# Régression linéaire simple : à la défense du $R^2$

**Résumé** L'élaboration d'une droite d'étalonnage s'effectue la plupart du temps en régression linéaire simple par la méthode des moindres carrés ordinaires. La difficulté résidant dans l'établissement d'une valeur limite inférieure acceptable, cet article présente une procédure permettant de l'estimer, en se basant sur l'exemple courant d'une manipulation en spectrophotométrie UV-visible.

**Mots-clés** **Limite $R^2$, régression linéaire, chimie, incertitude.**

Après avoir été pendant des décennies la référence ultime d'une bonne linéarité lors de l'établissement d'une droite de régression linéaire simple, le coefficient de corrélation de Pearson mis au carré ($R^2$) est aujourd'hui largement délaissé, pour ne pas dire méprisé. Or cette attitude est excessive, car il apporte des informations, notamment dans un travail de routine.

## Que reproche-t-on exactement au coefficient $R^2$ ?

1. Un des premiers arguments des détracteurs est son manque de fiabilité. Reprenant le célèbre quartet d'Anscombe [1], Internet regorge de vidéos désopilantes dans lesquelles différents nuages de points totalement aléatoires possèdent le même $R^2$.
2. Le $R^2$ n'est pas robuste, trop sensible aux points atypiques.
3. L'estimateur $\hat{r}$ est un estimateur biaisé [2] : $E(\hat{r}) = r - \frac{r(1-r^2)}{2n}$, et surestime la réalité.
4. La restriction du nuage de points sur une plage plus petite modifie la valeur du R [3].
5. $R^2$ ne rend pas compte des éventuels biais des couples (xi,yi), reproche récurrent en statistique inférentielle.
6. $R^2$ ne permet pas de comparer deux modèles.

Tous ces reproches sont bien réels, mais ne nous concernent pas vraiment lors d'un travail de routine. En effet, au début du XX$^e$ siècle, les précurseurs que sont Ronald Fisher ou Karl Pearson tentaient, en tant que généticiens, de mettre en lumière des corrélations. Ils ont depuis été rejoints par les économistes qui ont les mêmes préoccupations.

## Mais en physique et en chimie ?

La différence fondamentale réside dans la question : cherche-t-on à prouver une linéarité, ou bien le modèle de linéarité étant avéré, à obtenir une droite la plus précise possible ? Dans ce dernier cas, les choses diffèrent totalement : U = RI ; U et I sont proportionnels et les points doivent être alignés. De même pour la loi de Beer-Lambert, ou alors il est urgent de revoir tous les programmes de physique et de chimie. Et par expérience, un $R^2$ de l'ordre de 0,98-0,99 est attendu.

Nous ne sommes pas dans la même gamme de $R^2$. Théoriquement, des points parfaitement alignés correspondent à un $R^2$ égal à 1 ; les incertitudes expérimentales, voire définitionnelles, abaissent légèrement sa valeur, mais peu [4]. De même qu'on peut faire un spectre infrarouge sans porter de matériel de radioprotection – et pourtant il s'agit d'ondes électromagnétiques… –, les longueurs d'onde ne sont pas au même endroit de l'échelle que les rayons gammas.

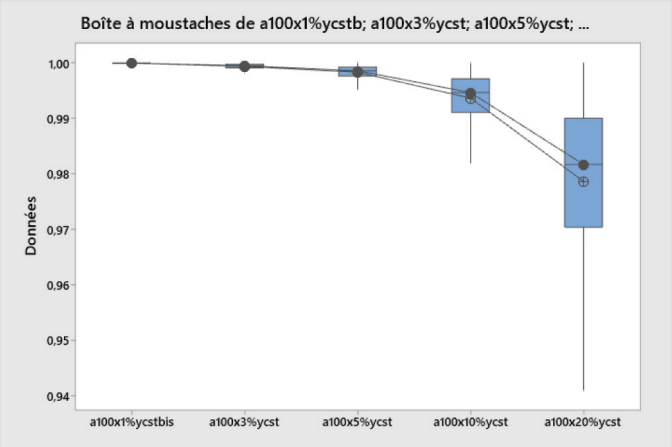



Figure 1 - Simulations $R^2$ incertitude sur x croissante ; constante sur y ; pente 100.

Partant de cette position assez résolument bayesienne, à titre d'exemple, une simulation Monte Carlo sur 450 000 valeurs a été réalisée dans le cas « incertitude sur x de 1 à 20 %, incertitude sur y constante (fixée à 0,2 % de la valeur max de y), pente de 100 ». Les boites à moustaches de Tukey des $R^2$ obtenus sont regroupées *figure 1*. Sur cette simulation(1), le $R^2$ décroit quand l'incertitude sur x augmente, mais reste tout de même, dans le pire des cas, supérieur à 0,94, même pour des incertitudes de 20 %. Le point 1 des reproches ne nous concerne donc pas ; Anscombe lui-même ayant effectué ses simulations sur un $R^2$ de 0,66.

De même, les reproches 4, 5 et 6 tombent d'eux-mêmes : lors d'un travail de routine, la plage de travail est connue et répertoriée. Il ne s'agit pas de comparer deux modèles, mais deux séries de résultats sur une plage donnée, le modèle de linéarité étant avéré et le biais(2) ayant été préalablement corrigé.

## Robustesse du $R^2$

Le statisticien dispose d'un certain nombre d'indicateurs qui lui permettent de déterminer si un point est atypique, voire aberrant [5], et donc d'estimer lui-même la validité du reproche 2 (voir *annexe** A1).

## Le $R^2$, un excellent indicateur ?

En réalité, les deux reproches (d'importance) qu'un physicien ou un chimiste peuvent émettre sont d'un autre ordre : $R^2$ ne possède aucune loi à densité, et il ne permet en rien de prédire une valeur inconnue. En cela c'est un indicateur compliqué.

Le fait que $R^2$ n'ait pas de loi à densité a une conséquence dramatique : il est impossible de fixer de façon universelle

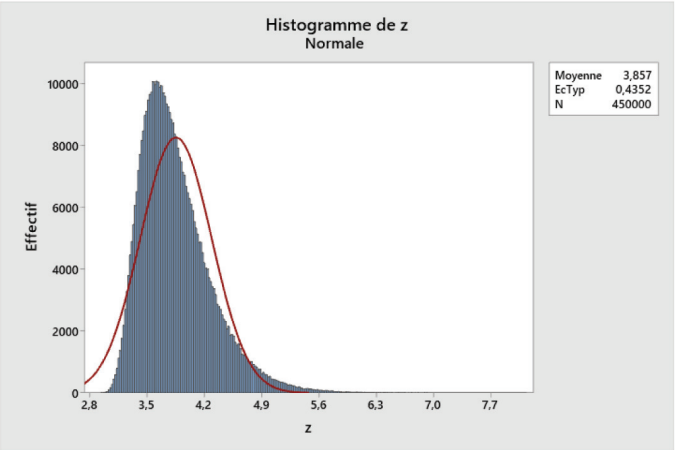



Figure 2 - Exemple des valeurs de z pour l'exemple en chimie.

une limite en dessous de laquelle on peut estimer que la valeur obtenue soit acceptable. Les limites de significativité classiques obtenues par la loi de Student[3] sont bien trop basses pour être utilisables (par exemple, le seuil de significativité à 95 % de confiance de $R^2$ est de 0,658 pour 6 points).

De même, pour tous les cas qui ont été testés par simulation Monte Carlo (incertitudes croissantes sur x et y ; constantes ; décroissantes ; croissantes sur l'un, constantes sur l'autre, etc.), la transformation classique de Fisher[4] ne fonctionne pas pour 5 ou 6 points (*figure 2*) : z ne suit pas une loi normale[5].

Mais alors, comment savoir si une valeur est acceptable ? Il n'y a pas d'universalité ; chaque cas est différent et la question est mal posée. Il n'est pas possible de répondre par l'affirmative, mais on peut raisonner par la négative. Revenons à la *figure 1* : dans cette configuration d'incertitudes, un $R^2$ expérimental de 0,98 par exemple ne nous permet pas d'estimer l'incertitude sur x, mais nous permet d'être certain qu'elle ne peut pas être de 1 à 5 %. Un $R^2$ de 0,999 ne permet pas de savoir quelle est l'incertitude puisqu'il est probable dans tous les cas – on pourrait éventuellement, si on obtient plusieurs $R^2$ de 0,999, estimer une probabilité que l'incertitude sur x ne soit pas de 20 % par une simple loi binomiale, voire une loi de Poisson.

## Un exemple en chimie

L'exemple choisi l'a été en chimie car les déterminations d'incertitude sont bien plus délicates qu'en physique.

Trente-trois étudiants de BTS Métiers de la chimie en fin de formation ont réalisé l'expérience suivante :

Une gamme étalon de cinq concentrations croissantes de dichromate de potassium, chaque modalité ayant été répliquée trois fois, soit quinze fioles de 50 mL à préparer (toutes verreries de classe A) à partir d'une solution préalablement étalonnée de dichromate de potassium de concentration Cmère = 0,010390 ± 0,000030 $mol.L^{-1}$. L'absorbance a ensuite été mesurée à 450 nm dans un spectrophotomètre UV-visible préalablement étalonné en absorbance et en longueur d'onde.

| Vmère (mL) | 2 | 4 | 6 | 8 | 10 |
|---|---|---|---|---|---|
| Cf ($mol.L^{-1}$) | 0,000416 | 0,000831 | 0,001247 | 0,001662 | 0,002080 |

Nous avons en réalité effectué deux séries d'expérience sur du dichromate de potassium et du sulfate de nickel, en dépit de leur toxicité, de façon à raisonner sur une pente importante (environ 415) et une pente faible (environ 5). Nous ne présentons ici que le premier cas, aucune différence significative n'ayant été détectée. Une pente inférieure à 5 ne semble pas raisonnable : d'une part les incertitudes sur la mesure prenant une trop grande importance relative, d'autre part une pente tendant vers 0 étant caractéristique d'une absence de corrélation.

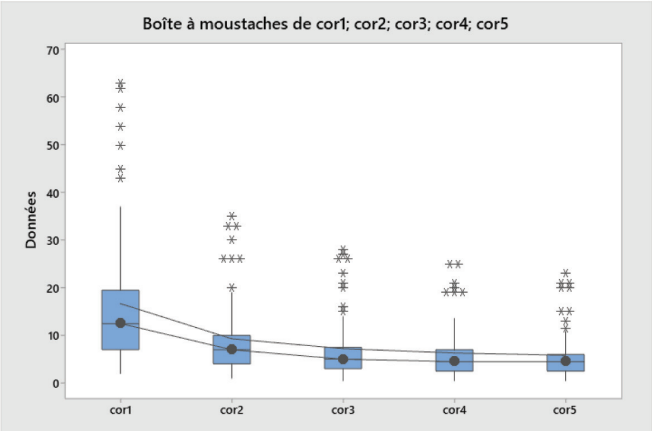
Après avoir retiré les valeurs jugées aberrantes selon le § 8.3 de la norme ISO 5725-2 : « Méthode de base pour la détermination de la répétabilité et de la reproductibilité d'une méthode de mesure normalisée » par l'intermédiaire des tests de Cochran et Grubbs[6], les valeurs obtenues ont été comparées à celles issues d'une simulation Monte Carlo. Cette simulation nécessite au préalable d'estimer au mieux les incertitudes sur les concentrations et les absorbances.

Les incertitudes sur l'absorbance, très clairement expliquées par la théorie [5], ont été finalement maximisées respectivement à ± 0,0020 : cette valeur d'incertitude est compatible avec les tolérances des appareils de moyenne gamme de prix.

