BLAISE PASCAL
PT 2018-2019

Outils pour la physique

Régression linéaire

Au programme

Extrait du programme officiel de PTSI : partie « Formation expérimentale », bloc 1 « Mesures et incertitudes ».

Notions et contenus	Capacités exigibles
Vérification d'une loi physique ou validation d'un modèle ; ajustement de données expérimentales à l'aide d'une fonction de référence modélisant le phénomène.	Utiliser un logiciel de régression linéaire. Expliquer en quoi le coefficient de corrélation n'est pas un outil adapté pour juger de la validité d'un modèle linéaire. Juger qualitativement si des données expérimentales avec incertitudes sont en accord avec un modèle linéaire. Extraire à l'aide d'un logiciel les incertitudes sur la pente et sur l'ordonnée à l'origine dans le cas de données en accord avec un modèle linéaire.

En **gras**, les points devant faire l'objet d'une approche expérimentale.

Plan du cours

Étienne Thibierge, 10 août 2018, www.etienne-thibierge.fr

I - Objectifs

Imaginons qu'un modèle théorique prédise que deux grandeurs physiques x et y sont reliées par une loi s'écrivant mathématiquement sous la forme $y = f(x)$ avec f une fonction quelconque connue. Deux situations expérimentales se rencontrent fréquemment :

- ▷ on cherche à savoir si le modèle est juste, donc à savoir si la relation $y = f(x)$ est bien vérifiée ;
- ▷ on sait (ou on suppose) que le modèle est juste, mais la fonction f dépend d'un paramètre que l'on cherche à estimer.

Pour ce faire, on reproduit l'expérience dans N configurations différentes : on choisit plusieurs valeurs x_n ou y_n ($1 \leq n \leq N$) et on mesure la valeur de y_n ou x_n qui correspond.

> ***Exemples :***
> - ▷ *Les fréquences propres de la corde de Melde sont données par $f_n = nc/2L$, et en particulier la fréquence du fondamental vaut $f_1 = c/2L$. On cherche à vérifier si la fréquence du fondamental $f_1 = y$ est bel et bien proportionnelle à $1/L = x$. Pour cela, on modifie la longueur de la corde et on mesure à chaque fois la fréquence du fondamental.*
> - ▷ *La loi d'Ohm indique que la tension U et l'intensité I traversant une résistance R sont reliées par $U = RI$. On cherche la valeur de R. On peut alors modifier la valeur de la tension et mesurer à chaque fois l'intensité.*

II - Principe de la régression linéaire

Dans ce paragraphe on se focalise exclusivement sur le cas où la fonction f est affine :

$$y = f(x) = ax + b\,.$$

Ce cas particulier est important non seulement pour lui-même, mais surtout car on cherchera à s'y ramener systématiquement, comme on le verra au paragraphe IV. L'objectif de la régression linéaire est donc de trouver les valeurs de a et b qui correspondent le mieux aux mesures, éventuellement pour les comparer aux prédictions du modèle.

II.1 - Traduction graphique

Graphiquement, une fonction affine est représentée par une droite dont a est la **pente** (ou **coefficient directeur**) et b l'**ordonnée à l'origine**. La recherche des valeurs de a et b se traduit donc graphiquement par la recherche de « la meilleure droite », c'est-à-dire celle qui passe au plus près des points expérimentaux dont les coordonnées sont (x_n, y_n).

