La régression linéaire et ses conditions d'application

par R. JOURNEAUX
GHDSO/LIREST
Université Paris XI, 91400 Orsay

Dans un article récent publié dans le Bulletin [1] sont abordées les techniques de régression et particulièrement le problème posé quand les grandeurs expérimentales mises en jeu sont affectées d'incertitudes. Le présent article propose une approche qui, tout en reposant sur des justifications théoriques voisines, fait plus appel au sens physique et permet une approche moins axiomatique pour les étudiants ou les enseignants. Par ailleurs, des compléments sont donnés au sujet des incertitudes sur les coefficients donnés par la régression.

1. LE PROBLÈME DE LA RÉGRESSION

Dans les sciences expérimentales, il est très fréquent de se trouver devant la situation suivante : quand on fait varier une grandeur, d'autres grandeurs varient, et il est intéressant de savoir s'il est légitime de traduire ces variations par une relation fonctionnelle.

Deux cas peuvent d'ailleurs se présenter. Le premier est celui où la relation fonctionnelle est inconnue. Il s'agit alors de faire des essais et de présenter le relation qui paraît la plus satisfaisante. Il est donc important de disposer de critère de jugement qui justifient le choix et permettent à la communauté scientifique d'apprécier le résultat publié. La loi empirique ainsi trouvée est assortie d'un taux de confiance. Le second cas se présente quand un modèle théorique est censé régir les variations mesurées. Il s'agit alors de confronter ce modèle avec les résultats expérimentaux et de se donner ici encore des critères de jugement. La réponse consiste alors à dire si le modèle envisagé est compatible avec les résultats, avec toujours un certain degré de confiance. Une variante consiste à tester plusieurs modèles concurrents et à choisir celui qui est le plus en accord avec les résultats. Dans les deux cas enfin, le traitement des données conduit à la détermination de

paramètres du modèles qui sont obtenus eux aussi avec un certain intervalle de confiance.

1.1. **Les critères utilisés dans la régression**

En abordant les problèmes de régression, on est amené à faire appel à un vocabulaire flou : quelle est la «meilleure loi», comment faire passer «au mieux» une courbe par un ensemble de points, quelle est la «meilleure représentation» d'un ensemble de résultats ? Ces activités sont en effet appliquées à des résultats qui sont affectés d'incertitudes ; il n'est donc pas question de trouver des lois exactes mais bien de trouver des lois avec un certain degré de confiance et des paramètres avec un certain intervalle de confiance associé. On retrouve le vocabulaire typique des traitements statistiques.

La nécessité d'un critère

Le point de départ de toute technique de régression consiste à se donner un critère, c'est-à-dire une règle du jeu permettant de répondre quantitativement à la question floue du départ. Ce critère est essentiel : son choix conditionne toute la suite des opérations, et peut être éventuellement remis en cause.

Dans le cas général, la relation mathématique testée est de la forme $F(u,v) = 0$ en se restreignant à deux variables. S'il est possible de la mettre sous la forme $g(v) = h(u)$, il est alors possible de raissonner sur une fonction $y = f(x)$ par un changement éventuel de variable sur v et/ou sur u. Le cas général est souvent résolu par des méthodes numériques complexes et itératives. Nous restreindrons la suite au cas particulier $y = f(x)$.

Soit une fonction $f(x)$ comportant m paramètre tels que P_k. On cherche la valeur des m paramètres en appliquant le critère de régression choisi. On peut tout de suite remarquer que chercher la fonction $f(x)$ dissymétrise le problème ; la variable x apparaît comme la variable «cause» et y la variable «effet» ; on dit alors qu'on étudie la regression de y par rapport à x, mais ce choix est arbitraire et peut-être inversé.

Les choix possibles

La grandeur de départ pour la recherche du critère est constituée par la différence entre la valeur expérimentale y_i et la valeur $f(x_i)$ donnée

par la fonction cherchée, i variant de 1 à n (n nombre de couples expérimentaux x_i, y_i). On peut par exemple se demander si toutes ces différences peuvent être nulles. On obtient alors n relations avec m inconnues qui n'a en général de solution unique que si m = n. Cette technique particulière d'ajustement est intéressante si le nombre de points n'est pas trop élevé et peut servir pour un étalonnage ; elle fait l'objet de développements dans les ouvrages spécialisés et ne sera pas aborbée ici.

Si la relation cherchée est représentée par une droite, un critère évident consiste à minimiser la somme des distances des points à la droite. Ce critère n'est pas utilisé pour deux raisons. La première tient au fait que le résultat obtenu dépend des échelles. La seconde résulte de la complexité des calculs mathématiques mis en jeu. On dissymétrise le problème en prenant la distance «verticale» entre la droite et les points.