Les incertitudes sur les concentrations ont été déterminées en utilisant la loi de propagation des incertitudes appliquée à $C = \frac{A-b}{a}$ :

$$uc^2 = \frac{uA^2}{a^2} + \frac{ub^2}{a^2} + \left[\frac{(A-b)}{a^2}\right]^2 ua^2$$

Les incertitudes relatives sur les cinq concentrations ne suivent pas une loi normale et présentent toutes une asymétrie à droite (*figure 3*) : il n'est pas possible d'effectuer des simulations sur une valeur moyenne. Diverses simulations ont donc été réalisées, en prenant des valeurs d'incertitude comprises entre les premier et troisième quartiles.




| | Moyenne | Q1 | Médiane | Q3 |
|---|---|---|---|---|
| Concentration 1 | 16,7 | 7 | 12,5 | 19,5 |
| Concentration 2 | 9,3 | 4 | 7 | 10 |
| Concentration 3 | 7,2 | 3 | 5 | 7,5 |
| Concentration 4 | 6,3 | 2,5 | 4,2 | 7 |
| Concentration 5 | 5,8 | 2,5 | 3,9 | 6 |

Figure 3 - Incertitudes relatives en % pour les cinq concentrations de dichromate de potassium.

Il n'y a pas une différence significative entre les deux graphes (*figure 4*), si ce n'est qu'obtenir un $R^2$ compris entre 0,990 et 0,998 indique que les incertitudes sur x sont plus proches de la valeur médiane, donc caractéristiques d'une manipulation moins rigoureuse. On peut également constater que les troisième et quatrième chiffres significatifs prennent de l'importance lorsque les incertitudes diminuent.

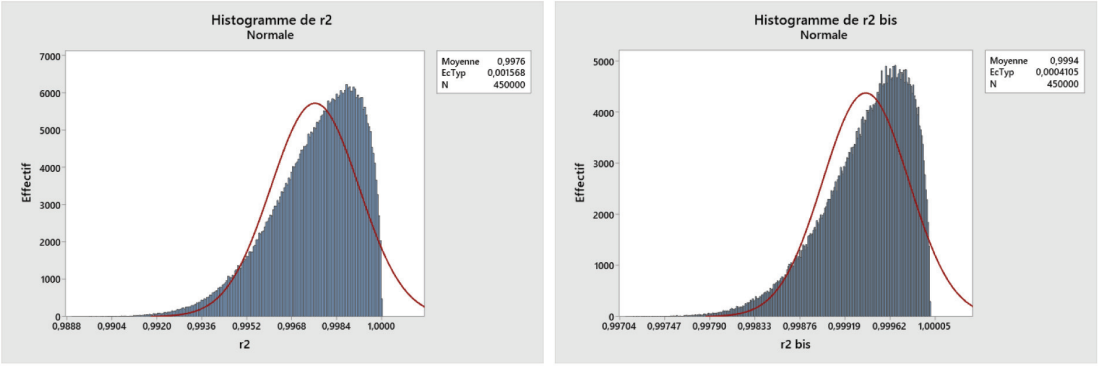



Figure 4 - À gauche : les $R^2$ théoriques obtenus avec les valeurs médianes. À droite : un exemple avec les valeurs médianes divisées par 2.

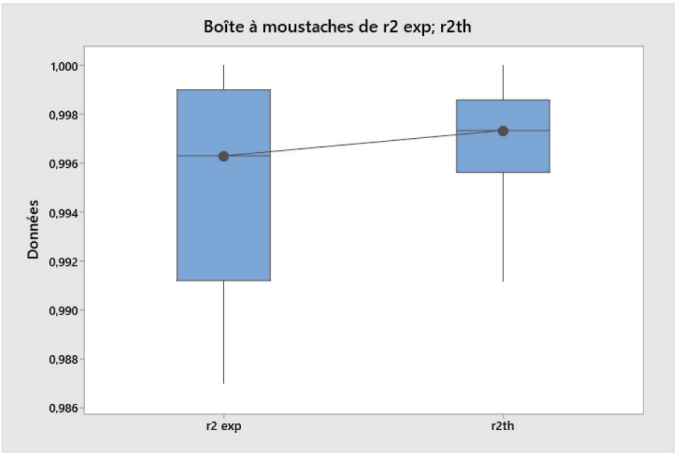



Figure 5 - $R^2$ expérimentaux (71 valeurs) comparées à $R^2$ théorique (450 000 valeurs).

En se référant aux valeurs théoriques (*figure 5*, à droite) obtenues avec les valeurs médianes, on peut émettre l'hypothèse qu'un $R^2$ doive valoir environ 0,991 dans cette manipulation précise, avec cette verrerie. Toutes les valeurs expérimentales en dessous sont suspectes, et il faut observer de plus près la droite obtenue par régression linéaire.
En fin de compte, cela revient à utiliser la règle empirique classique en chimie : « Trois neuf, c'est très bien ; deux neufs, ça va ; un seul, il y a trop d'incertitudes ».
Tout ce travail préliminaire une fois effectué, il devient alors possible de se donner un ordre de grandeur d'une limite, et le fait que la valeur obtenue soit biaisée (reproche 3) n'a plus une grande importance car on compare des estimations.

## Autres indicateurs de la linéarité

D'autres coefficients de corrélation existent [2, 6] :
- Celui de Spearman, qui n'a pas dans notre cas d'intérêt : la variable aléatoire est le rang des observations, et les nôtres sont déjà classées dans l'ordre croissant, sauf grossière erreur de manipulation.
- Le « $R^2$ ajusté », qui peut être très utile en physique car il permet de comparer une droite à 5 points avec une droite à 6 points. En chimie, il est plus délicat d'utilisation car les incertitudes dépendent des conditions expérimentales et du choix des points – comme on l'a vu – et ne présentent que rarement des lois (contrairement par exemple à un multimètre).
- Le « $R^2$ prédit », qui n'apporte rien si on a retiré les points atypiques par l'utilisation des résidus studentisés supprimés.
Le $R^2$ semble un bon compromis. Cependant, entraîné à la méfiance concernant les chiffres significatifs, l'esprit se heurte à une validation basée sur le troisième. Il est alors possible d'utiliser le F de Fisher, dont les valeurs sont plus « naturelles ». Ainsi, un $R^2$ limite de 0,991 correspond à une valeur de F limite d'environ 90, pour 6 points. Dans tous les cas, il n'existe aucune loi universelle qui nous permette de donner une valeur limite acceptable. Ce n'est pas pour autant qu'il faut renier $R^2$. Il permet de comparer plusieurs résultats, dans un travail de routine, voire de proposer des modes opératoires meilleurs, à coût égal [7].

*Les graphes ont été effectués avec le logiciel Minitab.*
** Le fichier des annexes est téléchargeable librement sur www.lactualite-chimique.org (page liée à cet article).*

(1) L'utilisation des boites à moustaches, qui suppose une distribution à tout le moins symétrique, reste valable : les valeurs « aberrantes » ne représentent que 2,1 % du total des valeurs, inférieur aux 5 % communément adoptés, et peuvent donc être retirées.
(2) Biais : d'après le théorème de Bienaymé-Tchebychev ou, plus généralement, de par la définition de l'écart quadratique moyen, un estimateur $\widehat{\theta n}$ est un bon estimateur de θ si : il est sans biais : $E(\widehat{\theta n}) = \theta$ ; il est convergent : $\lim V(\widehat{\theta n}) = 0$.
(3) Loi de Student : est utilisée lorsque la variable suit une loi normale(5), mais que l'écart type n'est pas connu mais estimé généralement à partir de celui d'un échantillon.
(4) Transformation de Fisher : le coefficient de corrélation ne suivant aucune loi répertoriée, Fisher a proposé la transformation de variable $Z = 1/2 \ln(\frac{1+r}{1-r})$. Si n est assez grand, Z suit une loi normale d'écart type constant $\frac{1}{\sqrt{n-3}}$, et peut être utilisé pour confirmer la significativité de la différence entre deux coefficients de corrélations. Ce n'est pas le cas ici.
(5) Loi normale : également appelée Loi de Laplace-Gauss, c'est une loi de probabilité continue caractérisée par deux paramètres : son espérance et son écart type σ. La courbe de la densité de probabilité est la célèbre courbe en cloche.
(6) Tests de Cochran et Grubbs : dans notre cas, le test de Cochran permet de déterminer si l'écart type des résultats d'un laboratoire est trop élevé par rapport aux résultats d'ensemble. Le test de Grubbs identifie les laboratoires dont la moyenne est trop éloignée de la moyenne d'ensemble. Les tests ont été effectués en autonomie par les étudiants qui, par un processus itératif « Cochran puis Grubbs », ont décidé par eux-mêmes des laboratoires à éliminer dans la limite préconisée des 2/9e maximum, soit sept laboratoires dans notre cas.

[1] https://fr.wikipedia.org/wiki/Quartet_d%27Anscombe (consulté le 29/11/2021).
[2] https://eric.univ-lyon2.fr/~ricco/cours/cours/Analyse_de_Correlation.pdf, p. 13 (consulté le 15/11/2021).
[3] https://delladata.fr/regression-lineaire-simple-le-r%C2%B2-info-ou-intox (consulté le 28/11/2021).
[4] F. Grégis, La valeur de l'incertitude : l'évaluation de la précision des mesures physiques et les limites de la connaissance expérimentale, Thèse de doctorat, Paris 7, **2016**, p. 227.
[5] https://metrologie-francaise.lne.fr/sites/default/files/media/document/rfm41-1601.pdf
[6] https://statisticsbyjim.com/regression/interpret-adjusted-r-squared-predicted-r-squared-regression/(consulté le 15/11/2021).
[7] Voir articles des mêmes auteurs en *annexe** : Régression linéaire simple : 1. Variante de la pratique courante : linéarité ; 2. Variante de la pratique courante : prévision.

**Christophe ROUSSEL***, professeur de synthèse en BTS Métiers de la chimie, Lycée Jean Perrin, Marseille, et **Alexis ROUSSEL**, étudiant en Master 1 Instrumentation, Mesure, Métrologie, Aix Marseille Université.

*christophe.roussel@free.fr ; xi.roussel@gmail.com

# Compléments à l'article « Régression linéaire simple : à la défense du $R^2$ », par C. Roussel et A. Roussel (*L'Act. Chim.*, 2022, *470*, p. 42)

## Annexe A1 : Recherche de points atypiques : indicateurs

### Effet levier

$$\begin{cases} Y = a0 + a1x + \varepsilon \\ y1 = a0 + a1x1 + \varepsilon 1 \\ \vdots \\ yn = a0 + a1xn + \varepsilon n \end{cases}$$

En termes vectoriel, $\begin{pmatrix} y1 \\ \vdots \\ yn \end{pmatrix} = a0 \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} + a1 \begin{pmatrix} x1 \\ \vdots \\ n \end{pmatrix} + \begin{pmatrix} \varepsilon 1 \\ \vdots \\ \varepsilon n \end{pmatrix}$, où $\mathbf{Y} = a0\ \mathbf{1} + a1\ \mathbf{X} + \boldsymbol{\varepsilon}$

**Y** appartient donc à l'espace vectoriel (**1** ; **X** ; **ε**) (attention : ces trois vecteurs ne sont pas orthogonaux)

$\widehat{\boldsymbol{Y}} = \widehat{a0}\ \mathbf{1} + \widehat{a1}\ \mathbf{X}$, combinaison linéaire de **1** et **X** appartient alors au sous-espace vectoriel engendré par **1** et **X**, et est donc la projection orthogonale du vecteur Y sur (**1** ; **X**).

On peut écrire $\hat{Y} = Px\ Y$, ou $\hat{Y} = \hat{H}\ Y$

$$\begin{pmatrix} \hat{y}1 \\ \vdots \\ \hat{y}n \end{pmatrix} = \begin{pmatrix} h11 & \cdots & h1n \\ \vdots & \ddots & \vdots \\ hn1 & \cdots & hnn \end{pmatrix} \begin{pmatrix} y1 \\ \vdots \\ yn \end{pmatrix}$$

On démontre que la matrice $\hat{H} = X(X'X)^{-1}X'$, X étant la matrice d'expériences, composée de deuc colonnes représentant les vecteurs **1** $\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$ et **X** $\begin{pmatrix} x1 \\ \vdots \\ xn \end{pmatrix}$ : $X = \begin{pmatrix} 1 & x1 \\ \vdots & \vdots \\ 1 & xn \end{pmatrix}$

En écriture matricielle : $\begin{pmatrix} y1 \\ \vdots \\ yn \end{pmatrix} = \begin{pmatrix} 1 & x1 \\ \vdots & \vdots \\ 1 & xn \end{pmatrix} \begin{pmatrix} a0 \\ a1 \end{pmatrix} + \begin{pmatrix} \varepsilon 1 \\ \vdots \\ \varepsilon n \end{pmatrix}$ soit $Y = X\beta + \varepsilon$

Remarque : les vecteurs **1** et **X** ne sont pas orthogonaux.

