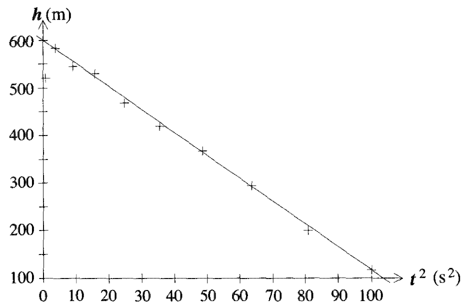
Introduction (3) — Régression linéaire

L'exploitation de résultats expérimentaux, que ce soit en physique ou en chimie nécessite très souvent l'emploi de la **régression linéaire**. Cet outil permet notamment de valider un modèle théorique grâce aux données expérimentales.
Puisque la droite est la fonction mathématique la plus simple à manipuler, on cherchera toujours à tracer graphiquement une grandeur Y en fonction d'une grandeur X de telle sorte que le graphe $Y = f(X)$ soit une droite — donc d'équation $Y = aX + B$.

Exemple : La chute libre d'un corps sans vitesse initiale se traduit théoriquement par une variation de l'altitude h avec le temps t de la forme : $h = a.t^2 + b$ (a et b sont des constantes liées à l'accélération de pesanteur g et à l'altitude initiale). Un dispositif expérimental permet la mesure de l'altitude h en fonction du temps t écoulé. On obtient le tableau de valeurs suivant :

t (s)	0	1	2	3	4	5	6	7	8	9	10
h (m)	600	520	582	545	468	420	368	295	200	118	

La méthode graphique la plus simple pour vérifier que les données expérimentales corroborent le modèle est de tracer la courbe $h = f(t^2)$. Si le modèle théorique est valide, les points expérimentaux doivent s'aligner sur le graphe. On peut alors déterminer « manuellement » la droite qui passe « au plus près » de l'ensemble des points expérimentaux.
→ Cf. graphe ci-contre.



On voit bien sur l'exemple proposé que certains points peuvent être aberrants ($t = 1\ s$, $h = 520\ m$).
Il peut s'agir d'un dysfonctionnement ponctuel d'un appareil de mesure, ou, plus fréquemment, d'un mauvais report de valeur de la part de l'expérimentateur... Il convient alors de ne plus tenir compte de ce(s) point(s) pour la suite du traitement.[1]
On déduit la valeur de a du coefficient directeur de la droite, et celle de b de son ordonnée à l'origine. Pour l'exemple donné, on trouve $a \simeq -5\ m.s^{-2}$ et $b \simeq 600m$.

Toute cette étude « à la main » est théoriquement nécessaire, notamment pour repérer les points aberrants[2], mais un complément **DOIT** être effectué à l'aide d'une calculatrice (ou d'un ordinateur).
Il manque en effet un paramètre important à l'étude, qui permet de quantifier l'alignement plus ou moins bon des points expérimentaux. Quelle que soit la calculatrice utilisée, le principe reste le même :

> Il suffit d'éditer un tableau de données (en ayant pris la précaution d'ôter les points aberrants...), puis de demander le calcul de la régression linéaire. Si la calculatrice est bien configurée, elle donne immédiatement les valeurs de a et b (coefficient directeur et ordonnée à l'origine), ainsi que le coefficient de corrélation linéaire r et son carré r^2.

▮ Propriété : Plus une corrélation (ou régression) linéaire est convenable (i.e. plus les points sont alignés), plus le coefficient de corrélation r est proche de 1. Par habitude, on reporte sur la copie la valeur de r^2.

1. Certains verront, à tort, une forme de « triche » dans le fait de ne pas tenir compte du ou des points aberrants. C'est pourtant le bon sens scientifique qui pilote cette mise à l'écart.
2.

Rq : En physique ou en chimie, la valeur de r^2 doit être comprise entre $0,99$ et 1 pour pouvoir considérer une bonne adéquation entre les points expérimentaux et la droite d'extrapolation.

Dans le cas de l'exemple précédent, en ôtant le point aberrant, on obtient les paramètres suivants pour la corrélation linéaire :

$$a = -4,81\ m.s^{-2} \quad b = 598\ m \quad r^2 = 0,998$$

La corrélation est donc correcte, mais pas exceptionnelle.

Attention : La valeur de r^2 se donne avec tous les chiffres « 9 » après la virgule ainsi que le premier chiffre différent de « 9 » sans arrondi, même s'il faut mettre pour cela un grand nombre de chiffres significatifs !
Exemples : $0,990$ ou $0,9999996$
Contre-exemples : $0,999$ ou $0,9999564$.

Avec la **TI-89** :
→ Éditeur de données
→ 3 : Nouveau
→ Entrer les données X dans la colonne c1 et les données Y dans la colonne c2
→ Touche F5 (Calculs)
→ Sélectionner Type de calculs : 5 : RegLin ; Enter
→ Indiquer c1 pour X et c2 pour Y ; Enter ; Enter
→ Une fenêtre VARS STAT s'ouvre et donne a, b, r (corr) et r^2 (R^2)

Avec la **CASIO graph65** :
→ Menu Stat, Exe
→ Entrer les données X dans list 1 et les données Y dans list 2
→ Faire CALC REG X

Avec la **TI84$^+$** et la **TI83$^+$** :
→ Aller dans le menu STAT, puis EDIT (L1,L2)
→ Ensuite aller dans STAT, puis CALC (le 2$^{\text{ème}}$ onglet) et enfin choisir 4 :LinReg(ax+b) ou 8 :LinReg(a+bx).
Compléments :
• en ligne : http ://www.ti83plus.online.fr/articles.php ?id=20
• Pour afficher le coefficient de corrélation avec les calculatrices **TI83$^+$**, il faut aller dans Catalog, puis, DiagnosticOn.
Autre possibilité : une fois la régression faite, faire :
VARS > statistics > (eqn ?) > r
dans le menu Vars, toutes les variables systèmes sont accessibles.

- **Une erreur classique** et fréquente est d'inverser la liste des abscisses et celle des ordonnées. Il faut bien maitriser sa machine. Pour voir si on se trompe ou pas, il suffit de rentrer les couples de valeurs $(x = 0, y = 0)$ et $(x = 1, y = 2)$. Si on trouve une pente 2, on a bon, si on trouve une pente $0,5$ c'est qu'on a inversé abscisses et ordonnées.
- **Attention aux conventions :** Les Français écrivent les équations de droite $ax + b$, mais les anglo-saxons préfèrent $a + bx$. Bien vérifier ce que a et b signifient pour votre calculatrice.