Dans le cas général, le critère peut donc être recherché en raisonnant sur la somme des carrés des écarts $S = \Sigma\,(f(x_i) - y_i)^2$. Ce critère classique est utilisé depuis longtemps et il conduit, dans le cas de la régression linéaire, à la droite des «moindres carrés» bien connue. C'est d'ailleurs le critère utilisé dans les calculatrices.

1.2. **Les limites du critère quadratique classique**

Depuis quelques années, ce critère est remis en cause car il accorde la même importance à tous les points expérimentaux, ce qui est contraire au bon sens : il paraît important de privilégier les points pour lesquels la confiance est maximum, c'est-à-dire les points qui sont obtenus avec la meilleure précision. On arrive alors à la notion de **pondération** qui va consister à attribuer un «poids» à chacun des termes de la somme, poids qui varie dans le même sens que la précision du point envisagé [2, 3, 4]. C'est la même opération que celle qui consiste à attribuer un coefficient pour la détermination d'un barycentre, par exemple pour trouver un centre de masse.

Le problème qui reste à résoudre est de trouver ce poids à partir des précisions avec lesquelles sont connues les grandeurs x_i et y_i, c'est-à-dire à partir des variances $s_{x_i}^2$ et $s_{y_i}^2$ caractérisant chaque résultat de mesure. Ce poids va être choisi inversement proportionnel à la variance du terme $f(x_i) - y_i$. Si les grandeurs x_i et y_i sont indépendantes, il faut ajouter les variances des termes de la somme soit

$\mathrm{Var}(f(x_i)) + \mathrm{Var}(y_i) = s_i^2 = s_{f_{x_i}}^2 + s_{y_i}^2$ d'où le facteur de pondération $g_i = \dfrac{1}{s_{f_{x_i}}^2 + s_{y_i}^2}$. On retrouve le critère issu du principe du **Maximum de Vraisemblance** évoqué dans la référence (1) justifié ici par des considérations physiques. Si on a fait des changements de variable à partir de u et v, il faut appliquer les relations de propagation des erreurs pour déterminer chaque variance. Le choix de ce poids, outre son caractère logique, fait apparaître chaque terme de la somme comme le rapport du carré d'une grandeur et de sa variance. On reconnait la manière de construire une variable obéissant à une loi statistique classique : la loi de χ^2. Outre son rôle de donner la solution au problème cherché, le critère de régression fait intervenir une grandeur dont la loi de probabilité est connue. On verra qu'on pourra en tirer des renseignements sur la qualité de la régression et donc aller plus loin que la simple détermination des coefficients.

Les m paramètres P_k sont donc tels qu'ils minimisent la somme :

$$\chi^2 = \sum g_i\,[f(x_i) - y_i]^2 \qquad \text{soit} \qquad \frac{\partial \chi^2}{\partial P_k} = 0 \text{ pour k de 1 à m.}$$

On obtient m relations à m inconnues : le problème a donc en général une solution. Mais dans le cas le plus général, ces solutions ne sont pas obtenues de façon littérale, il faut donc les déterminer par voie numérique. En fait, on cherche directement le minimum de χ^2 par des techniques informatiques spéciales.

2. CAS DE LA RÉGRESSION LINÉAIRE

La fonction envisagée est $y = ax + b$. On en déduit $s_i^2 = a^2 s_{x_i}^2 + s_{y_i}^2$ d'où la valeur du poids g_i pour chaque point expérimental. Le critère, après dérivation par rapport à a et b, ne donne pas de relations simples qui permettent de trouver une solution littérale. Seuls des cas particulers vont fournir des résultats simples.

2.1. Cas nº 1

Les x_i sont connus avec une précision très supérieure aux y_i. Les y_i sont caractérisés chacun par la même précision. La quantité s_i^2 est constante et peut donc être sortie de la somme : c'est la méthode des

moindres carrés classique. Il faut noter que l'expression «les x_i sont connus avec une précision très supérieure aux y_i» est ambiguë. C'est en fait **la quantité $a^2 s_{x_i}^2$ qui doit être inférieure à $s_{y_i}^2$**. En effet, si a est grand, une faible incertitude sur x_i entraine une grande incertitude sur la position du point sur la droite qui est presque «verticale». Une estimation initiale de a peut donc seule permettre d'apprécier la validité de l'approximation.