En résumé, les valeurs prédites $\hat{y}i$ se calculent selon $\hat{Y} = \hat{H}\ Y$, soit :

$$\hat{y}i = hii\ yi + \sum_{j \neq i} hijyj$$

Par construction, les valeurs hij sont totalement indépendantes des valeurs de Y et ne dépendent que des valeurs de x, c'est-à-dire de nos conditions expérimentales et non pas des résultats que nous avons obtenus.

Les termes diagonaux ont alors une grande importance : ils traduisent l'influence de la ième mesure sur sa propre estimation. Plus hii est grand (on démontre par ailleurs qu'ils sont tous positifs inférieurs à 1), plus la valeur de yi a du poids pour l'estimation de $\hat{y}i$ et donc pour l'estimation de la droite de régression : le point devient influent.

La valeur limite de hii correspondant à un point influent fait débat. Un point est dit avoir un effet levier dans le cas d'une régression linéaire simple si :

- hii> 4/n selon Hoaglin et Welsh, 1978.
- Hii> 0,5 selon Huber, 1981.

Quelle que soit la limite que l'on choisisse, on s'aperçoit rapidement que les points ayant le plus fort effet levier, donc les plus influents, sont les points de xi extrêmes, éloignés de la moyenne $\bar{x}$ (d'où le nom d'effet levier).

### Résidus, résidus normalisés ou studentisés ?

Parmi les hypothèses de base de la régression linéaire, les erreurs ne sont pas corrélées et ont la même variance (homoscédasticité).

Ces deux hypothèses peuvent se modéliser par une matrice de variance covariance

$$V(\varepsilon) = \begin{pmatrix} \sigma^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma^2 \end{pmatrix} = \sigma^2 \begin{pmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{pmatrix} = \sigma^2$$ In avec In matrice identité (n,n).

Mais nous n'avons pas connaissance des erreurs . Nous avons accès à leurs réalisations $\hat{\varepsilon}$ lors des différentes expériences.

Or $\hat{\varepsilon} = Y - \hat{Y} = Y - \hat{H}\ Y = InY - \hat{H}\ Y = (In - \hat{H})Y$

D'où il s'ensuit que $V(\hat{\varepsilon}) = \sigma^2\ (In - \hat{H})$

Ce dernier résultat, dont la démonstration importe peu ici, conduit à un résultat important :
$V(\hat{\varepsilon} i = {}^{2}(1\text{-hii})$
Ce qui signifie, d'une part l'hétéroscédasticité, mais aussi et surtout que les résidus dépendent de la position du point sur la droite. Ils sont sous-estimés pour les points extrêmes. Il n'existe aucune théorie concernant les résidus susceptibles de nous fournir une limite au-delà de laquelle le point serait atypique, voire aberrant, car la variable t = – ne suit ni une loi normale ni de Student.
Le résidu normalisé $r = \frac{\hat{\varepsilon} i}{\sqrt{V(\hat{\varepsilon} i)}} = \frac{\hat{\varepsilon} i}{\sigma\sqrt{(1-\text{hii})}}$ est bien plus réaliste que le résidu, mais possède un inconvénient majeur : est inconnu.

Ce problème peut se résoudre en remplaçant par son estimation $\hat{\sigma}$, mais la nouvelle variable alors définie devient $t = \frac{\hat{\varepsilon} i}{\hat{\sigma}\sqrt{(1-\text{hii})}}$ qui suivrait une loi de student, d'où son nom de résidu studentisé.

Ce résidu amélioré ne suit toujours pas une loi de Student. C'est pourquoi on définit le seul résidu qui suive une loi de Student [1], dit « résidu studentisé par validation croisée », ou « résidu studentisé supprimé », par $t^* = \frac{\hat{\varepsilon} i}{\hat{\sigma}(i)\sqrt{(1-\text{hii})}} \rightsquigarrow \mathcal{T}_{n-3}$.

Ce résidu est obtenu par la technique du « leave-one-out » et est proposé par tous les logiciels de statistiques classiques. Ladite technique consiste, pour simplifier, à retirer le point considéré des données, recalculer la droite de régression et estimer si le point qu'on a retiré entre dans l'intervalle de prévision de la nouvelle droite. Ainsi, $\hat{\sigma}(i)$ est l'estimation de sur la droite dont le point i considéré a été retiré, d'où le fait que t* suive une loi $\mathcal{T}_{(n-1)-2} = \mathcal{T}_{n-3}$.
Nous avons maintenant un résidu qui peut nous permettre d'estimer une limite (que, par commodité, on estime souvent égale à tort à 2).
Néanmoins, cet indicateur, aussi bon soit-il, doit être pris avec une grande précaution, et doit toujours être confronté à la réalité de la droite, tel le cas ci-dessous :

| x | y | T SUPPR |
|---|---|---|
| 1 | 2 | - 0,5 |
| 2 | 4 | - 0,4 |
| 3 | 6,05 | 633714,9 |
| 4 | 8 | - 0,4 |
| 5 | 10 | - 0,5 |

Ce cas est passablement limite et peut faire prendre conscience que l'étude de la régression par ses différents indicateurs ne s'impose pas nécessairement lorsque les points sont presque parfaitement alignés.

**Distance de Cook et DFFITS**

Nous allons peut-être pouvoir donner maintenant une réponse à la question : puis-je raisonnablement estimer que ce point est aberrant, et donc l'enlever sans remords ?
Assez naturellement, on tendrait à enlever un point qui posséderait un résidu studentisé supprimé et un effet levier important. Deux indicateurs existent, qui rassemblent ces deux informations, en indiquant l'influence globale d'un point sur la régression. Ces deux indicateurs sont proposés également dans les logiciels de statistique.
- La distance de Cook, dont Cook lui-même a proposé comme limite inférieure préoccupante la valeur $F^2_{n-2}(0,5)$ (quantile de p > 0,5) :
Ces limites sont maintenant largement controversées : pour certains statisticiens, une valeur supérieure à 1 est atypique quel que soit le nombre de points, d'autres utilisent une limite de 4/n, d'autres encore 4/(n-3).
Une tentative de conciliation entre toutes ses différentes limites nous conduit à estimer qu'un D compris entre 0,8 et 1 mérite qu'on s'y attarde, et D > 1 anormal. En effet, encore une fois nous ne cherchons pas à établir une relation linéaire entre deux variables, celle-ci est établie. Les points devraient donc être alignés, aux incertitudes près, et tout point qui s'écarte trop de la linéarité est représentatif d'un problème.
- Le DFFITS dont la limite inférieure consensuelle est fixée à $\frac{3,46}{\sqrt{n}}$
Tous ces indicateurs peuvent être calculés par Excel [2].

Tableau résumant les valeurs limites : valeurs inférieures des différents indicateurs.

| Indicateur | | 4 points | 5 points | 6 points |
|---|---|---|---|---|
| Effet levier h | Hoaglin et Welsh | 1 | 0,8 | 0,67 |
| | Huber | 0,5 | 0,5 | 0,5 |
| Résidu studentisé supprimé \|t*\| | $t_{n-3,0,975}$ | 12,71 | 4,30 | 3,18 |
| Distance de Cook D | Cook | 1 | 0,88 | 0,82 |
| | 4/n | 1 | 0,8 | 0,67 |
| DFFITS | $\frac{3,46}{\sqrt{n}}$ | 2 | 1,79 | 1,63 |

## Annexe A2 : Différents indicateurs de la linéarité

$$\underset{\text{SCT}}{\sum_{i=1}^{n}(yi-\bar{y})^2} = \underset{\text{SCR}}{\sum_{i=1}^{n}(yi-\hat{y}i)^2} + \underset{\text{SCE}}{\sum_{i=1}^{n}(\hat{y}i-\bar{y})^2}$$

- coefficient de détermination $R^2 = \frac{SCE}{SCT} = \frac{\sum_{i=1}^{n}(\hat{y}i-\bar{y})^2}{\sum_{i=1}^{n}(yi-\bar{y})^2} = r^2$ le coefficient de corrélation au carré dans le cas d'une régression linéaire simple.

- $F = \frac{\frac{SCE}{1}}{\frac{SCR}{n-2}} = \frac{SCE}{SCT-SCE}(n\text{-}2) = \frac{R^2}{1-R^2}(n\text{-}2)$

- $R^2$ ajusté $= 1 - \frac{(n-1)(1-R^2)}{n-2}$, avec n nombre de points.

### Références


[1] P.A. Cornillon, E. Matzner Lober, *Régression, théorie et applications*, Springer, **2007**.
[2] http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/fr_Tanagra_Regression_Excel.pdf


# Régression linéaire simple (partie 1)

## Variante de la pratique courante : linéarité

La réalisation pratique d'un étalonnage par une « droite d'étalonnage » fait appel à la théorie de la régression linéaire simple par la méthode des moindres carrés ordinaire (MMCO). Cet article propose donc une amélioration possible de la pratique de l'élaboration d'une droite d'étalonnage, appuyée par des réflexions mathématiques simples.
L'exemple sera celui d'un étalonnage de routine classique en chimie, avec l'utilisation conjointe d'un spectrophotomètre UV-visible et de la loi de Beer-Lambert $A = \varepsilon lc$. La pratique courante consiste à préparer soi-même une gamme de différentes solutions de concentrations pratiquement équidistantes, puis à tracer A = f(C).

### Amélioration de la procédure

La régression linéaire simple consiste, à partir d'un couple de n données $(x_i,y_i)$, i variant de 1 à n, à modéliser une relation affine entre les deux variables $y_i = ax_i + b$.
La méthode des moindres carrés ordinaires (MMCO) conduit aux résultats [1] :

$$\hat{a} = \frac{\sum_{i=1}^{n}(xi-\bar{x}i)(yi-\bar{y})}{\sum_{i=1}^{n}(xi-\bar{x})^2} \text{ et } V(\hat{a}) = \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{(n-2)\sum_{i=1}^{n}(x_i-\bar{x})^2}$$
$$\hat{b} = \bar{y} - \hat{a}\bar{x} \text{ et } V(\hat{b}) = \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-2}\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)$$

Ces deux estimateurs sont de bons estimateurs, car sans biais et convergents (voir *annexe 1*). Ce sont d'ailleurs les meilleurs estimateurs possibles.
Une amélioration du procédé consiste à diminuer des variances de $\hat{a}$ et $\hat{b}$, c'est-à-dire dans cette étude, à augmenter le dénominateur $\sum_{i=1}^{n}(x_i - \bar{x})^2$ : pour ce faire, il suffit d'utiliser plus de points de concentration éloignée de la moyenne.
Si on prépare classiquement une gamme de cinq concentrations, au lieu de préparer cinq solutions de concentrations régulièrement espacées [12345], il est mathématiquement plus opportun de préparer cinq solutions [11155] ou [11555], à coût égal.
Or, la norme ISO 11843-2 (2000) « Méthodologie de l'étalonnage linéaire », § 4.2 préconise :
« *Il y a lieu de choisir les états de référence de sorte que les valeurs de la variable nette d'état* […] *soient approximativement équidistantes dans la plage comprise entre la valeur la plus petite et la valeur la plus grande.*
Cette préconisation est reprise dans la norme ISO 11095 (1996) « Étalonnage linéaire utilisant des matériaux de référence », § 5.3.2.
Il paraît effectivement plus raisonnable d'espacer régulièrement les valeurs de concentration, afin de confirmer la linéarité sur l'ensemble de la plage. Mais est-il véritablement nécessaire d'effectuer cette vérification lors d'un travail de routine ? Sur une plage donnée et répertoriée, $A = \varepsilon lc$, A et C sont théoriquement proportionnels et les points alignés.
Cette approche (assez résolument bayesienne [2], voir *annexe 2*) conduit à tester la méthode[11355], après avoir dans un premier temps déterminé la plage de linéarité, bien évidemment. Il s'agit en fin de compte d'une variante de celle proposée au § 8.2 de la norme ISO 11095, qui propose une droite d'étalonnage en un point, « lorsqu'il n'y a aucun doute sur la linéarité de la fonction sur une plage donnée ». Cette variante proposée par la norme ISO paraît un peu réductrice, l'expérience montrant que la linéarité est vérifiée, mais que la droite obtenue par la MMCO est affine et non linéaire, de par les incertitudes cumulées. Estimer qu'elle passe nécessairement par l'origine ne semble expérimentalement pas être une évidence.