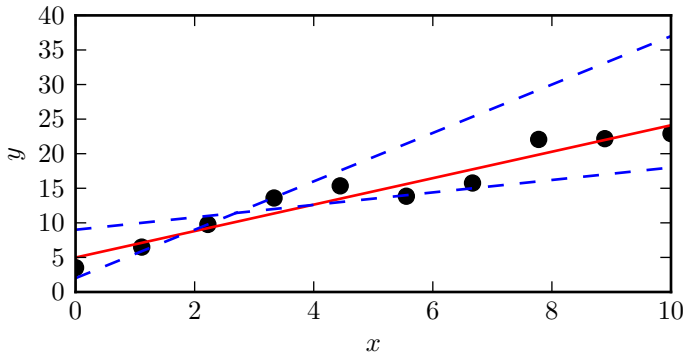


Figure 1 – **Illustration de la notion de meilleure droite.** La droite en traits pleins est celle qui passe au plus près des points noirs. La droite pointillée de plus grande pente en est clairement éloignée. La droite pointillée de plus faible pente pourrait sembler raisonnable, mais tous les points sont en dessous pour $x < 5$ et au dessus pour $x > 5$: elle est donc moins bien que la droite en traits pleins.

Règle d'or de la régression linéaire :
La première chose à faire avant toute régression linéaire est donc de représenter ces points expérimentaux.

Le cerveau humain étant doué pour repérer les droites, cela permet en un coup d'œil de repérer si les données semblent bien se superposer sur une droite, ou plus souvent de détecter un éventuel « point aberrant », c'est-à-dire une valeur sur laquelle vous vous êtes trompé lors de la mesure.

Par ailleurs, si jamais vous aviez à faire une régression linéaire sans ordinateur [1], la meilleure droite peut être approximée à la règle et ses paramètres déterminés par simple lecture graphique sur papier millimétré.

1. Malheureusement, c'est systématiquement le cas aux TP de la banque PT.

II.2 - Méthode des moindres carrés

> ***Remarque :*** *Paragraphe hors-programme, donc pas à connaître.*

La méthode utilisée par les logiciels pour trouver les valeurs optimales de a et b est appelée méthode des moindres carrés. Elle repose sur la minimisation d'une fonction appelée **fonction d'écart**

$$C(a,b) = \sum_{n=1}^{N} \left[y_n - (ax_n + b)\right]^2 .$$

Dans cette expression y_n est la valeur expérimentale, et $ax_n + b$ la valeur prévue par le modèle de coefficients a et b. Le terme $\left[y_n - (ax_n + b)\right]^2$ représente alors schématiquement la distance entre le point expérimental et la droite modèle, la fonction d'écart $C(a,b)$ étant la somme de ces distances. Plus la valeur de $C(a,b)$ est faible, plus les points expérimentaux sont proches de la droite. Les valeurs optimales de a et b sont donc celles qui minimisent la valeur de $C(a,b)$. L'ordinateur (ou la calculatrice) exécute un algorithme qui renvoie les valeurs optimales de a et b.

II.3 - Validation du modèle

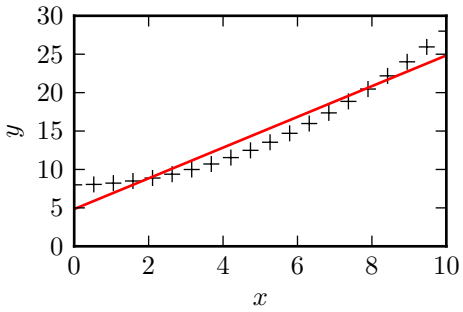
Attention ! Le fait que le logiciel renvoie des valeurs de a et b n'est certainement pas une garantie de succès : les calculs aboutissent toujours à quelque chose, même si la relation linéaire n'est pas du tout vérifiée.

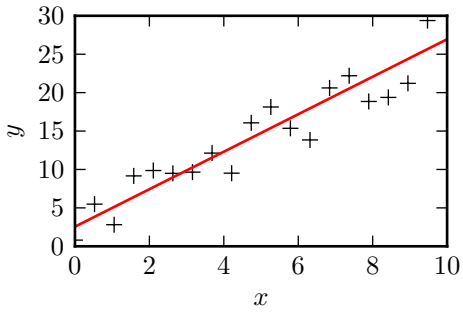
Au niveau CPGE, on retient :

L'unique façon de vérifier la validité d'une régression linéaire est de tracer les résultats de mesure et la courbe de régression et de les comparer visuellement.

