Modélisation de données expérimentales

En TP, vous aurez parfois à traiter un relevé de n valeurs expérimentales (x_i, y_i) et leurs incertitudes (Δx_i, Δy_i) associées (les erreurs systématiques ne sont pas traitées ici).

On attend de vous que vous cherchiez à faire 2 choses :

1 – valider un modèle, c'est-à-dire répondre à la question suivante : la loi de variation $y(x)$ observée expérimentalement peut-elle, compte tenu des incertitudes, se représenter par la fonction modèle f ?

2 – trouver les k coefficients de la fonction modèle f et leur incertitude.

Par exemple, si un modèle théorique prévoit une variation $y = f(x)$ linéaire, vous essayerez de valider le modèle $f(x) = a\,x$ et de trouver le coefficient $a \pm \Delta a$.

La réponse à ces questions dépend bien sûr de la dispersion des points expérimentaux autour de la fonction modèle, mais aussi de la valeur des incertitudes de mesure : si on réalise une mesure très précise, il sera plus difficile de trouver une fonction modèle représentant correctement les points expérimentaux, mais une fois la fonction trouvée, l'incertitude sur ses paramètres sera faible.

Si les incertitudes expérimentales sont connues, les logiciels de traitement de données (*Synchronie*, *Kaleidagraph*, …) utilisent en général la méthode du « χ^2 » qui consiste à chercher les k coefficients (a, b, …) de la fonction f choisie qui minimisent la quantité :

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - f(x_i))^2}{\sigma_i^2}$$

où σ_i est la variance liée à l'incertitude sur le point (x_i, y_i). On voit qu'il s'agit d'une somme pondérée : plus un point expérimental est précis, plus sont influence est importante.

La minimisation de cette fonction par le logiciel conduit à l'affichage des résultats suivants :

- paramètres optimum de la fonction f : $a \pm \Delta a$, $b \pm \Delta b$, …
- validation du choix de la fonction f : χ^2 ou $C_m = \chi^2/(n\text{-}k)$

C_m est appelé critère d'optimisation de la modélisation. On peut considérer que le choix de la fonction modèle f est correct si :

$$\boxed{\chi^2 \approx n\text{-}k, \text{ ou } C_m \approx 1 \quad (\text{l'erreur sur } \chi^2 \text{ est } \Delta\chi^2 = \sqrt{2(n-k)})}$$

Une valeur trop importante du χ^2 signifie que le modèle n'est pas valide. Une valeur trop faible signifie que les incertitudes ont été surévaluées.

Remarque : Les logiciels donnent aussi souvent le coefficient de régression linéaire R ou l'écart quadratique moyen E (moyenne du carré des écarts entre les points expérimentaux et la fonction f). Ces deux coefficients ne dépendent que de la dispersion des points expérimentaux autour de f, pas de leur précision. Ils ne permettent donc pas de valider le modèle ni d'obtenir la précision sur les coefficients de la fonction f.

- Utilisation du logiciel Synchronie :

Si on utilise une fonction prédéfinie, ou si on ne précise pas les incertitudes Δx et Δy, le logiciel affiche seulement les valeurs des coefficients a, b, … et l'écart quadratique E.

Pour obtenir la valeur des incertitudes sur les coefficients Δa, Δb, … ainsi que le critère d'optimisation C_m, il faut choisir une fonction utilisateur (sélectionner : *autre fonction*) et entrer les valeurs des incertitudes Δx et Δy (identiques pour tous les points).

Le logiciel réclame alors une estimation des paramètres a, b, … pour l'aider dans sa modélisation. Si les estimations de ces paramètres ne sont pas assez précises, il est possible que la modélisation échoue. Seuls les paramètres *actifs* sont affinés. Si, alors que tous les paramètres sont *actifs*, la modélisation échoue, recommencer en rendant *actifs* les paramètres un par un.

- Utilisation du logiciel Kaleidagraph :

Ici encore, si on utilise une fonction prédéfinie (ex. : *Curve fit – Linear*), le logiciel affiche seulement les valeurs des coefficients a, b, … et R. Si on veut obtenir les incertitudes Δa, Δb, … et le χ^2, il faut définir soi-même la fonction modèle (passer par : *Curve fit – General – Fit 1 – Define ...*).