C'est d'ailleurs cette dernière considération qui conduit à utiliser la théorie de la MMCO avec une relation du type y = ax + b et non y = ax, les deux théories ne présentant que peu de différences significatives sur le plan de l'analyse des résultats [3].

**Dangers inhérents à la méthode [11555]**

La concentration minimale devant être bien évidemment supérieure à la limite de quantification (cf *annexe 1*) communément admise $10\frac{s(\hat{b})}{\hat{a}}$ de la méthode utilisée, et la maximale dans le domaine de linéarité, ces deux limites ayant été déterminées lors de manipulations précédentes.
Par ailleurs, les points extrêmes ont un fort effet levier, et même s'il est toujours possible de déterminer leur caractère « aberrant » (cf *annexe 2*), un point central s'impose pour vérifier la linéarité, d'où le choix d'une comparaison [12345] nommée par la suite 51 avec une gamme [11355], nommée 212.

**Réalisation pratique**

Ainsi, trente-trois étudiants de Brevet de Technicien Supérieur Métiers de la chimie en fin de formation ont réalisé deux expériences :
- une gamme étalon de cinq concentrations croissantes de dichromate de potassium, chaque modalité ayant été répliquée trois fois, soit quinze fioles à préparer à partir d'une solution préalablement étalonnée de dichromate de potassium de concentration Cmère = 0,010390 ± 0,000030 mol.L$^{-1}$. L'absorbance a ensuite été mesurée à 450 nm dans un spectrophotomètre UV-visible préalablement étalonné en absorbance et en longueur d'onde.

| Vmère (mL) | 2 | 4 | 6 | 8 | 10 |
|---|---|---|---|---|---|
| Cf (mol.L$^{-1}$) | 0,000416 | 0,000831 | 0,001247 | 0,001662 | 0,002080 |

- une gamme étalon de sept concentrations (seules les première et dernière modalités ont été répliquées une fois, soit sept fioles à préparer) de sulfate de nickel de concentration Cmère = 0,20120 ± 0,00020 mol.L$^{-1}$. Les mesures ont été effectuées à 394 nm.

| Vmère (mL) | 2 | 4 | 6 | 8 | 10 |
|---|---|---|---|---|---|
| Cf (mol.L$^{-1}$) | 0,008048 | 0,016096 | 0,024144 | 0,032192 | 0,040240 |

Dans les deux cas, les dilutions ont été réalisées à l'eau bipermutée avec de la verrerie de classe A. Les fioles ont un volume de 50 mL.
Ces deux solutions ont été sélectionnées, car si l'ion dichromate absorbe bien (absorbances mesurées entre 0,1 et 1), le sulfate de nickel absorbe entre 0,05 et 0,1 sur notre plage de concentrations, la pente est donc faible, ce qui augmente l'incertitude des mesures. C'est d'ailleurs la raison pour laquelle ces deux espèces ont été sélectionnées : quoique toxiques, elles nous permettent de couvrir une large plage de pentes.

**Résultats et discussion**

Après avoir retiré les valeurs jugées aberrantes selon le § 8.3 de la norme ISO 5725-2 : « Méthode de base pour la détermination de la répétabilité et de la reproductibilité d'une méthode de mesure normalisée » par l'intermédiaire des tests de Cochran et Grubbs (cf *annexe 3*), les résultats sont résumés dans le graphe de la *figure 1* :

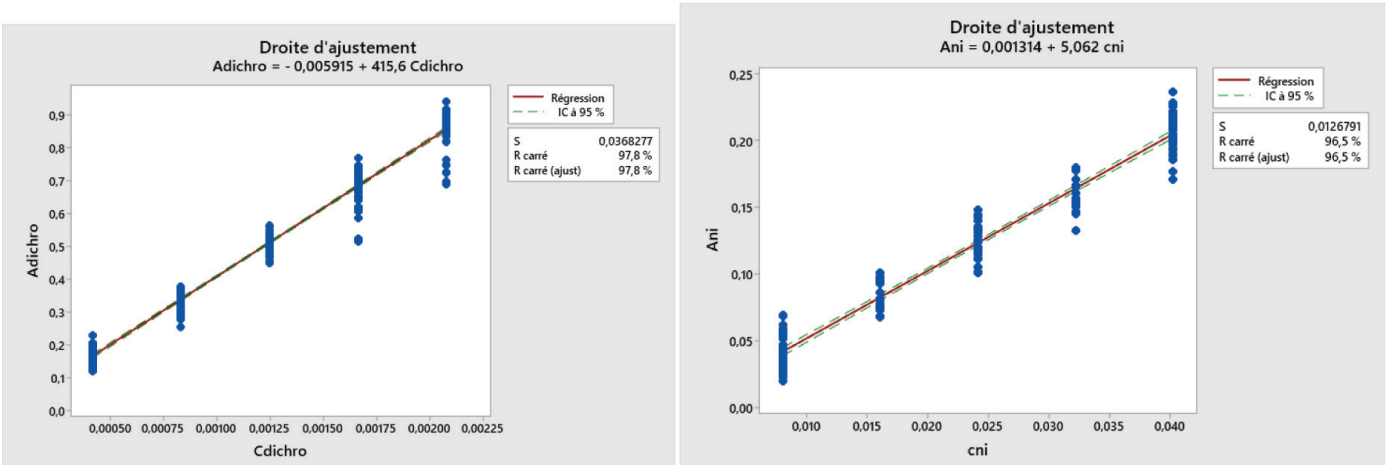



Figure 1 - f(C) pour a) des solutions de dichromate de potassium et b) pour du sulfate de nickel.

Solutions de dichromate de potassium (*figure 1a*) : les concentrations les plus basses C1 suivent une loi normale, ainsi que les deux suivantes (confirmé par un test de Shapiro Wilk). En revanche, ni les concentrations C4 ni C5 ne suivent une loi normale. Les trois premières concentrations possèdent la même variabilité (ce qu'on appelle l'homoscédasticité, confirmée par un test de Bonett), mais n'ont pas la même que C4 et C5 qui ont chacun une variabilité supérieure. Tout ceci est largement visible sur les distributions bleues de la *figure 2*.
Solutions de sulfate de nickel (*figure 1b*) : C1 ne suit pas une loi normale, ni C5. C2 a une variabilité inférieure à celle des quatre autres.
La normalité n'est pas une hypothèse de base de la régression linéaire ; elle permet de préciser des bornes, car elle possède une loi mathématique à densité.
Par contre, l'homoscédasticité est une hypothèse de base : tous les points doivent avoir la même variabilité.
Pourtant, elle est rarement vérifiée dans notre cas : d'une part, l'incertitude sur les faibles concentrations est évidente avec une verrerie classique (volumes prélevés de solution mère faibles, erreur relative importante) ; d'autre part, même si l'incertitude sur les grandes concentrations est faible (volume prélevé plus important, erreur relative plus faible), cette incertitude se propage sur l'absorbance par la multiplication par la pente.
Passant donc outre, les droites ont été tracées et les valeurs des pentes, ordonnées à l'origine, écarts types ont été listés pour les méthodes 51 et 212, et sont résumés dans les *tableaux Ia* et *Ib*).

| | **51** | **212** | **Test T** | **test F** | **Conclusion** |
|---|---|---|---|---|---|
| a | 416,065 | 414,692 | 0,82 | | pas de différence |
| s(a) | 19,798 | 9,443 | | 0,000004 | différence |
| b | 0,000 | -0,014 | 0,14 | | pas de différence |
| s(b) | 0,027 | 0,012 | | 0 | différence |

Tableau Ia - Comparaison des pentes, ordonnées à l'origine et écarts types dichromate de potassium (71 droites de type 51 comparées à 56 droites de type 212).

| | **51** | **212** | **Test T** | **test F** | **Conclusion** |
|---|---|---|---|---|---|
| a | 4,990 | 5,096 | 0,39 | | pas de différence |
| s(a) | 0,341 | 0,004 | | 0,000004 | différence |
| b | 0,001 | 0,004 | 0,62 | | pas de différence |
| s(b) | 0,009 | 0,005 | | 0,01 | différence |

Tableau Ib - Comparaison des pentes, ordonnées à l'origine et écarts types sulfate de nickel (27 droites de type 51 comparées à 26 droites de type 212).

Les comparaisons des ordonnées à l'origine n'a pas de signification, les différentes valeurs n'étant pas significativement différentes de zéro. En ce qui concerne la pente, il n'y a heureusement pas de différence significative (test T). Par contre, et c'est le plus important, on observe bien une diminution significative des différents écarts type (test F), ce qui était attendu.

**Comment évaluer la qualité de l'ajustement ?**

Une première manière d'apprécier la qualité d'un ajustement est l'intervalle de confiance à 95% de la droite. Un intervalle de confiance petit rend compte de la réduction des paramètres de la droite [1], ce qui augmente donc la fiabilité que l'on peut lui accorder. Les droites obtenues par la MMCO sont donc, en ce sens, plus fiables dans le cas de la méthode 212.
Une seconde manière d'apprécier la qualité de l'ajustement consiste à se fier aux indicateurs classiques. Afin d'éviter toute polémique sur l'opportunité d'utiliser le coefficient de corrélation au carré $r^2$, souvent injustement décrié, la qualité de l'ajustement a été évaluée à partir de la valeur du F de Fisher, qui, par sa définition même, permet de comparer deux résultats à n égal (cf *annexe 3*). Pour résumer, plus F est grand, meilleure est la linéarité.
Les deux boîtes à moustaches des F expérimentaux, notés Fexp_51 et Fexp_212, des *figures 2a* et *2b*) confirment bien ce que la diminution des variabilités des pentes et ordonnées à l'origine avaient laissé présager : la linéarité est meilleure pour la méthode 212.
Par ailleurs, on peut tenter de comparer avec des résultats théoriques. « Tenter » est employé à dessein, aucune loi répertoriée, donc aucun test, ne pouvant être utilisés, on ne peut que passer par une simulation Monte Carlo, ce qui a été effectué sur 450 000 valeurs.
Cette simulation nécessite au préalable d'estimer au mieux les incertitudes sur les concentrations et les absorbances.
Les incertitudes sur l'absorbance, très clairement expliquées par la théorie [4], ont été finalement maximisées respectivement à ± 0,0020 pour l'ion dichromate et ± 0,0010 pour le sulfate de nickel pour lequel les absorbances mesurées sont plus basses (entre 0,05 et 0,2).Ces valeurs d'incertitude sont par ailleurs compatibles avec les tolérances des appareils de moyenne gamme de prix.
Les incertitudes sur les concentrations ont été déterminées en utilisant la loi de propagation des incertitudes appliquée à $C = \frac{A-b}{a}$ :

$$uc^2 = \frac{uA^2}{a^2} + \frac{ub^2}{a^2} + \left[\frac{(A-b)}{a^2}\right]^2 ua^2$$

Encore une fois, les incertitudes relatives sur les cinq concentrations ne suivent pas une loi normale et présentent toutes une asymétrie avec une queue allongée à droite (cf *tableau II*, cas du dichromate de potassium).

| | Moyenne | Q1 | médiane | Q3 |
|---|---|---|---|---|
| Concentration 1 | 16,7 | 7 | 12,5 | 19,5 |
| Concentration 2 | 9,3 | 4 | 7 | 10 |
| Concentration 3 | 7,2 | 3 | 5 | 7,5 |
| Concentration 4 | 6,3 | 2,5 | 4,2 | 7 |
| concentration 5 | 5,8 | 2,5 | 3,9 | 6 |

Tableau II - Erreurs relatives en % pour les cinq concentrations de dichromate de potassium.