Deux critères sont à vérifier :
- ▷ les points doivent être proches de la courbe ... ce qui peut être délicat à juger à cause des erreurs aléatoires ;
- ▷ il ne doit pas y avoir de courbure nette dans l'écart des points à la courbe.

> ***Exemples :***





- ▷ *Courbe de gauche : Les mesures ne sont donc pas compatible avec un modèle linéaire, on observe une nette tendance parabolique.*
- ▷ *Courbe de droite : Les points sont plus écartés de la courbe, mais il n'y a pas de tendance nette dans les écarts. Les mesures pourraient donc être compatibles avec un modèle linéaire : il manque les incertitudes pour vraiment conclure.*

II.4 - Coefficient de corrélation linéaire

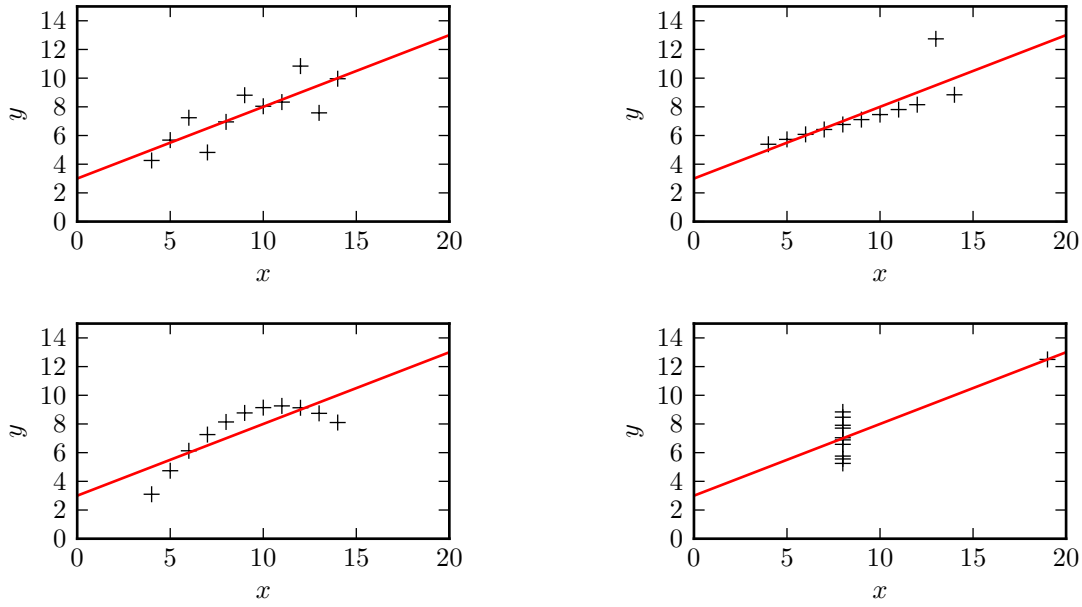
Pour quantifier l'écart entre les points expérimentaux et la droite de régression, les calculatrices et logiciels renvoient généralement la valeur du **coefficient de corrélation linéaire r**, ou parfois son carré r^2. Sa définition exacte est trop technique pour être donnée ici.

Le coefficient r est un nombre compris par définition entre -1 et 1, du même signe que a, qui mesure l'existence d'un lien linéaire entre x et y : plus r^2 est proche de 1, mieux les points sont alignés. La valeur $r^2 = 1$ signifie que tous les points expérimentaux sont exactement sur la droite. En pratique, pour un petit nombre de mesures comme en TP, il est fréquent d'obtenir $r^2 > 0{,}9$ même si les points expérimentaux sont clairement éloignés de la droite.

Une question légitime est de savoir si le coefficient de corrélation permet ou non de conclure à la validité du modèle. L'exemple suivant permet de répondre par la négative.

