Le B-A-BA de statistiques : les subtilités de la régression linéaire

G. Paturel, Observatoire de Lyon

Résumé : *Dans ce dernier article, nous donnons les grands principes à respecter pour faire correctement une régression linéaire. Nous donnons les relations de la régression linéaire par la méthode des moindres carrés. Nous traitons ensuite un exemple complet.*

Introduction

Nous allons continuer cette brève introduction aux statistiques par la présentation d'une méthode très utilisée : la régression linéaire par les moindres carrés. Le but est de trouver une relation linéaire entre deux grandeurs sous la forme : $Y = AX + B$, où A et B sont les deux constantes que nous voulons déterminer, tandis que Y et X sont les quantités entre lesquelles nous cherchons une corrélation. Les quantités Y et X se présentent comme une collection de mesures (Y_i, X_i). Donnons un exemple concret. Vous voulez savoir si la période de variation lumineuse P d'une étoile Céphéide dépend de sa magnitude absolue M. Vous testez, par exemple, la relation : $M = AP + B$ avec un échantillon de Céphéides, pour lesquelles vous connaissez les magnitudes absolues M_i et les périodes P_i, (i = 1,N).

Les lois ne sont pas toujours linéaires, mais il est souvent facile de se ramener au cas linéaire en prenant le logarithme ou l'exponentielle d'une des grandeurs. Dans l'exemple des Céphéides, la relation théorique est proche de : $P = C.10^{\alpha M}$. La régression linéaire sera recherchée sous la forme $\log P = \log C + \alpha M$, qui peut aussi se mettre sous la forme habituelle :

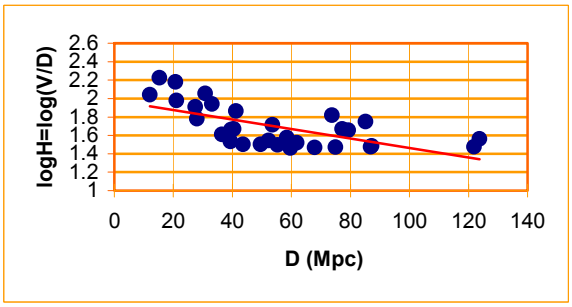
$$M = A \log P + B,$$

avec $B = -\frac{1}{\alpha}\log C$ et $A = \frac{1}{\alpha}$.

Il est donc nécessaire de ne tester que des relations linéaires (ce qui paraît évident compte tenu du nom de la méthode...) en choisissant bien la forme des grandeurs à tester.

Un piège à éviter

Un piège dangereux peut vous faire trouver des corrélations qui n'existent pas. La simulation qui va suivre est inspirée d'une publication réelle. Par exemple, vous voulez savoir si la constante de Hubble H, qui mesure l'expansion de l'Univers, dépend de la distance D. Cette interrogation est bien légitime. L'idée toute simple est de tester avec une régression linéaire la relation : $\log H = AD + B$. (le "log" sert à "linéariser" l'équation). On s'attend à trouver $A=0$, si H est constant quelle que soit la distance.



Faisons une simulation avec un tableur. La constante de Hubble, H, est fixée exactement à 50 (km/s)/Mpc. La vraie distance D' est tirée au hasard entre 10 et 160 Mpc. La distance observée D est égale à $D' + E(D')$, où $E(D')$ est l'incertitude, fonction de D'. Cette incertitude est très grande, compte tenu de la difficulté des mesures de distance. Nous prenons : $E(D') = 1{,}5D'(N-0.5)$, N étant une variable aléatoire entre 0 et 1.

Le graphique $logH = f(D)$ semble montrer que la constante de Hubble décroît avec la distance alors que par construction elle était constante.

L'explication est simple : quand D est surestimé, par suite de son incertitude, $H=V/D$ est sous-estimé et, inversement, si D est sous-estimé, H est surestimé. Finalement, pour les petites valeurs de D

nous aurons plus de chance d'avoir H grand, et réciproquement. Nous aurons une tendance artificielle à trouver une décroissance de H en fonction de D, comme le montre la droite de régression, sensée donner objectivement la tendance.