La dérivation de la somme par rapport à a et b conduit à :

$$\sum_{i=1}^{i=N} x_i (y_i - a x_i - b) = 0 \quad (1) \qquad \sum_{i=1}^{i=N} (y_1 - a x_i - b) = 0 \quad (2)$$

Les coefficients a et b sont donnés par les relations suivantes :

$$D = n \sum x_i^2 - (\Sigma x_i)^2$$

$$a = \frac{1}{D} (n \Sigma x_i y_i - \Sigma x_i \Sigma y_i)$$

$$b = \frac{1}{D} (\Sigma y_i \sum x_i^2 - \Sigma x_i \Sigma x_i y_i)$$

On démontre dans ce cas que la droite obtenue passe par le point μ_x , μ_y , μ_x et μ_y étant les valeurs moyennes des x_i et y_i.

2.2. **Cas nº 2**

C'est le cas précédent mais chaque y_i est caractérisé par une variance $s_{y_i}^2$, donc $s_i^2 = s_{y_i}^2$. Un calcul analogue au précédent conduit à :

$$\Delta = \Sigma \frac{1}{s_i^2} \Sigma \frac{x_i^2}{s_i^2} - \left(\Sigma \frac{x_i}{s_i^2} \right)^2$$

$$a = \frac{1}{\Delta} \left(\Sigma \frac{1}{s_i^2} \Sigma \frac{x_i y_i}{s_i^2} - \Sigma \frac{x_i}{s_i^2} \Sigma \frac{y_i}{s_i^2} \right)$$

$$b = \frac{1}{\Delta} \left(\Sigma \frac{x_i^2}{s_i^2} \Sigma \frac{y_i}{s_i^2} - \Sigma \frac{x_i}{s_i^2} \Sigma \frac{x_i y_i}{s_i^2} \right)$$

2.3. **Cas nº 3**

C'est le cas le plus général où $s_i^2 = a^2 s_{x_i}^2 + s_{y_i}^2$ doit être utilisé. Les dérivations par rapport à a et b ne donnent plus un système linéaire. La méthode la plus rapide utilise les itérations suivantes :

– on détermine une estimation de a par la méthode du cas 1, ou on utilise une valeur si on a une idée de la solution par une étude auxiliaire,

– on déduit une estimation des quantité s_i^2 qui sont alors considérées comme des constantes,

– on refait une estimation de a et b par la méthode du cas 2,

– on itère le processus jusqu'à ce que les variations des estimations de a et b soient considérées comme négligeables par rapport aux intervalles de confiance de ces coefficients (voir 3. pour leur estimation).

L'expérience montre que la première itération donne déjà un résultat très acceptable.

3. **PRÉCISIONS SUR LES COEFFICIENTS**

Les coefficients a et b ainsi déterminés doivent être affectés d'un intervalle de confiance. Pour cela, il faut connaître leurs variances compte tenu des précisions des variables x_i et y_i. Dans la suite, on considère que les quantités x_i d'une part, y_i d'autre part sont indépendantes.

3.1. **Variance de a et b**

Dans les cas 1 et 2, seules les grandeurs y_i sont entachées d'incertitudes caractérisées par les variances $s_{y_i}^2$. Par ailleurs, a et b sont des fonctions relativement simples des y_i. Il est donc possible d'appliquer le théorème de propagation des erreurs et on trouve alors :

$$s_a^2 = \Sigma \left(\frac{\partial a}{\partial_{y_i}} \right)^2 s_{y_i}^2 \quad \text{et} \quad s_b^2 = \Sigma \left(\frac{\partial b}{\partial_{y_i}} \right)^2 s_{y_i}^2.$$

On trouve alors les résultats :

$$s_a^2 = \frac{1}{\Delta} \Sigma \frac{1}{s_{y_i}^2} \text{ et } s_b^2 = \frac{1}{\Delta} \Sigma \frac{x_i^2}{s_{y_i}^2}$$

Ces expressions se simplifient dans le cas 1 où tous les $s_{y_i}^2$ sont égaux (on appellera s_{yx}^2 cette valeur commune) pour donner :

$$s_a^2 = \frac{n}{D} s_{yx}^2 \text{ et } s_b^2 = \frac{s_{yx}^2}{D} \Sigma x_i^2$$

Dans le cas général (cas 3), la théorie donne (3, 4) :

$$s_a^2 = \frac{1}{\Delta} \Sigma \frac{1}{s_1^2} \text{ et } s_b^2 = \frac{1}{\Delta} \Sigma \frac{x_i^2}{s_i^2}$$

3.2. **La covariance entre a et b**

Par ailleurs, la covariance entre a et b est importante à considérer car les coefficients sont fortement corrélés par les x_i et y_i. Dans le cas 1 et 2, en appliquant la définition de la covariance $\text{Cov}(a,b) = \Sigma \left(\frac{\partial a}{\partial y_i}\right)\left(\frac{\partial b}{\partial y_i}\right) s_{y_i}^2$, on trouve :

$$\text{Cov}(a,b) = -\frac{1}{\Delta} \Sigma \frac{x_i}{s_{y_i}^2}$$

qui se réduit à $\text{Cov}(a,b) = -\frac{s_{yx}^2}{D} \Sigma x_i$ dans le cas 1.