Exemple du quartet d'Anscombe :

Le quartet d'Anscombe est constitué de quatre ensembles de données qui ont les mêmes propriétés statistiques simples (moyennes, écarts types, équation de la régression linéaire, coefficient de corrélation linéaire) mais qui sont en réalité très différents, ce qui se voit facilement lorsqu'on les représente sous forme de graphiques. Ils ont été construits en 1973 par le statisticien Francis Anscombe dans le but de démontrer l'importance de tracer des graphiques avant d'analyser des données.



La droite de régression linéaire ne semble décrire correctement que le jeu de données représenté en haut à gauche.

Ainsi, fixer un critère arbitraire sur le coefficient de corrélation afin de valider ou pas le modèle (par exemple $r^2 > 0{,}99$) n'est pas pertinent. En effet, la présence d'un point aberrant (souvent issu d'une erreur de manipulation) ou de fortes erreurs aléatoires peut diminuer considérablement la valeur de r^2 même si la loi reste vérifiée. Réciproquement, la valeur du coefficient de corrélation linéaire peut demeurer très élevée en dépit d'une courbure nettement visible dans les données.

La seule façon valable de conclure à la validité d'une régression linéaire est une représentation graphique.

Il existe en réalité des outils statistiques plus fiables que r^2, mais d'une part ils exigent un recul conséquent tant en statistiques que sur l'expérience réalisée, et d'autre part ils demandent un nombre important de mesures pour être efficaces.

III - Prise en compte des incertitudes

Considérons maintenant que les points de mesure sont assortis d'une incertitude, pouvant éventuellement dépendre de n : $(x_n \pm \Delta x_n, y_n \pm \Delta y_n)$. Comment la prendre en compte dans la régression ?

III.1 - Moindres carrés pondérés

Remarque : *Paragraphe hors-programme, donc pas à connaître.*

Il est logique d'accorder plus d'importance à ce que la droite de régression passe près des points dont l'incertitude est la plus faible par rapport à ceux dont l'incertitude est grande. Dans la fonction d'écart C, cela se traduit par l'introduction d'une pondération sur chacun des points dépendant des incertitudes :

$$C(a,b) = \sum_{n=1}^{N} \frac{1}{\Delta y_n^2 + a^2 \Delta x_n^2} \left[y_n - (a x_n + b)\right]^2 .$$

Le coefficient de pondération $1/\left(\Delta y_n^2 + a^2 \Delta x_n^2\right)$ est d'autant plus grand que les incertitudes sont faibles.

L'algorithme est analogue au cas sans incertitude, et le logiciel renvoie les valeurs optimales de a et b qui minimisent la fonction d'écart.

III.2 - Validation du modèle

La validation du modèle est plus simple à comprendre lorsque les incertitudes sont précisées. Elle repose encore sur une représentation graphique des données superposées à la droite de régression.

Pour pouvoir conclure à la validité d'une relation linéaire, il faut
- ▷ que la droite de régression passe par les barres d'erreur de tous les points ;
- ▷ que l'écart entre les points expérimentaux et la droite de régression ne présente pas de courbure.

Exemples : *voir figure 2.*

- ▷ *Courbe de gauche : les barres d'erreurs passent par tous les points, mais une courbure parabolique apparaît nettement ; le modèle n'est donc pas compatible avec les données.*
- ▷ *Courbe centrale : il n'y a pas de courbure, mais cette fois la droite ne passe pas du tout par les barres d'erreur ; le modèle ne décrit pas plus les données.*
- ▷ *Courbe de droite : il n'y a pas de courbure, et la droite passe par toutes les barres d'erreurs ; le modèle est compatible les données.*