Pour préciser l'incertitude σ_i attachée à chaque point de mesure, sélectionner *Weight Data*. Le logiciel demande alors de préciser quelle est la colonne contenant les valeurs de σ_i (attention : si *Weight Data* n'est pas sélectionné, le calcul est fait avec une valeur constante arbitraire $\sigma_i = 1$). Si l'incertitude sur x est négligeable, on prend $\sigma_i = \Delta y_i$, qui peut dépendre du point i considéré.

Attention : l'affichage des barres d'erreur sur un graphique (*Plot – Error bars* …) est indépendante de la modélisation.

- Utilisation du logiciel Excel :

La modélisation de base d'Excel n'affiche que les coefficients a, b, … et le coefficient de régression linéaire R, insuffisant pour nous.

Exemple :

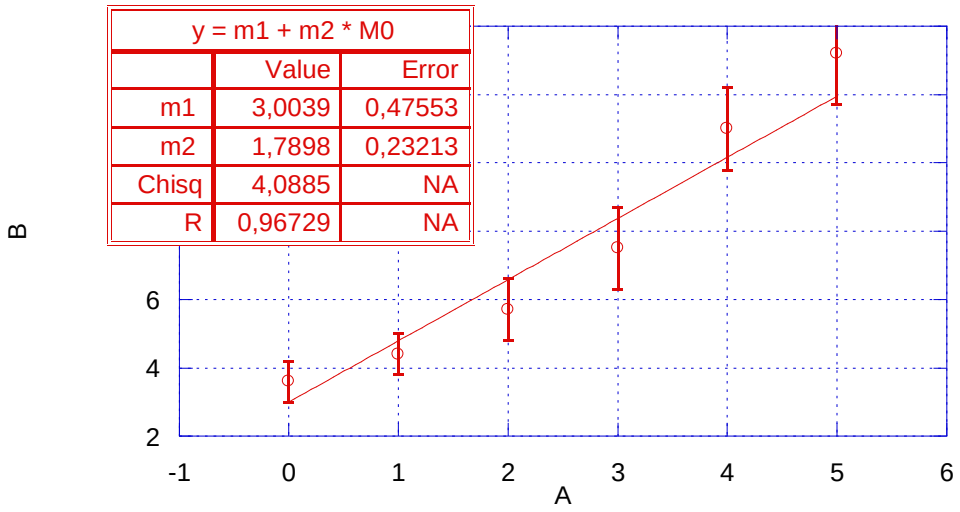
Les quatre courbes suivantes ont été tracées sur *Kaleidagraph*, avec $n = 6$ points de mesure, identiques dans tous les cas. Les barres d'erreur tracées correspondent aux poids σ_i utilisés pour la modélisation : elles sont trois fois plus faibles dans les cas 2 et 3 que dans le cas 1 et 4.

Cas 1 : Grandes barres d'erreur, modélisation par une droite.

On trouve $\chi^2 = 4{,}0885$ au lieu de la valeur idéale 6 - 2 = 4 (avec une erreur ≈ 2,8).

La modélisation est donc bonne. Par contre, on voit que les coefficients de la droite sont donnés avec une précision assez médiocre (par exemple, la pente est donnée avec une précision de 13 %).

1 - incertitudes importantes, fit linéaire



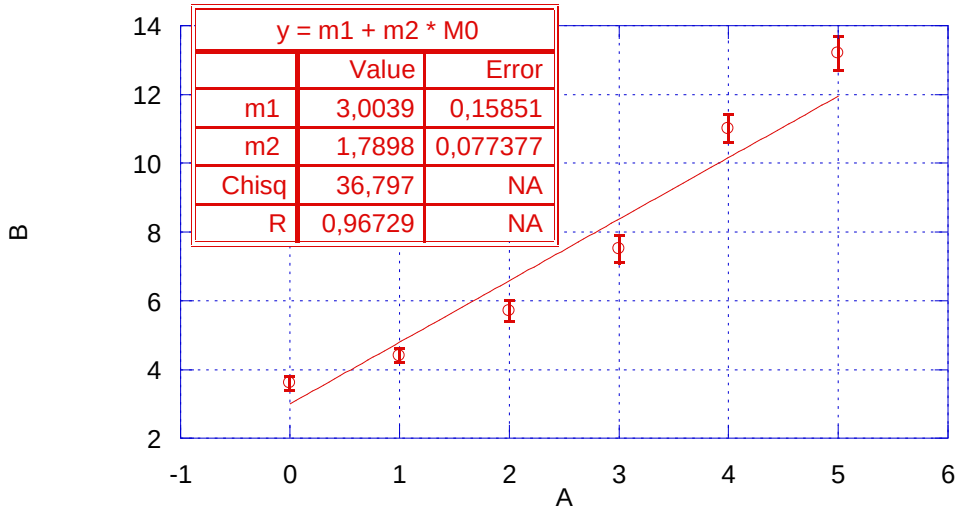
***Cas 2* : Faibles barres d'erreur, modélisation par une droite.**