Dans le cas général (cas 3), on trouve :

$$\text{Cov}(a,b) = -\frac{1}{\Delta} \Sigma \frac{x_i}{s_i^2}$$

3.3. **Retour sur le cas 1**

Les quantités ci-dessus nécessitent la connaissance des variances de toutes les grandeurs expérimentales. Un cas particulier est intéressant, c'est le cas 1 où toutes les variances inconnues $s_{y_i}^2$ sont égales. La seule quantité à déterminer expérimentalement est donc s_{yx}^2. Ceci peut se faire en effectuant plusieurs mesures pour un x donné, en faisant éventuellement varier x pour tester la validiter de l'hypothèse. Si cette hypothèse est jugée valable par une analyse sérieuse des causes et amplitudes des incertitudes de mesure, on peut se dispenser de cette

multiplication des relevés expérimentaux. En effet on démontre que s_{yx}^2 est estimée par la quantité :

$$s_{yx}^2 = \frac{1}{n-2} \Sigma (a\, x_i + b - y_i)^2$$

Cette relation peut facilement se comprendre en revenant au sens des hypothèses. Dire que $s_{y_i}^2$ ne dépend pas de x signifie que les y_i expérimentaux peuvent être considérés comme extraits de populations de même variance mais de valeurs moyennes différentes. Si on fait «glisser» les points par la pensée parallèlement à la droite de régression pour les amener à la même valeur de x_i, on va ébaucher pour ce point la courbe de répartition des y_i. On est donc amené à calculer une variance sur les valeurs $ax_i + b - y_i$ qui représente bien la différence entre la même valeur de y_i et la valeur estimée sur la droite de régression. Il reste à comprendre le sens du terme n – 2, nombre de degrés de liberté pour le problème envisagé. On peut le faire en remarquant que si n = 2, il passe une droite par ces points, donc la somme des carrés est nulle, donc s_{yx}^2 est aussi nulle si le dénominateur est différent de zéro, ce qui signifie qu'on peut obtenir un résultat parfaitement précis avec deux points entachés d'incertitude. La seule façon d'éviter cette absurdité est de mettre n – 2 au dénominateur, d'où indétermination, ce qui est plus conforme à la situation. (On peut reparquer que cette situation est la même que celle qui est rencontrée pour l'estimation de la variance de n mesures et qui conduit à mettre n – 1 au dénominateur). Ces considérations, si elles ne constituent pas des démonstrations, sont utiles pour démythifier auprès des élèves des formules qui peuvent paraître «parachutées» mais que le simple bon sens permet de rendre logiques.

3.4. **Passage à l'intervalle de confiance**

Les variances estimées de a et b permettent d'en déduire les intervalles de confiance de ces coefficients en utilisant le coefficient de Student. Par exemple, la pente inconnue possède $100 - \alpha$ chances de se trouver dans l'intervalle $a \pm t_\alpha s_a$, t_α étant le coefficient de Student à n – 2 degrés de liberté au taux de confiance α choisi.

3.5. **Le cas de l'interpolation**

Quand la régression est effeectuée, on est amené à en tirer la valeur de y_0 à partir d'une valeur de x_0 : c'est l'interpolation si x est dans le

domaine des x_i de départ. La quantité $y_o = ax_0 + b$ est également qualifiée par sa variance s_0 qu'on peut obtenir par propagation des erreurs (on néglige l'incertitude sur x_0) :

$$s_0 = x_0^2 s_a^2 + s_b^2 + 2x_0 \, Cov(a,b)$$

La précision de y_0 dépend donc de x_0. Dans le cas de régression 1, on montre facilement que cette précision est maximum (s_0 minimum) pour $x_0 = \mu_x$ (moyenne des x). Ceci reste approximativement vrai dans le cas général, sauf si les précisions sur les points sont très différentes. Le domaine de confiance de y_0 pour un taux de confiance donné est donc situé entre deux courbes qui s'éloignent de la droite de régression quand on s'écarte de μ_x, ce qui explique en particulier que l'extrapolation est toujours plus imprécise qu'une interpolation (indépendamment du fait que l'extrapolation suppose que le modèle est accepté en dehors du domaine étudié).