En ce qui concerne le sulfate de nickel, les résultats sont sensiblement les mêmes, toutes les valeurs du tableau précédent étant systématiquement supérieures de 0,5 % à celles du dichromate de potassium. Ce dernier résultat est d'ailleurs très étonnant et sera commenté dans un prochain article.
Dans ces conditions, différentes simulations ont été tentées, en faisant varier les différentes incertitudes relatives des différentes concentrations de 2 % autour de la médiane. Les résultats ne varient pas de façon significative. En ce qui concerne la boîte à moustaches Fthéorique 51, ils ne varient pratiquement pas. En ce qui concerne la boîte Fthéorique 212, la taille varie légèrement. Nous présentons *figure 2* les résultats obtenus avec les valeurs médianes.

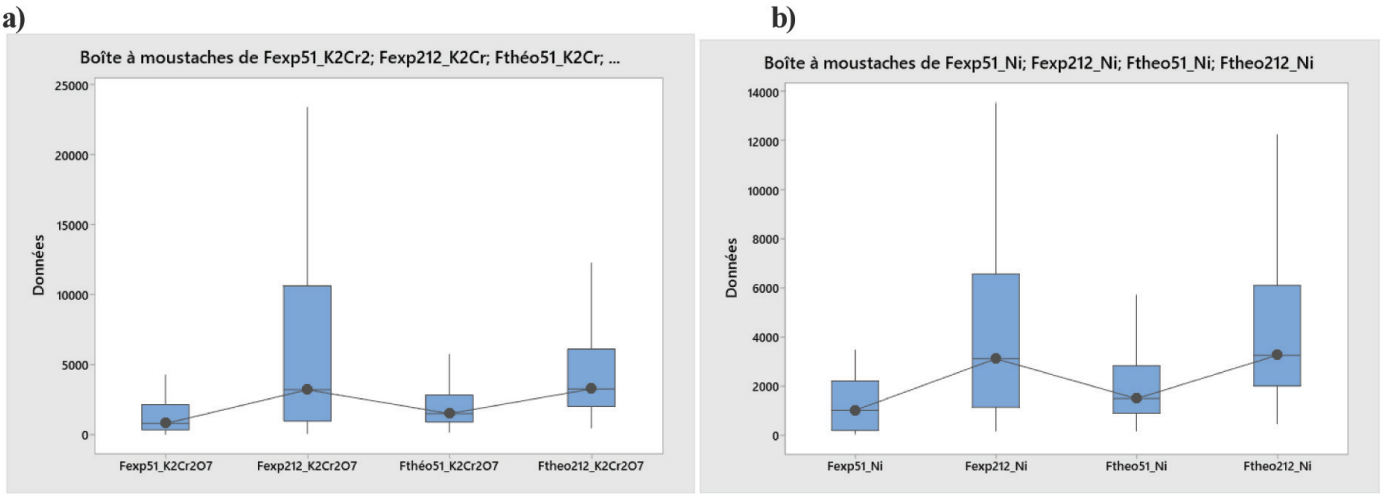



Figure 2 - a) Dichromate de potassium ; b) sulfate de nickel.
De gauche à droite : F expérimental méthode 51 ; F expérimental méthode 212 ; F Monte Carlo méthode 51 ; F Monte Carlo méthode 212.

La similitude évidente entre les boîtes à moustaches expérimentales et celles obtenues par simulation Monte Carlo confirme les résultats expérimentaux et assure la cohérence de l'ensemble : l'ordre de grandeur des concentrations, la supériorité de la méthode 212 et l'utilisation de F comme indicateur de la linéarité, dans ce cas précis.

**Efficacité de la méthode dérivée [11355]**

La variante intermédiaire de celles, classiques, proposées par la norme ISO 11843-2, qui consiste à établir une gamme d'étalonnage « deux petites valeurs, une moyenne, deux grandes » minimise les variabilités des pentes et ordonnées à l'origine, assurant ainsi une meilleure linéarité ainsi qu'une diminution des limites de détection et de quantification.
Elle peut être utilisée lors d'un travail de routine, après avoir déterminé le domaine de linéarité exploitable de travail.

## Annexes

### 1. Limites de détection et de quantification

La limite de quantification $10\frac{s(\hat{b})}{\hat{a}}$ est empirique, en tout cas n'a pas à notre connaissance de démonstration mathématique.
L'élaboration de la limite de détection tente de concilier une approche empirique et une démonstration rigoureuse telle qu'explicitée dans la norme ISO 11843-2. Elle conduit à :

$$xc \approx 2t_{n-2,\,0,95}\frac{s(\hat{b})}{\hat{a}}$$

| Nombre de points | 4 | 5 | 6 |
|---|---|---|---|
| $t_{n-2,\,0,95}$ | 2,923 | 2,353 | 2,132 |

Tableau des valeurs.

La définition de $s(\hat{b})$ est en *annexe 3*.

## 2. Recherche de points atypiques : indicateurs

### *Effet levier*

Les valeurs prédites $\hat{y}i$ se calculent à partir des valeurs mesurées yi selon $\hat{Y} = \hat{H}\ Y$, avec la « hat matrice » $\hat{H} = X(X'X)^{-1}X'$.
Par construction, les valeurs hij sont totalement indépendantes des valeurs de Y, et ne dépendent que des valeurs de x, c'est-à-dire de nos conditions expérimentales et non pas des résultats que nous avons obtenus. Néanmoins :

$$\hat{y}i = hii\ yi + \sum_{j \neq i} hijyj$$

Les termes diagonaux traduisent l'influence de la ième mesure sur sa propre estimation. Plus $h_{ii}$ est grand (on démontre par ailleurs qu'ils sont tous positifs inférieurs à 1), plus la valeur de yi a du poids pour l'estimation de $\hat{y}i$ et donc pour l'estimation de la droite de régression : le point devient influent.
La valeur limite de $h_{ii}$ correspondant à un point influent fait débat. Un point est dit avoir un effet levier dans le cas d'une régression linéaire simple si :
- $h_{ii} > 4/n$ selon Hoaglin et Welsh, 1978.
- $h_{ii} > 0,5$ selon Huber, 1981.

Quelle que soit la limite que l'on choisisse, on s'aperçoit rapidement que les points ayant le plus fort effet levier, donc les plus influents, sont les points de xi extrêmes, éloignés de la moyenne $\bar{x}$ (d'où le nom d'effet levier). Ils doivent donc être soumis à une attention particulièrement soutenue, notamment en termes d'écart type.

### *Résidus, résidus normalisés ou studentisés ?*

Parmi les hypothèses de base de la régression linéaire, les erreurs ne sont pas corrélées et ont la même variance (homoscédasticité). Mais nous n'avons pas connaissance des erreurs . Nous avons accès à leurs réalisations $\hat{\varepsilon}$ lors des différentes expériences.
Or $\hat{\varepsilon} = Y - \hat{Y} = Y - \hat{H}\ Y = InY - \hat{H}\ Y = (In - \hat{H})Y$
D'où il s'ensuit que $V(\hat{\varepsilon} = {}^{2}(In - \hat{H})$
Soit $V(\hat{\varepsilon}\, i) = {}^{2}(1\text{-}h_{ii})$
Ce qui signifie, d'une part l'hétéroscédasticité, mais aussi et surtout que les résidus dépendent de la position du point sur la droite. Sachant que les points extrêmes ont un effet levier plus important que les autres, leurs résidus sont systématiquement sous-estimés.
L'unique résidu suivant une loi connue, et permettant donc de fixer une quelconque limite (souvent estimée par commodité à tort égale à 2) n'est ni le résidu normalisé $t = \frac{\hat{\varepsilon}\, i}{\hat{\sigma}\sqrt{(1-hii)}}$, car nous ne connaissons pas ; ni le résidu studentisé $r = \frac{\hat{\varepsilon}\, i}{\hat{\sigma}\sqrt{(1-hii)}}$ qui ne suit toujours pas une loi de Student ; mais le « résidu studentisé par validation croisée », ou « résidu studentisé supprimé » $t^* = \frac{\hat{\varepsilon}\, i}{\hat{\sigma}(i)\sqrt{(1-hii)}} \rightsquigarrow \mathcal{T}_{n-3}$ qui est le seul de tous ces résidus à suivre une loi de Student [2].
Ce résidu est obtenu par validation croisée, selon la technique du « leave-one-out » et est proposé par tous les logiciels de statistiques classiques. $\hat{\sigma}(i)$ est l'estimation de sur la droite dont le point i considéré a été retiré, d'où le fait que t* suive une loi $\mathcal{T}_{(n-1)-2} = \mathcal{T}_{n-3}$.
Cette technique nous conduit à une remarque importante : il est toujours nécessaire de confronter le chiffre obtenu à la réalité de la droite, ainsi que l'atteste l'exemple ci-dessous :

| x | y | T SUPPR |
|---|---|---|
| 1 | 2 | - 0,5 |
| 2 | 4 | - 0,4 |
| 3 | 6,05 | 633714,9 |
| 4 | 8 | - 0,4 |
| 5 | 10 | - 0,5 |

Ce cas limite peut faire prendre conscience que l'étude de la régression par ses différents indicateurs ne s'impose pas nécessairement lorsque les points sont presque parfaitement alignés.

### *Distance de Cook et DFFITS*

Est-il possible de raisonnablement estimer qu'un point est aberrant, et donc l'enlever sans remords ?
Deux indicateurs existent, qui rassemblent les deux informations « résidu studentisé supprimé » et « effet levier » importants, et indiquent l'influence globale d'un point sur la régression. Ces deux indicateurs sont proposés également dans les logiciels de statistique :
- la distance de Cook, dont la limite diverge selon les sources ;
- le DFFITS, dont la limite inférieure consensuelle est fixée à $\frac{3,46}{\sqrt{n}}$.

| Indicateur | | 4 points | 5 points | 6 points |
|---|---|---|---|---|
| Effet levier h | Hoaglin et Welsh | 1 | 0,8 | 0,67 |
| | Huber | 0,5 | 0,5 | 0,5 |
| Résidu studentisé supprimé \|t*\| | $t_{n-3,0,975}$ | 12,71 | 4,30 | 3,18 |
| Distance de Cook D | Cook | 1 | 0,88 | 0,82 |
| | 4/n | 1 | 0,8 | 0,67 |
| DFFITS | $\frac{3,46}{\sqrt{n}}$ | 2 | 1,79 | 1,63 |

Tableau - Valeurs inférieures des différents indicateurs.
Tous ces indicateurs peuvent être calculés par Excel [5].

### 3. Justifications mathématiques

#### *Écart type*

L'écart type est défini par $s^2 = V$, donc $s^2(\hat{a}) = V(\hat{a})$ et $s^2(\hat{b}) = V(\hat{b})$.

#### *Coefficient de détermination*

$$\underset{SCT}{\sum_{i=1}^{n}(yi-\bar{y})^2} = \underset{SCR}{\sum_{i=1}^{n}(yi-\hat{y}i)^2} + \underset{SCE}{\sum_{i=1}^{n}(\hat{y}i-\bar{y})^2}$$

Par définition, le coefficient de détermination $R^2 = \frac{SCE}{SCT} = \frac{\sum_{i=1}^{n}(\hat{y}i-\bar{y})^2}{\sum_{i=1}^{n}(yi-\bar{y})^2}$.

On démontre que, dans le cas de la régression linéaire simple $r^2 = R^2$.