On trouve $\chi^2 = 36{,}797$ au lieu de la valeur idéale 6 - 2 = 4 (avec une erreur ≈ 2,8).

La modélisation est donc maintenant mauvaise (une mesure précise est plus difficile à modéliser), bien que les coefficients de la droite soient donnés avec une assez bonne précision (par exemple, la pente est maintenant donnée avec une précision de 4 %). On voit dans cet exemple que pour valider un modèle, il ne faut pas se fier à la précision des coefficients de la modélisation : seule la valeur du χ^2 (ou du C_m) peut valider le choix de la fonction modèle.

Remarquez aussi que le coefficient de régression R est le même dans les cas 1 et 2 (il est indépendant des incertitudes) : ce coefficient ne peut pas non plus renseigner sur la validité de la modélisation.

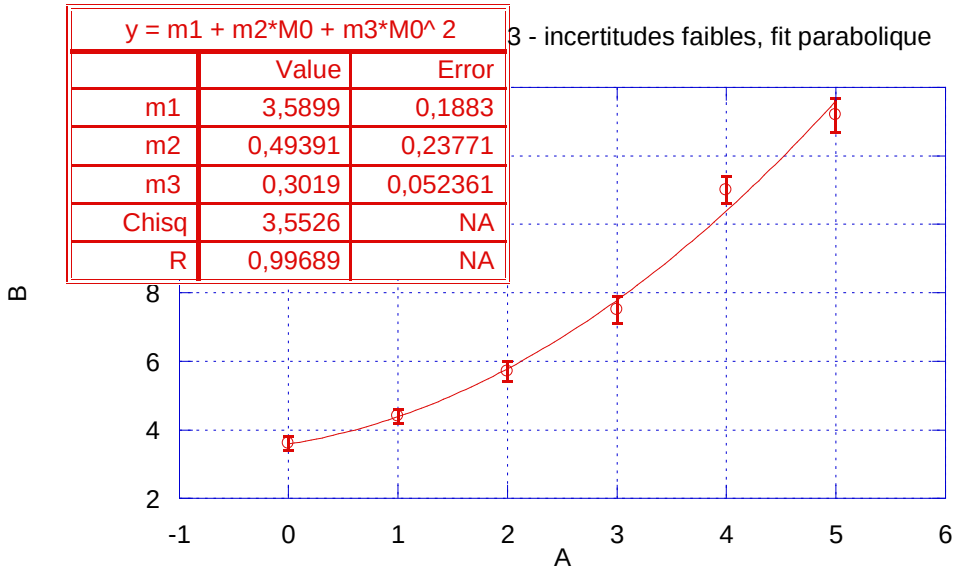
2 - incertitudes faibles, fit linéaire



<u>Cas 3</u> : Faibles barres d'erreur, modélisation par une parabole.

On trouve $\chi^2 = 3{,}5526$ au lieu de la valeur idéale 6 - 3 = 3 (avec une erreur ≈ 2,4).

La modélisation par une parabole est donc bonne : on peut dire que les résultats expérimentaux sont bien représentés par une loi parabolique, <u>compte tenu des incertitudes expérimentales</u> (dans le cas 1, les incertitudes étant plus importantes, une loi linéaire suffisait).



<u>Cas 4</u> : Grandes barres d'erreur, modélisation par une parabole.

On trouve $\chi^2 = 0{,}39473$ au lieu de la valeur idéale 6 - 3 = 3 (avec une erreur ≈ 2,4).

Ici, la valeur du χ^2 est plutôt faible, ce qui signifie que la modélisation est valide mais peu sensible : la parabole convient, mais pas mieux que la droite ! D'ailleurs, on peut noter que l'incertitude sur les paramètres de la modélisation est très grande.