Que fallait-il faire ? Dans notre exemple, le problème provenait de la corrélation des erreurs entre les deux axes. En effet, la même quantité D intervenait sur les deux axes ($Y=V/D$) et ($X=D$) et, de plus, cette grandeur était affectée d'une incertitude très grande. Il fallait donc tester simplement V en fonction de D et voir si la relation présentait un écart à la linéarité.

Il est donc nécessaire de choisir les variables de telle manière que leurs erreurs soient indépendantes. Nous reviendrons sur la question des erreurs attachées aux variables.

La méthode des moindres carrés

L'idée est de trouver la droite qui passe au milieu des couples de points (Y_i, X_i) (i=1,N), en minimisant les "écarts" entre les points et la droite. Comment résoudre ce problème ?

Pour un X_i donné, le Y correspondant de la droite est $Y=A\ X_i\ +B$, alors que la mesure est Y_i. Ce que nous appellerons "l'écart" sera le carré :

$$\Delta_i=(Y-Y_i)^2=(AX_i+B-Y_i)^2$$

Pourquoi le carré ? Simplement pour éviter qu'il y ait une compensation arithmétique. C'est vraiment la distance géométrique que nous voulons réduire.
Naturellement cette réduction doit se faire pour tous les points (Y_i, X_i). Nous allons donc chercher à minimiser la somme :

$$S=\sum_{i=1}^{N}\Delta_i=\sum_{i=1}^{N}(Y-Y_i)^2=\sum_{i=1}^{N}(AX_i+B-Y_i)^2$$

La méthode générale pour trouver le minimum d'une expression, quand celle-ci varie en fonction d'une quantité, c'est de calculer la dérivée de cette expression en fonction de la quantité. Quand la dérivée est nulle, l'expression est à son extremum (ici le minimum). Nous avons une expression, S, et deux quantités inconnues, A et B. Il nous suffit d'écrire que S doit être minimum quand A et B varient. Nous aurons donc deux relations à satisfaire simultanément :

$$\frac{\partial S}{\partial A}=0 \text{ et } \frac{\partial S}{\partial B}=0 \qquad (1)$$

Or comme nous avons deux inconnues (A et B), notre problème est un système de deux équations à deux inconnues. Il ne reste qu'à résoudre (voir l'encadré à la page suivante).

Un problème insoluble

Dans la méthode que nous venons d'exposer, nous avons minimisé les écarts verticaux, c'est-à-dire selon l'axe Y. Est-ce à dire que les mesures X_i sont sans incertitudes ? Évidemment, dans bien des cas la réponse est non. Donc, la méthode standard des moindres carrés (méthode directe) n'est applicable que dans le cas où seules les grandeurs Y_i sont affectées d'incertitudes. Ce cas arrive, par exemple, quand on étudie la magnitude apparente $Y\equiv m$ d'une étoile en fonction de $X\equiv\sec\zeta=1/\cos\zeta$ pour déterminer l'extinction atmosphérique au cours d'une nuit (C'est la méthode de Bouguer). Dans ce cas l'erreur sur la mesure de la distance zénithale ζ est négligeable devant l'erreur sur la mesure de magnitude. La méthode directe des moindres carrés est tout à fait justifiée. Si nous avons la situation inverse (incertitude sur X seulement), la solution est facile à trouver : il suffit de permuter les axes, pour mettre en Y la quantité qui a une grande incertitude et en X celle dont l'incertitude est négligeable.
Mais qu'en est-il du cas où les deux quantités X et Y sont toutes deux affectées d'incertitudes de mesure ? Nous pouvons faire une régression "directe" (incertitude selon Y) et une régression "inverse" (incertitude selon X). Nous obtenons deux solutions différentes, disons : A et B pour la solution directe et A' et B' pour la régression inverse[1].