#### *F de Fisher*

$$F = \frac{\frac{SCE}{1}}{\frac{SCR}{n-2}} = \frac{SCE}{SCT-SCE}(n-2) = \frac{R^2}{1-R^2}(n-2)$$

Le test de significativité de Fisher Snedecor ne présente aucun intérêt, la valeur critique de $F_{(1,4,0.95)}$ valant 7,71, ce qui est ridiculement bas dans notre cas, puisque la linéarité est pratiquement assurée.

### Références


[1] P.A. Cornillon, E. Matzner Lober, *Régression, théorie et applications*, Springer, **2007**, p. 4-21.
[2] F. Grégis, La valeur de l'incertitude: l'évaluation de la précision des mesures physiques et les limites de la connaissance expérimentale, Thèse de doctorat, Paris 7, **2016**, chap. 5.
[3] P. Dagnelie, *Statistique théorique et appliquée T1. Statistique descriptive et bases de l'inférence statistique*, de Boek, **2011**, p. 145-149.
[4] https://metrologie-francaise.lne.fr/sites/default/files/media/document/rfm41-1601.pdf (consulté le 25/11/2021).
[5] http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/fr_Tanagra_Regression_Excel.pdf (consulté le 15/11/2021)


# Régression linéaire simple (partie 2)

## Variante de la pratique courante : prévision

L'élaboration d'une droite d'étalonnage en chimie s'effectue par la méthode des moindres carrés ordinaires. Cette droite d'étalonnage a souvent pour but de déterminer une concentration inconnue, c'est-à-dire d'effectuer une prédiction sur une valeur inconnue. Nous avons proposé dans l'article précédent une variante permettant d'obtenir une meilleure linéarité, cet article propose désormais une variante permettant de déterminer plus simplement un intervalle de prédiction pour la concentration inconnue.

La méthode classique consiste à tracer une droite d'étalonnage Y = f(C), C étant la concentration d'une gamme étalon préparée par le laboratoire et Y étant la mesure : absorbance, surface, intensité… ; puis, à partir d'une valeur mesurée de Yinconnue (Yinc), en déduire une valeur prédite de Cinconnue (Cinc) en inversant la formule obtenue :

$$Yinc = \hat{b} + \hat{a}Cinc \quad (1)$$

L'exemple choisi ici sera celui, classique, d'une droite d'étalonnage effectuée par mesure d'absorbance A en spectrophotométrie UV-visible.

La théorie qui présente les valeurs prédites est bien connue [1]. Néanmoins, une question se pose : fondamentalement, ce qui nous intéresse, c'est la concentration inconnue Cinc. Alors, pourquoi ne pas tracer C = f(A) directement ?

Cette approche n'est pas contraire aux définitions du VIM :

Étalonnage : « *Opération qui, dans des conditions spécifiées, établit en une première étape une relation entre les valeurs et les incertitudes de mesure associées qui sont fournies par des étalons et les indications correspondantes avec les incertitudes associées, puis utilise en une seconde étape cette information pour établir un résultat de mesure à partir d'une indication.* »

Mesurande : « *Grandeur que l'on veut mesurer.* »

Or, dans notre cas, la grandeur que l'on souhaite mesurer est la concentration de la solution inconnue, et la relation établie entre les valeurs et les indications devient alors C = aA + b.
La théorie mathématique qui propose de tracer x = f(y) est bien connue et consiste [2], par l'intermédiaire d'une matrice Hessienne, à minimiser la quantité $S'(a,b) = \sum( x_i - a'yi - b')^2$, ce qui revient à minimiser une somme des carrés des écarts à la droite horizontale, et non plus verticale. Le raisonnement reste encore focalisé sur A.
L'optique présentée dans cet article est totalement différente. Elle propose de considérer les couples de valeurs obtenus d'un point de vue statistique, et à envisager de tracer directement C = f(A). L'intérêt majeur consiste à accéder sans calcul parfois erroné à l'incertitude sur la concentration, ainsi que nous allons le voir. Trois arguments d'inégales importances amènent à ce changement de point de vue.
Dans toute la suite, la méthode $M_1$ est la méthode classique qui consiste à déterminer l'équation Ymes = f(Ximposé), puis à l'inverser pour déterminer Xinc à partir de Yinc mesuré.
La méthode $M_2$ est la méthode alternative : tracer directement X = f(Y) et obtenir directement Xinc à partir de Yinc mesuré.

**Non inversibilité de la formule (1)**

Déduire la concentration inconnue de la droite revient à effectuer l'inversion :

$\hat{y}inc = \hat{b} + \hat{a}xinc \leftrightarrow \text{xinc} = \frac{\hat{y}inc - \hat{b}}{\hat{a}}$.

Or, cette équivalence, vraie pour des points parfaitement alignés, *ie* un $R^2$ égal à un, l'est de moins en moins lorsque $R^2$ diminue. Par exemple, pour les deux droites issues de $(x,y_1)$ et $(x,y_2)$ :

| x | $y_1$ | $y_2$ |
|---|---|---|
| 1 | 3,1 | 3,9 |
| 2 | 4,9 | 4,1 |
| 3 | 7,3 | 8 |
| 4 | 8,9 | 8,1 |
| 5 | 11,2 | 12 |

Les résultats, qui se passent de commentaires, sont :

| $R^2$ | 0,997 | 0,907 |
|---|---|---|
| Obtenu par $M_1$ | $x = - 0,505 + 0,4950\ y_1$ | $x = - 0574 + 0,4950\ y_2$ |
| Obtenu par $M_2$ | $x = - 0,494 + 0,4930\ y_1$ | $x = - 0,242 + 0,4490\ y_2$ |

En fin de compte, refuser d'inverser A = εIC revient à la situation ubuesque de préférer inverser une relation affine qui ne l'est pas plutôt que d'inverser une relation qui l'est…

**Première hypothèse de base de la régression linéaire non respectée**

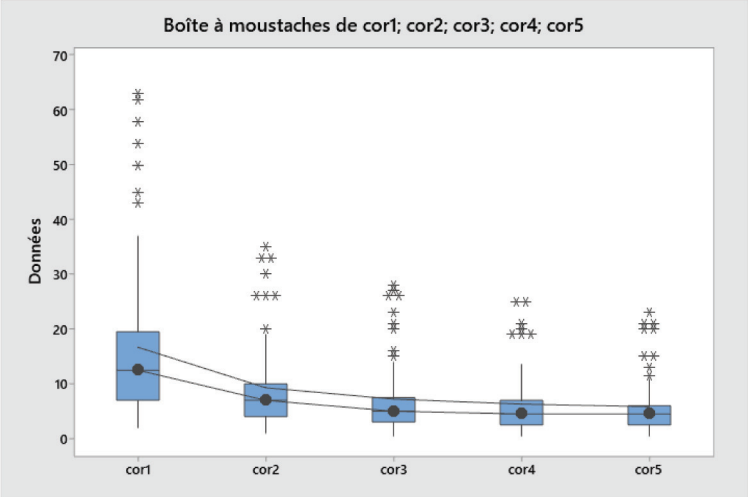
Cette section ne concerne pas le cas d'étalons déjà dilués fournis par une entreprise spécialisée, mais celui courant d'une gamme d'étalonnage préparée en laboratoire.
La réalisation d'une régression linéaire consiste, à partir de n couples de valeurs $(x_i,y_i)$, à minimiser une quantité qui ne porte que sur les $y_i$, les $x_i$ étant considérés par hypothèse constant et fixe.
Les valeurs $x_i$ et $y_i$ sont les réalisations de deux variables aléatoires $X_i$ et $Y_i$, que nous supposerons suivant une loi normale. La relation devient $Y_i = aX_i + b + \varepsilon_i$, $\varepsilon_i$ étant nommée erreur, variable aléatoire qui rend compte du fait que $X_i$ et $Y_i$ le sont également.
La première hypothèse est que $x_i$ n'est pas une variable aléatoire. Il représente une valeur que nous avons imposée. Le modèle de la régression linéaire devient : $Y_i = ax_i + b + \varepsilon_i$
Les deux seules variables aléatoires sont alors $Y_i$ et $\varepsilon_i$, et donc :
- Première hypothèse : les xi étant parfaitement déterminés, $\varepsilon_i \rightsquigarrow \mathcal{N}( 0 , \sigma i^2 )$.
- Deuxième hypothèse : les erreurs sont décorrélées et de même variance. $\varepsilon_i \rightsquigarrow \mathcal{N}( 0 , \sigma^2 )$ : elles sont dites iid (indépendantes identiquement distribuées). ( cf *annexe 1*).
La question est donc : peut-on réellement considérer que l'incertitude relative sur les concentrations des solutions étalons est nulle, ou au pire largement inférieure à celle des réponses mesurées ?
Pour répondre à cette question, trente-trois étudiants de Brevet de Technicien Supérieur Métiers de la Chimie en fin de formation ont réalisé l'expérience suivante :
Une gamme étalon de cinq concentrations croissantes de dichromate de potassium, chaque modalité ayant été répliquée trois fois, soit quinze fioles de 50 mL (toutes verreries de classe A) à préparer à partir d'une solution préalablement étalonnée de dichromate de potassium de concentration Cmère = 0,010390 ± 0,000030 $mol.L^{-1}$. L'absorbance a ensuite été mesurée à 450 nm dans un spectrophotomètre UV-visible préalablement étalonné en absorbance et en longueur d'onde.

| Vmère (mL) | 2 | 4 | 6 | 8 | 10 |
|---|---|---|---|---|---|
| Cf ($mol.L^{-1}$) | 0,000416 | 0,000831 | 0,001247 | 0,001662 | 0,002080 |

La dilution a été effectuée à l'eau distillée.

Les incertitudes sur l'absorbance ont été maximisées à ± 0,0020, ce qui représente sur une échelle d'absorbances mesurées de 0,1 à 1 une incertitude relative allant de 2 % (pour les plus petites valeurs mesurées) à 0,2 % (pour les plus grandes) : cette valeur d'incertitude est compatible avec les tolérances des appareils de moyenne gamme de prix.
Les incertitudes sur les concentrations ont été déterminées en utilisant la loi de propagation des incertitudes appliquée à $C = \frac{A-b}{a}$ :

$$uc^2 = \frac{uA^2}{a^2} + \frac{ub^2}{a^2} + \left[\frac{(A-b)}{a^2}\right]^2 ua^2$$




| | Moyenne | Q1 | Médiane | Q3 |
|---|---|---|---|---|
| Concentration 1 | 16,7 | 7 | 12,5 | 19,5 |
| Concentration 2 | 9,3 | 4 | 7 | 10 |
| Concentration 3 | 7,2 | 3 | 5 | 7,5 |
| Concentration 4 | 6,3 | 2,5 | 4,2 | 7 |
| Concentration 5 | 5,8 | 2,5 | 3,9 | 6 |

Figure 3 - Erreurs relatives en % pour les cinq concentrations de dichromate de potassium.

Les incertitudes relatives sur les cinq concentrations ne suivent pas une loi normale et présentent toutes une asymétrie à droite (cf *figure 3*). Néanmoins, les valeurs obtenues sont largement au-dessus des incertitudes relatives aux absorbances.
Les résultats obtenus par les étudiants sont compilés selon le graphe classique *figure 2*.

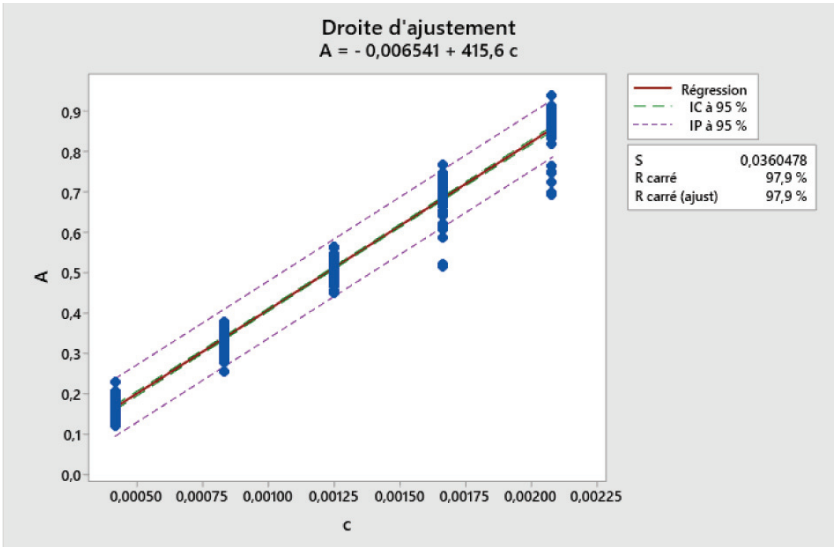



Figure 2 - A = f(C) sur 71 manipulations.