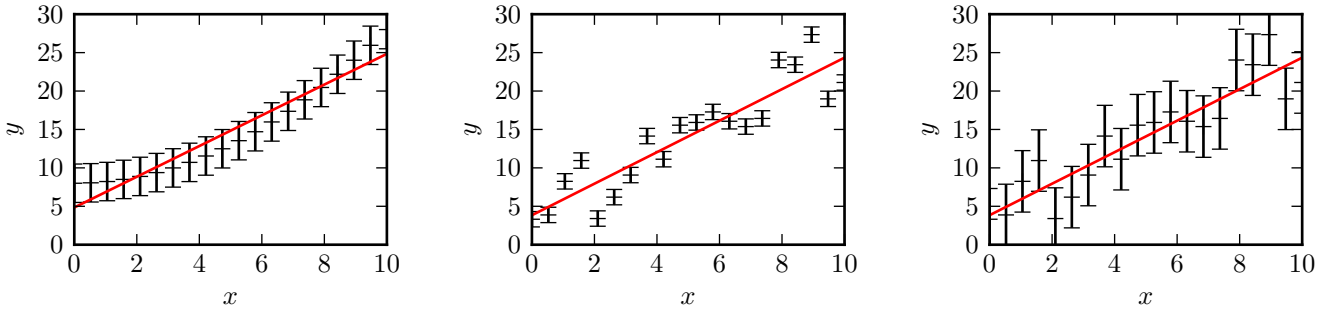


Figure 2 – Régression linéaire avec barres d'erreur. Pour que les figures soient plus lisibles, seules les valeurs de y ont été représentées avec une barre d'erreur.

III.3 - Incertitudes sur les paramètres de la régression

Si une incertitude est attachée aux données sur lesquelles la régression est réalisée, alors une incertitude doit également être associée aux paramètres a et b de la régression. La plupart des logiciels de régression permettent d'avoir accès à ces incertitudes.

Une estimation manuelle (ou graphique) est également possible en jouant sur les paramètres de la régression.
- ▷ Chercher d'abord la valeur *maximale* de la pente et la valeur *minimale* de l'ordonnée à l'origine qui permettent d'obtenir une droite passant par toutes les barres d'erreurs ;
- ▷ Procéder de même pour déterminer la valeur *minimale* de la pente et la valeur *maximale* de l'ordonnée à l'origine ;
- ▷ Les incertitudes sont données par les différences entre ces valeurs extrêmes.

IV - Cas général : régression non-linéaire

Exemple : *Pour une lentille de distance focale image f', on souhaite vérifier la validité de la relation de conjugaison de Descartes ... qui est tout sauf linéaire :*

$$\frac{1}{\overline{OA'}} - \frac{1}{\overline{OA}} = \frac{1}{f'}\,.$$

Pour différents couples objet-image, on mesure les distances $\overline{OA}$ et $\overline{OA'}$.

- **Mauvaise méthode : régression non-linéaire**

Un peu de calcul permet de montrer que

$$\overline{OA'} = \frac{f' \times \overline{OA}}{f' + \overline{OA}}\,.$$

Une première idée pourrait donc être de faire chercher à l'ordinateur la meilleure valeur du paramètre a pour une régression des points expérimentaux par la fonction

$$y = f(x) = \frac{ax}{a + x}\,,$$

Étienne Thibierge, 10 août 2018, www.etienne-thibierge.fr

puis de la comparer à la valeur annoncée de la focale f'. Cette méthode est cependant **à éviter absolument** : d'une part la comparaison visuelle de la courbe de régression aux points expérimentaux n'est pas simple (qui sait reconnaître l'allure d'une telle fonction f ?), d'autre part la fonction d'écart C peut présenter plusieurs minimums, et donc rien ne garantit que les valeurs de a et b renvoyées sont celles du minimum global.

- **Bonne méthode : linéarisation**

> Pour tester la validité d'une loi non-linéaire, il faut la réécrire
> en termes de variables auxiliaires pour lesquelles elle prend une forme linéaire.

Cela sous-entend qu'il faut identifier correctement ces autres variables : cela ne peut se faire qu'au cas par cas. Ici, on peut introduire $y = 1/\overline{OA'}$ et $x = 1/\overline{OA}$. La relation de conjugaison s'écrit alors

$$y - x = \frac{1}{f'} \qquad \text{soit} \qquad y = x + \frac{1}{f'}\,.$$

En faisant calculer directement au logiciel les valeurs de x et y à partir des valeurs expérimentales, on peut sans difficulté procéder à une régression linéaire.