4. RÉGRESSION LINÉAIRE ET CALCULATRICE

De plus en plus, les calculatrices possèdent la fonction régression linéaire. **Il faut savoir que les formules utilisées correspondent au cas 1 et ne sont donc utilisables, en toute rigueur, que dans un nombre restreint de situations.** Outre les coefficients a et b, diverses autres données sont accessibles : coefficient de corrélation, Σx, Σx^2, Σxy, Σy^2, s_x , s_y.

Par contre, les quantité D et s_{yx} , nécessaires pour déterminer les variances des coefficients, ne sont pas directement accessibles. Il faut donc les évaluer à partir des données de la calculatrice.

La valeur de $D = n \Sigma x_i^2 - (\Sigma x_i)^2$ est directement calculable.

La quantité s_{yx} peut se calculer également en remarquant que :

$$s_{yx}^2 = \frac{1}{n-2} \sum (y_i - ax_i - b)^2$$

$$= \frac{1}{n-2} \left\{ \sum y_i(y_i - ax_i - b) - a \sum x_i (y_i - ax_i - b) - b \sum (y_i - ax_i - b) \right\}$$

Les deux dernières sommes sont nulles (voir relations 1 et 2 du § 2.), d'où : $s_{yx}{}^2 = \frac{1}{n-2}\left(\sum y_i{}^2 - a \sum x_i\, y_i - b \sum y_i\right)$ directement calculable.

5. LA VALIDITÉ DE LA RÉGRESSION

Effectuer une régression, en déduire des paramètres et leurs intervalles de confiance constituent une activité qui est incomplète pour le physicien. Il faut aussi avoir le moyen de répondre à la question : le modèle choisi est-il acceptable avec un taux de risque d'erreur qu'on est prêt à assumer ? Il peut aussi se formuler de la façon suivante : face à plusieurs modèles concurrents, quel est celui qui est le plus plausible, c'est-à-dire quel est celui qui correspond au risque minimum ?

Les statistiques permettent d'aborder le problème de divers façons. Nous en évoquerons rapidement deux : la corrélation et le χ^2.

5.1. Le coefficient de corrélation linéaire

Nous ne reviendrons pas sur son expression mais sur sa signification. Remarquons d'abord qu'il ne fait pas intervenir la pondération introduite pour la régression, ce qui limite à priori sont utilisation. Par ailleurs, il s'agit du coefficient de corrélation linéaire qui n'est pas applicable pour une fonction f(x) quelconque (il existe d'autres coefficients de corrélation). Le coefficient de corrélation permet de savoir si les variations d'une grandeur y peuvent être attribuées au hasard ou aux variations d'une autre grandeur x. Les tables dont on dispose donnent la probabilité pour que seul le hasard soit responsable des variations observées. En Physique-Chimie, de telles corrélations ne font en général pas de doute et conduisent toujours à des coefficients voisins de 1. Ceci est flagrant dans les exemples de la référence [1]. L'analyse du coefficient de corrélation est donc mal adaptée pour répondre aux questions ci-dessus. Par contre il sera très utile dans certains domaines, (économie, sociologie, psychologie...) où la corrélation n'est pas évidente et se révèle à elle seule une source d'information fondamentale à l'exclusion de toute relation fonctionnelle.

5.2. Le coefficient χ^2

La quantité χ^2 qui a servi de critère à la régression suit une loi de probabilité qui a été tabulée. Compte tenu des points expérimentaux

avec leurs incertitudes et de la relation mathématique choisie, les tables donnent la probabilité pour que ce coefficient dépasse une certaine valeur. Si on a choisi une probabilité, c'est-à-dire un risque d'erreur, le coefficient χ^2 donne une réponse immédiate à la question de l'acceptabilité du modèle. Il donne également la possibilité de trancher entre plusieurs modèles : le plus acceptable est celui qui donne le χ^2 le plus faible.

Il ne faut pas croire que le test du χ^2 dispense d'une réflexion supplémentaire. L'examen de la forme de ce coefficient montre en effet qu'il comporte deux termes :

– le numérateur qui dépend de l'écart des points expérimentaux au modèle choisi,

– le dénominateur qui traduit la qualité des mesures à travers les variances $s_{x_i}^2$ et $s_{y_i}^2$.