Certains des lecteurs de l'article précédent auront pu s'étonner de la désinvolture avec laquelle la seconde hypothèse de base (homoscédasticité des résidus) a été mise de côté.
C'est normal : le graphe obtenu *figure 2* n'a aucun sens, la première hypothèse de base n'étant pas respectée : non seulement les concentrations ne sont pas constantes, mais en plus leurs incertitudes sont supérieures !

L'impression d'une hétérogénéité des mesures, par exemple sur la cinquième concentration, est davantage due à l'incertitude sur la concentration que sur la mesure de l'absorbance. Par exemple, si on propage l'erreur en multipliant par la pente (dont on néglige l'incertitude) qui vaut environ 415, l'incertitude sur l'absorbance pour la cinquième concentration est de l'ordre de 415 x 3,9 % x 0,00208 = 0,034 environ, ce qui est largement supérieur à la tolérance de 0,0020 de l'appareil. Chaque modalité ne devrait pas être représentée par un segment de droite vertical mais par un rectangle étiré horizontalement.
Par ailleurs, les étudiants ont refait une deuxième série d'expériences sur du sulfate de nickel (les conditions sont explicitées dans le premier article). Les résultats obtenus sur l'incertitude sur les concentrations sont sensiblement les mêmes, supérieures d'environ 0,5 point. Ce résultat est extrêmement troublant : piqués au vif par leurs résultats précédents, encadrés de très près, ils se sont véritablement concentrés, et la précision devrait être nettement meilleure.
Comment expliquer ces mauvais résultats ? Les valeurs d'absorbance mesurées sont dans une plage de 0,05 et 0,2, la pente est faible, de l'ordre de 5 : les incertitudes sur la mesure de l'absorbance prennent une importance relative plus importante et les $R^2$ sont plus faibles. L'inversion $C = \frac{A-b}{a}$ est moins vraie, et donc les résultats sont plus sujets à caution.
Il est à noter que, après avoir déterminé les incertitudes sur les concentrations, une solution alternative possible à la régression linéaire simple serait d'effectuer une régression dite orthogonale de Deming, qui tient compte des incertitudes sur x et y [3], mais elle nécessite l'utilisation d'un logiciel de statistique adapté, et est donc peu réalisable en pratique dans un laboratoire non pourvu d'un tel logiciel.

**Difficulté pour déterminer une incertitude sur xinc**

Un dernier argument décisif plaide en la faveur d'un tracé C = f(A). En traçant A = f(C), il est mathématiquement compliqué de déterminer une incertitude sur C, ce qui est métrologiquement inacceptable.
Il est courant d'imaginer un intervalle des concentrations qui prendrait en compte l'intervalle de confiance (*figure 4*).

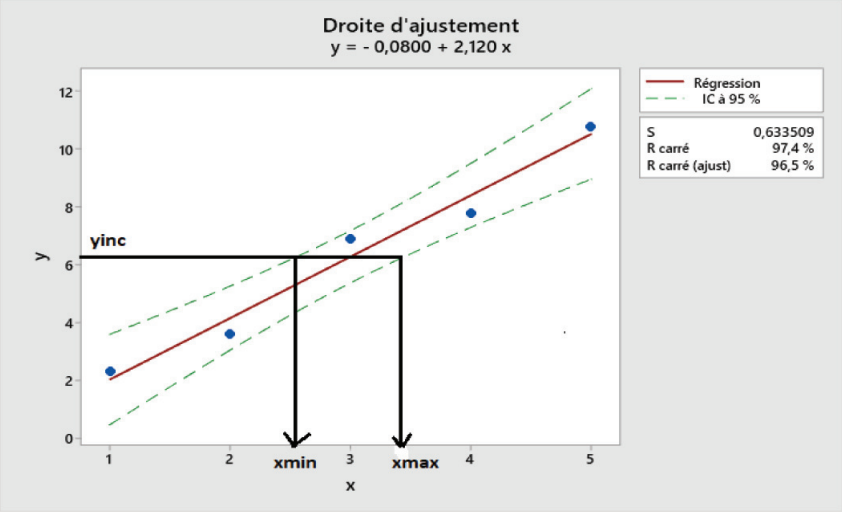



Figure 4 - Intervalle des concentrations en utilisant l'intervalle de confiance à 95 %.

Ce n'est pas le cas : pour une mesure prédite, l'intervalle de confiance est plus grand : il s'agit de l'intervalle de prévision à 95 % de confiance (cf *annexe 1*). Ainsi, on devrait plutôt imaginer la détermination de l'incertitude selon la *figure 5* :

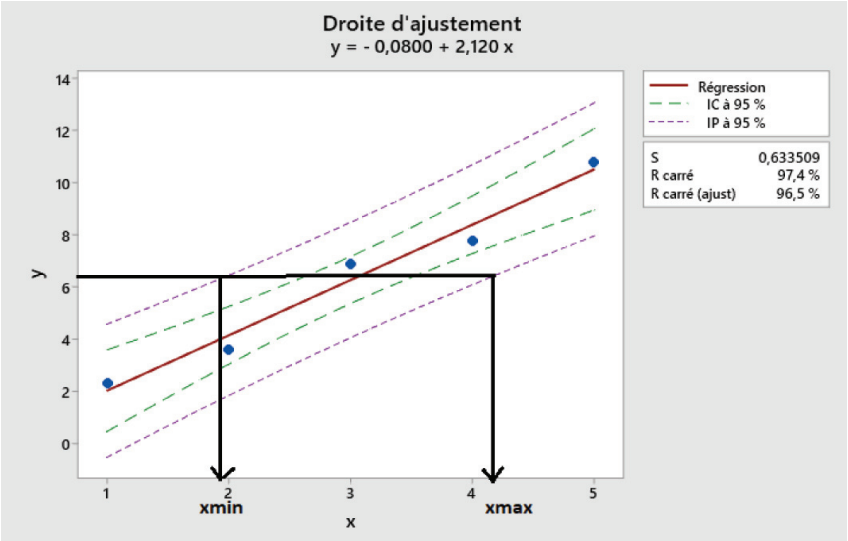



Figure 5 - Intervalle des concentrations en utilisant l'intervalle de prévision à 95 %.

Cette image est fausse : encore une fois, toute l'incertitude est sur y, et pas sur x : on ne peut raisonner que verticalement.
Une solution consiste à utiliser l'intervalle de prévision déterminé en *annexe 1* :

$$[\hat{y}_{n+1} - t_{n-2,0.975}\sqrt{\hat{\sigma}^2\left[1+\frac{1}{n}+\frac{(xn+1-\bar{x})^2}{\sum_{i=1}^{n}(xi-\bar{x})^2}\right]}\ ;\ \hat{y}_{n+1} + t_{n-2,0975}\sqrt{\hat{\sigma}^2\left[1+\frac{1}{n}+\frac{(xn+1-\bar{x})^2}{\sum_{i=1}^{n}(xi-\bar{x})^2}\right]} \quad (2)$$

En utilisant $\hat{y}_{n+1} = \hat{b} + \hat{a}x_{n+1}$, que l'on inverse, on trouve pour la concentration inconnue un intervalle de prévision de :

$$\left\{\frac{\hat{y}n+1-\hat{b}}{\hat{a}} - \frac{t_{n-2,0.975}}{\hat{a}}\sqrt{\hat{\sigma}^2\left[1+\frac{1}{n}+\frac{1}{\hat{a}^2}\frac{(yn+1-\bar{y})^2}{\sum_{i=1}^{n}(xi-\bar{x})^2}\right]}\ ;\right.$$

$$\left.\frac{\hat{y}n+1-\hat{b}}{\hat{a}} + \frac{t_{n-2,0.975}}{\hat{a}}\sqrt{\hat{\sigma}^2\left[1+\frac{1}{n}+\frac{1}{\hat{a}^2}\frac{(yn+1-\bar{y})^2}{\sum_{i=1}^{n}(xi-\bar{x})^2}\right]}\right\} \quad (3)$$

Les deux méthodes $M_1$ et $M_2$ ont été comparées (cf *annexe 2*), utilisant l'intervalle de prévision (3) pour $M_1$ et (2) pour $M_2$ :
Les résultats (*figure 6*) ont été obtenus sur 85 couples de cinq valeurs, les données étant brutes, les valeurs atypiques n'ayant été retirées ni en utilisant le résidu studentisé supprimé ni la distance de Cook (seuls ont été retirés les résultats franchement aberrants).
Sont notées Cmoy (M1), Cmin (M1) et Cmax (M1) et Cmoy (M1) (resp. M2) les valeurs moyennes ; minimales et maximales de l'intervalle de prévision à 95 % obtenus par la méthode classique M1 (resp. M2). Sont notées Cmoy, Cmin et Cmax les valeurs moyennes obtenues par les deux méthodes.
Étant donné que les valeurs sont brutes, elles ont été tracées en fonction de $R^2$ prédit et non de $R^2$ (cf *annexe 2*), ainsi qu'il est conseillé par la plupart des logiciels de statistique.

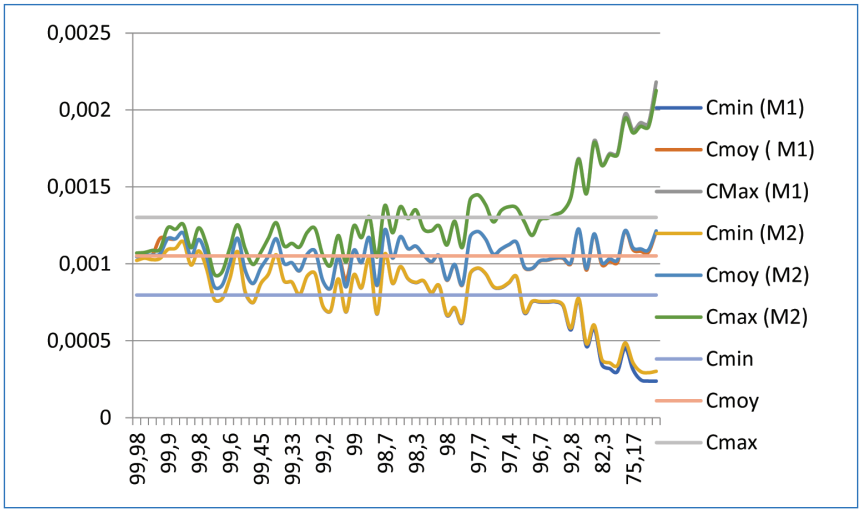



Figure 6 - Comparaison des méthodes $M_1$ et $M_2$ en fonction de $R^2$ prédit.

Les valeurs moyennes ne présentent pas de différence significative (excepté pour une seule régression comprenant un unique point nettement aberrant, qui aurait été normalement retiré).
Il en de même pour les valeurs minimales et maximales… tant que $R^2$ prédit n'est pas trop faible. Cela se retrouve *figure 7* où sont regroupées l'étendue de l'incertitude Cmax-Cmin ($M_1$) (resp. $M_2$) obtenues par la méthode $M_1$ (resp. $M_2$) ainsi que la différence delta entre les deux étendues (axe secondaire à droite).

