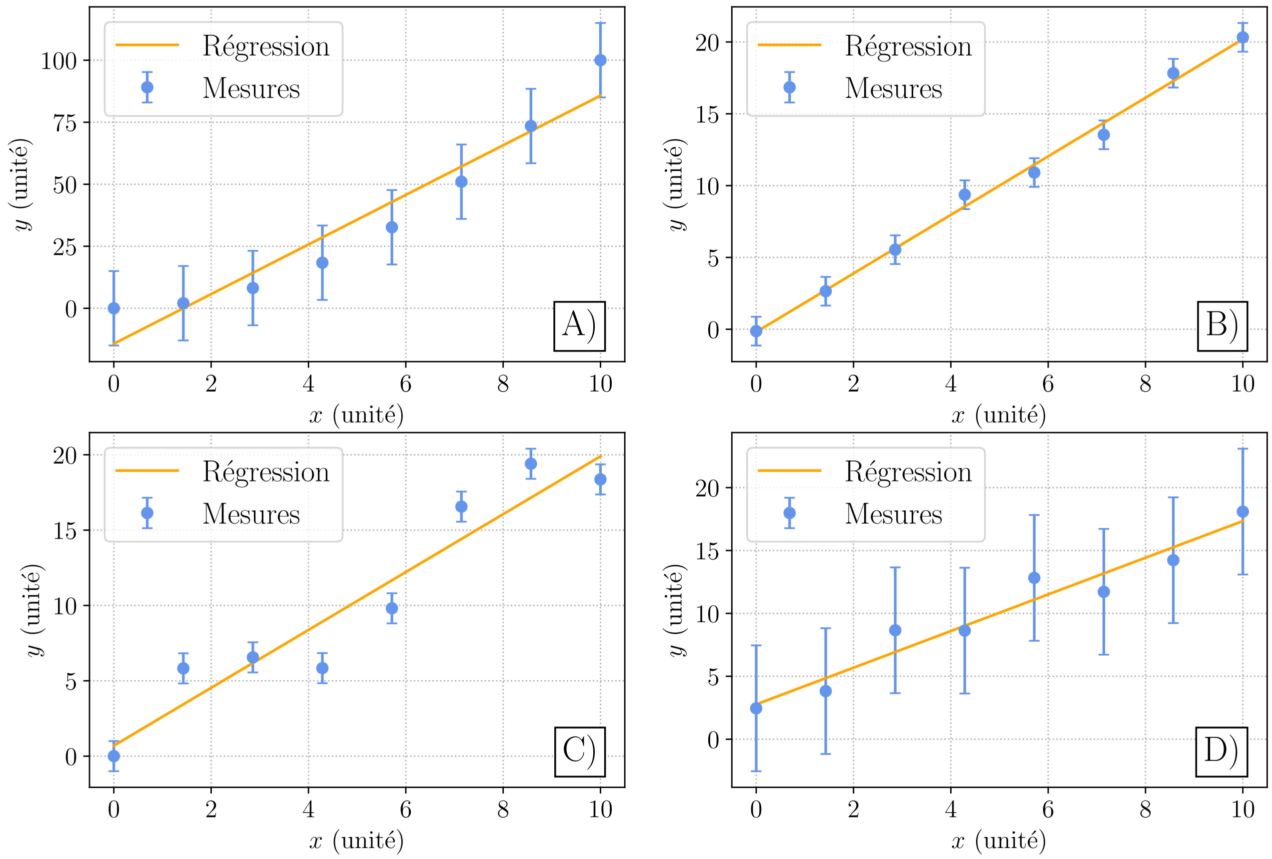

Figure F4.1 – Exemples de régressions linéaires.

A) **Non validée** : points ont une tendance courbée (même si droite passe par les incertitudes).

B) **Validée** : points répartis aléatoirement autour d'une droite, et la droite passe par les incertitudes.

C) **Non validée** : points répartis avec une fonction sinus ; droite est très éloignée de plusieurs points.

D) **Validée mais douteux** : pas de tendance particulière des points et droite passe par les incertitudes, mais elles sont probablement surestimées. Conclusion peu fiable.

III Autres cas

III/A Cas strictement linéaire

Si le modèle recherché implique $b = 0$, c'est-à-dire

$$y = ax$$

alors on cherche uniquement à estimer a. On peut donc calculer N valeurs de $a_i = y_i/x_i$, puis réaliser un traitement statistique de type A sur ces valeurs pour avoir $a = \bar{a} + u(\bar{a})$.

III/B Cas non linéaires

Dans de nombreux cas, la loi à vérifier n'est ni affine ni linéaire. C'est le cas de la loi de Snell-Descartes sur la réfraction entre deux milieux d'indice optique n_1 et n_2 :

$$n_1 \sin(i_1) = n_2 \sin(i_2)$$

Avec des mesures de i_1 et de i_2, il n'est pas possible d'obtenir directement une droite. Il faut dans ce cas tracer :

$$y = ax$$

$$\sin(i_1) \qquad n_2/n_1 \qquad \sin(i_2)$$