Cela signifie que si la précision des mesures est grande ($s_{x_i}^2$ et $s_{y_i}^2$ faibles), il faut que le numérateur soit faible pour que le modèle soit acceptable. Ce résultat est logique : plus les mesures sont précises, plus le modèle doit être bien adapté, sinon les écarts seront flagrants. C'est ce qui se constate quand on cherche à faire passer une droite par des points qu'on a agrémentés, de façon classique, de leur rectangle d'incertitude : si les précision sont bonnes, la relation linéaire peut se révéler peu plausible.

Cela signifie aussi que, devant un χ^2 trop élevé, il faut éviter de conclure trop rapidement à une inadéquation du modèle. Il faut aussi se demander si les incertitudes ont bien été évaluées, en particulier quand elles se fondent sur des données du constructeur pour un composant ou un appareil. Il faut traduire les précisions constructeur en terme de variance (voir BOEN de juillet 1987), ce qui introduit un certain arbitraire compte tenu en particulier de vieillissement des appareils. Enfin il faut s'assurer que les erreurs systématiques ont bien été recensées et corrigées. C'est la situation qui se produit quand il est difficile de faire passer une droite dans les rectangles d'incertitude sans qu'une tendance puisse laisser penser que le modèle est inadapté : il faut alors se demander si les incertitudes ne sont pas sous estimées.

Le test du χ^2 est donc un indicateur de qualité pour tester un modèle mais **il ne faut jamais oublier qu'il est un outil statistique et que l'analyse reste en dernier lieu aux mains du Physicien.**

6. UN EXEMPLE D'APPLICATION

6.1. Les conditions expérimentales

Nous avons mesuré l'impédance d'une portion de circuit composée d'un condensateur de capacité C (environ 0,5 μF) et d'un résistor de résistance R (environ 1000 Ω) en fonction de la fréquence f (entre 150 et 500 Hz). Deux séries de résultats ont été collectées.

La première série à consisté à mesuré V, I et f avec un multimètre pour chaque grandeur. La précision de la mesure repose donc sur les indications du constructeur. L'écart-type attribué à chaque mesure a été choisi arbitrairement égale au tiers de l'incertitude «classique» Δ (les appareils sont récents).

Dans la deuxiéme série, chaque grandeur V et I a été mesurée par 4 multimètres, la fréquence par 5 appareils. Chaque mesure de Z et f donnent donc une moyenne et un écart-type s. Le tableau ci-dessous donne les résultats de mesures pour f, V et I avec l'écart-type s pour chaque série de mesure, d'où la valeur de Z et de son écart-type. Pour la régression, l'écart-type des grandeurs f et Z est pris égal à $\frac{s}{\sqrt{n}}$, avec n = 4 ou 5.

Tableau des résultats pour la deuxième série de mesures.

f(Hz)	149,6	180,6	201,4	257,06	341,8	408	447
s_F(Hz)	0,89	0,89	1,1	1,5	1,6	2	2
V(V)	1,725	1,653	1,648	1,528	1,425	1,491	1,467
s_V(V)	0,0021	0,0019	0,0019	0,0019	0,0022	0,0022	0,0025
I(mA)	0,6550••	0,7318	0,7950	0,8783	0,9690	1,1102	1,1325
s_I(mA)	0,0035	0,0079	0,0035	0,0057	0,0088	0,0071	0,0085

L'impédance du dipôle étudié étant de la forme $Z^2 = R^2 + \frac{1}{C^2 \omega^2}$, on est amené à faire le changement de variable $y = Z^2$ et $x = \frac{1}{f^2}$. Les variances de x et y sont alors calculées en appliquant les théorèmes de propagation des erreurs à partir des variances de f et Z, soit :

$$s_y^2 = 4Z^2 s_Z^2 \text{ et } s_x^2 = \frac{4}{f^6} s_f^2. \text{ avec } s_Z^2 = \frac{1}{I^2} s_V^2 + \frac{V^2}{I^4} s_I^2.$$

6.2. **Les résultats**

a) Avec un seul appareil

– traitement classique

$$a = 1{,}3117\,10^{11} \quad b = 1{,}0171\,10^6 \quad s_a = 4{,}18\,10^8 \quad s_b = 9{,}96410^3$$

ce qui conduit à :

$$R = 1008{,}5\,\Omega \quad s_R = 5\,\Omega \quad C = 0{,}4394\,\mu F \quad s_C = 7\,10^{-4}\,\mu F$$