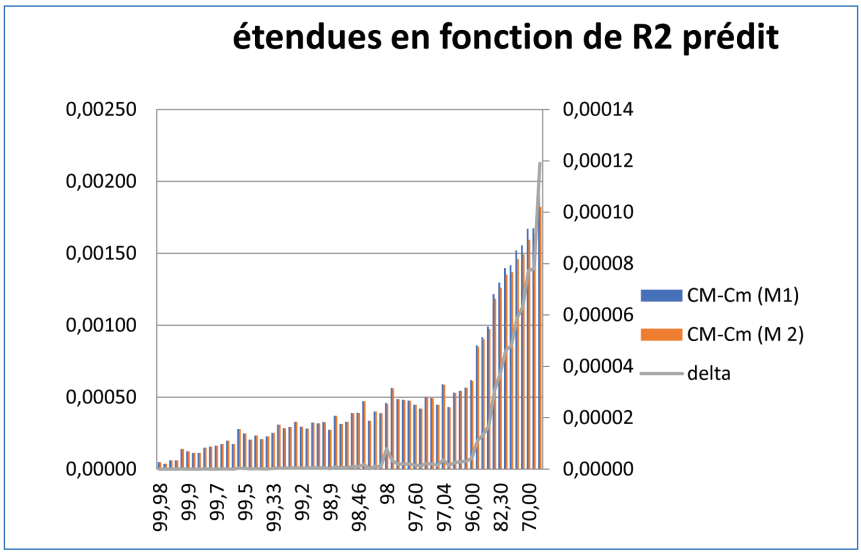



Figure 7 - Comparaison des deux étendues obtenues par les méthodes $M_1$ et $M_2$ en fonction de $R^2$ prédit.

La courbe verte delta montre qu'il n'y a aucune différence entre les deux méthodes, tant que le $R^2$ prédit conserve une valeur supérieure à environ 0,96. En dessous de cette valeur, la différence s'accroît et la méthode $M_2$ amène à une étendue de l'incertitude nettement plus faible. Cette différence apparaît entre les deux méthodes lorsque le $R^2$ diminue, c'est-à-dire lorsque l'inversion de la relation linéaire entre C et A est sujette à caution.
La supériorité de la méthode $M_2$ peut sembler à relativiser, elle n'apparaît que pour des valeurs de $R^2$ prédit telles que les intervalles de prévision soient déraisonnables, avec une incertitude sur la valeur de la concentration prédite au moins de l'ordre de 30 %. Néanmoins, il est acceptable, si la valeur inconnue a été répliquée, de considérer que la valeur Ainc est égale à la moyenne de ces répliques : l'intervalle à prendre en compte est celui de confiance à 95 %, bien plus étroit que l'intervalle de prévision à 95 % (intervalles en vert sur les *figures 2* et *5*) : la supériorité de la méthode $M_2$ est alors avérée.

### $R^2$ ou $R^2$ prédit ?

Si on trace le même graphique que celui de la *figure 7* en fonction de $R^2$, on obtient sensiblement le même résultat (*figure 8*), la différence entre les deux méthodes n'apparaissant qu'à partir d'un $R^2$ sensiblement égal à 0,99. C'est la limite en deçà de laquelle $R^2$ semble acceptable pour cette manipulation, ainsi qu'une simulation Monte Carlo l'a montré [8].

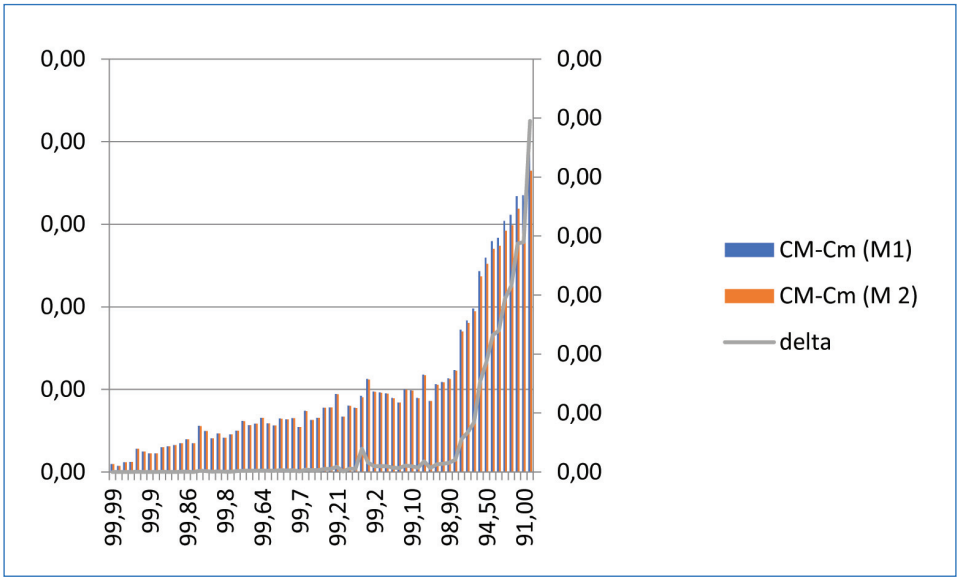



Figure 8 - Comparaison des deux étendues obtenues par les méthodes M1 et M2 en fonction de $R^2$.

La décroissance de $R^2$ prédit en fonction de celle de $R^2$ est assez brutale (*figure 9*). L'apparente linéarité est trompeuse : il vaut mieux, si une étude rigoureuse des points atypiques n'a pas été effectuée, utiliser le $R^2$ prédit qui tient compte des effets leviers, ce qui le rend plus sensible aux points extrêmes influents (cf *annexe 2*).

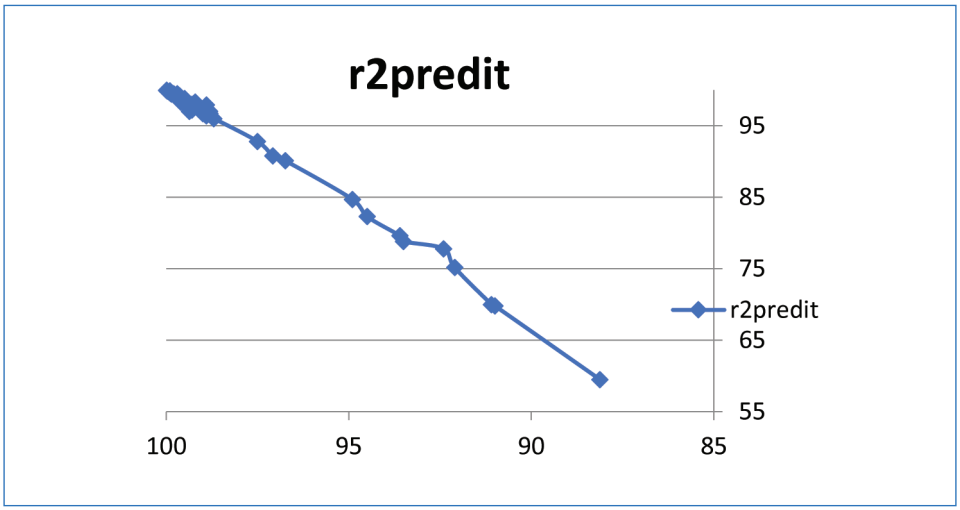



Figure 9 - $R^2$ prédit en fonction de $R^2$.

### A = f(C) ou C = f(A) ?

La méthode qui consiste à tracer C = f(A) conduit à des résultats au pire identiques, au mieux meilleurs. Elle semble donc supérieure car plus facile à mettre en œuvre, un simple programme sur Excel donnant aisément l'incertitude sur la concentration qui doit être déterminée, ce qui fait souvent défaut en pratique.

## Annexes

### 1. Intervalle de prévision

$$Y_{n+1} = ax_{n+1} + b + \varepsilon_{n+1}$$
$$\hat{Y}_{n+1} = \hat{a}x_{n+1} + \hat{b}$$
$\hat{\varepsilon}_{n+1} = Y_{n+1} - \hat{Y}_{n+1}$, et $V(\hat{\varepsilon}_{n+1}) = V(Y_{n+1}) + V(\hat{Y}_{n+1})$
$$V(\hat{\varepsilon}_{n+1}) = \sigma^2\left[\frac{1}{n} + \frac{(xn+1-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right] + \sigma^2 = \sigma^2\left[1 + \frac{1}{n} + \frac{(xn+1-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right]$$

L'erreur est supérieure pour une valeur prédite $\Longrightarrow$ intervalle de prédiction.
Si l'erreur suit une loi normale :

$$Y_{n+1} - Y_{n+1} \rightsquigarrow \mathcal{N}\left(0, \sigma^2\left[1 + \frac{1}{n} + \frac{(xn+1-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right]\right)$$

Alors :

$$\frac{yn+1 - yn+1}{\sqrt{\hat{\sigma}^2\left[1+\frac{1}{n}+\frac{(xn+1-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right]}} \rightsquigarrow \mathcal{T}_{n-2}$$

$y_{n+1}$ est donc, à 95 % de confiance, dans l'intervalle :

$$[\hat{y}_{n+1} - t_{n-2,0.975}\sqrt{\hat{\sigma}^2\left[1 + \frac{1}{n} + \frac{(xn+1-\bar{x})^2}{\sum_{i=1}^{n}(xi-\bar{x})^2}\right]}\,;\, y_{n+1} + t_{n-2,0975}\sqrt{\hat{\sigma}^2\left[1 + \frac{1}{n} + \frac{(xn+1-\bar{x})^2}{\sum_{i=1}^{n}(xi-\bar{x})^2}\right]}$$

L'intervalle de prédiction est donc plus grand que l'intervalle de confiance.

### 2. $R^2$ prédit

$$\underset{\text{SCT}}{\sum_{i=1}^{n}(yi-\bar{y})^2} = \underset{\text{SCR}}{\sum_{i=1}^{n}(yi-\hat{yi})^2} + \underset{\text{SCE}}{\sum_{i=1}^{n}(\hat{yi}-\bar{y})^2}$$

$$R^2 = \frac{SCE}{SCT} = \frac{\sum_{i=1}^{n}(\hat{y}i-\bar{y})^2}{\sum_{i=1}^{n}(yi-\bar{y})^2} = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT}$$

$$R^2 \text{ prédit} = 1 - \frac{PRESS}{SCT}$$

Avec $\text{PRESS} = \sum_{i=1}^{n}(yi - \hat{y}_{i,-i})^2$
Avec $\hat{y}_{i,-i}$ l'estimation de yi en enlevant le ième point selon la technique du « leave-one-out ». Le PRESS peut néanmoins se calculer simplement car on démontre [7] que :

$$\hat{y}_{i,-i} = \frac{hii}{1-hii}\hat{yi} - \frac{1}{1-hii}yi$$

Donc $i\,\hat{y}_{i,-i} - yi = \frac{\hat{y}i - yi}{1-hii}$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}i-yi)^2}{\sum_{i=1}^{n}(yi-\bar{y})^2} \text{ et } R^2 \text{ prédit} = 1 - \frac{PRESS}{\sum_{i=1}^{n}(yi-\bar{y})^2} \text{ avec PRESS} = \sum_{i=1}^{n}\left[\frac{\hat{y}i-yi}{1-hii}\right]^2$$

En fin de compte, il s'agit pratiquement de la même formule, le $R^2$ prédit étant nécessairement inférieur à $R^2$ puisque les points sont en quelque sorte pondérés par leur effet levier. Il peut y avoir une grande différence si le point influent a un grand effet levier.

### Références


[1] P.A. Cornillon, E. Matzner Lober, *Régression, théorie et applications*, Springer, **2007**, p 4-21
[2] A. Gamella-Mathieu, *Mathématiques du signal et statistiques*, Ellipses, **2019**, p. 419-423.
[3] P. Dagnelie, *Statistique théorique et appliquée. Tome 2 : Inférence statistique à une et à deux dimensions*, de Boek, **2011**, p. 467-471.
[5] http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/fr_Tanagra_Regression_Excel.pdf, (consulté le 15/11/2021).
[6] https://metrologie-francaise.lne.fr/sites/default/files/media/document/rfm41-1601.pdf
[7] www.lpsm.paris/pageperso/guyader/files/teaching/Regression.pdf, p. 87 (consulté le 20/12/2021).
[8] C. Roussel, A. Roussel, « À la défense du $R^2$ », *Bulletin de l'Union des physiciens*, à paraître.