– traitement rigoureux

$$a = 1{,}3176\,10^{11} \quad b = 1{,}0078\,10^6 \quad s_a = 2{,}32\,10^9 \quad s_b = 2{,}7110^4$$

$\chi^2 = 0{,}51$(la probabilité pour χ^2 d'être supérieure à cette valeur est de 99 %) ce qui conduit à :

$$R = 1004\,\Omega \quad s_R = 13\,\Omega \quad C = 0{,}4385\,\mu F \quad s_C = 4\,10^{-3}\,\mu F$$

b) Avec plusieurs appareils

– traitement classique

$$a = 1{,}3266\,10^{11} \quad b = 1{,}0208\,10^6 \quad s^a = 3{,}96\,10^8 \quad s_b = 9{,}34\,10^3$$

ce qui conduit à :

$$R = 1010{,}3\,\Omega \quad s_R = 4{,}6\,\Omega \quad C = 0{,}4370\,\mu F \quad s_C = 6{,}5\,10^{-4}\,\mu F$$

– traitement rigoureux

$$a = 1{,}3117\,10^{11} \quad b = 1{,}0171\,10^6 \quad s_a = 4{,}18\,10^8 \quad s_b = 9{,}964\,10^3$$

$\chi^2 = 1{,}34$(la probabilité pour χ^2 d'être supérieure à cette valeur est de 93 %) ce qui conduit à :

$$R = 1005{,}3\,\Omega \quad s_R = 5{,}6\,\Omega \quad C = 0{,}4361\,\mu F \quad s_C = 1{,}5\,10^{-3}\,\mu F$$

– cas où seules les incertitudes sur la fréquence sont prises en compte.

Nous avons envisagé le cas où les incertitudes sur la variables x sont les seules retenues (il faut remarquer qu'il serait alors plus logique de faire la régression de x en fonction de y, mais les résultats ne seraient pas très différents). On trouve alors :

$$R = 1004{,}1\,\Omega \quad s_R = 1{,}7\,\Omega \quad C = 0{,}4355\,\mu F \quad s_C = 7\,10^{-3}\,\mu F$$

avec un coefficient χ^2 qui vaut 11,2, ce qui correspond à une probabilité d'être supérieure à cette valeur de 5 % seulement. On voit qu'une mauvaise estimation des incertitudes peut conduire à un rejet du modèle, alors que les autres résultats obtenus avec une meilleure analyse permettent de l'accepter avec un taux de confiance élevé. Ici, il est en effet clair que les incertitudes sur la fréquence sont bien inférieures à celles sur l'impédance.

6.3. **Commentaires**

L'utilisation d'un seul appareil pose un certain nombre de problèmes. Le premier consiste à faire remarquer que les mesures d'une même grandeur ne sont pas indépendantes, en particulier si le calibre n'est pas changé (c'est le cas ici). En effet, la comparaison à un étalon montrerait certainement une erreur systématique (on a une idée de ce phénomène si on compare deux appareils : pour un calibre donné, leur différence d'indication est en général de signe constant). Pour des raisons économiques, le constructeur ne fait pas une telle étude mais se contente d'éliminer tout appareil qui ne respecte pas la marge d'incertitude affichée. Le deuxième est lié à l'arbitraire de la relation entre intervalle calculé par les données constructeur et écart-type attribué à la variable. Les résultats obtenus ne sont donc pas statistiquement rigoureux mais on peut considérer qu'ils fournissent une bonne indication sur la précision de la méthode et de la qualité du modèle choisi (valeur faible de χ^2).

La deuxième façon d'opérer est plus satisfaisante car on peut espérer, avec plusieurs appareils, réduire l'effet d'erreur systématique évoqué plus haut. On peut constater que la précision du résultat est améliorée par rapport à l'utilisation d'un seul appareil, (ce qui justifie la multiplication des mesures) mais ce qui se traduit également par une augmentation du χ^2.

La différence entre méthode classique et méthode rigoureuse n'est pas négligeable puisque, avec plusieurs appareils, l'écart entre les valeurs de R est d'un écart-type environ.

CONCLUSIONS

L'exemple de la régression linéaire illustre de façon presque carricaturale les problèmes posés par l'utilisation de l'outil statistique dans l'enseignement des Sciences Expérimentales. On peut les résumer en deux types principeux mais qui se complètent :

- ***Inventaire des conditions d'utilisation des outils*** (en particulier conditions d'indépendance des variables, de mises en commun de divers résultats...).

Si ces réflexions sont intéressantes, elles peuvent aussi conduire à la conclusion que la complexité des situations est telle qu'on ne sait pratiquement jamais résoudre le problème ; on se contente alors de quelques considérations vagues et la multiplication des mesures ne se justifie guère. Le problème qui se pose est alors de savoir si on cherche la rigueur, ce qui exclut souvent toute analyse statistique, ou si on utilise des outils en connaissant les limites d'utilisation mais en visant plus la discussion des résultats que leur valeur absolue. Une expérience, au lycée ou à l'université, n'a pas une vocation métrologique. Elle a pour but de faire prendre conscience aux étudiants des problèmes posés par le mesurage et l'analyse des résultats, les outils statistiques étant un moyen de mener cette analyse de façon plus efficace sans en être esclave.

- ***Choix de l'outil adapté aux conditions de recueil des données***

Ce dernier point peut conduire très vite à l'utilisation de logiciels de traitement de données sophistiqués mais générateurs magiques de résultats. Il faut donc être attentif à ce que les techniques ne soient pas un obstacle aux analyses et au sens critique des élèves devant les résultats.

Le problème des calculatrices se pose en ces termes. Si certaines conditions ne sont pas remplies, on sait que leurs indications ne sont pas les meilleures estimations de ce qu'on recherche. Mais le résultat, même constestable, n'est-il pas préférable à des évaluations grossières d'incertitudes ?

Il faut aussi savoir faire un traitement visuel des données quand cela est possible sans obligatoirement mettre en œuvre des outils sophistiqués. En effet, si les certitudes sont suffisamment importantes, les «barres d'erreur» traditionnelles (qui peuvent être la représentation de deux ou trois écart-type si on a fait plusieurs mesures) sont représentables graphiquement et on peut ainsi estimer grossièrement les incertitudes sur la pente et l'ordonnée à l'origine en traçant deux *«droites de confiance extrêmes»*. Il ne faut pas confondre ces droites avec les *«droites extrêmes»* qu'on trouve encore dans de nombreux fascicules et qui sont souvent très pessimistes. Il faut entendre par là les deux droites *au delà desquelles il paraît peu raisonnable de penser que la meilleure droite cherchée puisse se touver.*

Cette façon d'opérer présente plusieurs avantages :

– elle donne une estimation de l'incertitude sur a et b en terme de «taux de confiance» même si ce taux est impossible à chiffrer. Cela signifie que cet intervalle est raisonnable, qu'il pourrait être différent. On peut d'ailleurs sentir, en traçant ces deux droites, que le choix qu'on fait est lié au taux de risque qu'on est prêt à prendre, lequel peut varier selon le rôle et l'importance du résultat, illustrant ainsi cette dualité très importante : *un intervalle de confiance est toujours lié à un taux de confiance,* c'est-à-dire finalement à un enjeu.

– elle permet de tenir compte du poids des différents points en fonction de leur incertitude. Par exemple, si les points pour les grandes valeurs de x sont moins précis, on attachera plus d'importance aux points près de l'origine, ce qui se traduira par le fait que l'intersection des deux droites se trouvera plus près de cette origine.

– elle permet de visualiser le fait qu'une interpolation à partir de la droite de régression sera d'autant plus précise qu'elle se fait près de l'intersection des deux droites de confiance extrêmes, c'est-à-dire près du barycentre du nuage de points si les incertitudes sur ces différents points sont voisines. Par contre, dans le cas évoqué dans l'alinéa précédent, cela signifie que l'interpolation près de l'origine est de meilleurs qualité que dans les autres domaines.

Si la précision des mesures est telle que leur visualisation est impossible sur le graphe, l'estimation précédente est caduque. C'est là qu'un outil statistique est indispensable pour avoir une estimation des incertitudes sur les coefficients de la droite de régression. C'est là donc que la calculatrice, *même si on sait qu'elle ne donne pas un résultat*

rigoureux, peut être utilisée avantageusement pour obtenir un ordre de grandeur des incertitudes.

Quels que soient les moyens choisis, il est important que les étudiants aient conscience de l'enjeu de toute mesure, les outils ne venant que leur donner des résultats plus sûrs pour leurs conclusions.

Je tiens à remercier mes collègues P. FONTES et G. TORCHET qui m'ont incité à mettre au clair ces idées et qui m'ont apporté leurs critiques et suggestions.

BIBLIOGRAPHIE

[1] Y. CORTIAL B.U.P. nº 725 1990.

[2] CETAMA. Statistique appliquée à l'exploitation des mesures, 1986 Masson.

[3] K.S. KRANE, L. SCHECTER American Journal of Physics 50(1) 1982.

[4] J. OREAR American Journal of Physics 50(10) 1982.