La solution correcte se trouve donc entre ces deux solutions extrêmes. Si nous avons de bonnes raisons de penser que les deux axes ont des incertitudes identiques, alors la bonne solution est la moyenne géométrique entre les deux (bissectrice des deux droites de régression). Si nous n'avons aucune information sur les incertitudes, nous pouvons simplement dire que la solution se situe entre les deux solutions extrêmes. Autant dire que le problème est insoluble. Faisons une remarque pour conclure, le coefficient de corrélation, qui mesure la qualité de la régression, est donné par :

$$\rho=\sqrt{\frac{A}{A'}}\ .$$

Test d'hypothèse

Se pose alors le même problème que dans l'article précédent : la corrélation que je trouve entre les quantités Y et X est-elle réelle (on dit significative) ou non ? Il faut effectuer un test d'hypothèse comme nous l'avions vu la dernière fois.

[1] Le calcul de la régression inverse fournit la solution $X=aY+b$. Mais nous pouvons l'exprimer dans le sens direct $Y=A'X+B'$ en posant $A'=1/a$ et $B'=-b/a$.

Droites des moindres carrés

Nous partons des équations (1).

$$\frac{\partial S}{\partial A} = 2\sum_{i=1}^{N}(AX_i + B - Y_i)X_i = 0$$

$$\frac{\partial S}{\partial B} = 2\sum_{i=1}^{N}(AX_i + B - Y_i) = 0$$

Nous calculons d'abord les sommes des X_i, des X_i^2, des Y_i, des Y_i^2 et des X_iY_i. Nous les désignerons respectivement par *SX, X2, SY, Y2* et *XY*.

Les deux équations s'écrivent donc :

$A.X2 + B.X - XY = 0$

$A.X + B.N - Y = 0$ (*N* est le nombre de mesures)

La résolution de ce système nous donne *A* et *B*.

En posant :

$\alpha = XY - SX.SY / N$; $\beta = X2 - SX^2 / N$ et $\gamma = Y2 - SY^2 / N$

On trouve la pente et l'ordonnée à l'origine :

$$A = \frac{\alpha}{\beta} \qquad A' = \frac{\gamma}{\alpha}$$

$$B = \frac{SY}{N} - A\frac{SX}{N} \qquad B' = \frac{SY}{N} - A'\frac{SX}{N}$$

Nous donnons sans démonstration les erreurs moyennes sur *A* et *B* : $\overline{\sigma}_A = \sqrt{\frac{1}{N-2}\left(\frac{\gamma}{\beta} - A^2\right)}$ et $\overline{\sigma}_B = \overline{\sigma}_A\sqrt{\frac{X2}{N}}$

Dans l'encadré figurent les relations qui donnent les erreurs moyennes sur *A* et *B*, que nous noterons respectivement : $\overline{\sigma}_A$ et $\overline{\sigma}_B$.

Pour tester, par exemple, si la pente *A* est significativement différente de 1 et l'ordonnée à l'origine significativement différente de zéro on calcule :

$$t = \frac{|A-1|}{\overline{\sigma}_A} \text{ et } t = \frac{|B|}{\overline{\sigma}_B}$$

Si la valeur *t* est inférieure à $t_{0,01} \approx 2{,}58$[2] je peux conclure que ma valeur est compatible avec la valeur espérée (1 ou 0, respectivement) avec une probabilité d'erreur de 1% (ou, dit autrement, que la valeur trouvée ne diffère pas significativement de la valeur de l'hypothèse).

On peut chercher si le coefficient de corrélation ρ est significativement différent de zéro. L'erreur moyenne sur ρ s'approxime bien par : $\overline{\sigma}_\rho = \frac{1-\rho^2}{\sqrt{N}}$. Il suffit de calculer : $t = \frac{|\rho|}{\overline{\sigma}_\rho}$,et de comparer à $t_{0,01} \approx 2{,}58$ pour pouvoir conclure.

[2] Cette valeur est valable quand N>30. Si N<30, on peut approximer par : $t_{0,01} = 2{,}5 + \frac{7}{N}$

Un exemple pour vérifier

Il est utile d'avoir un exemple traité complètement pour vérifier un programme. Nous donnons l'exemple réel de la relation Période-Luminosité. Le tableau des mesures est le suivant :

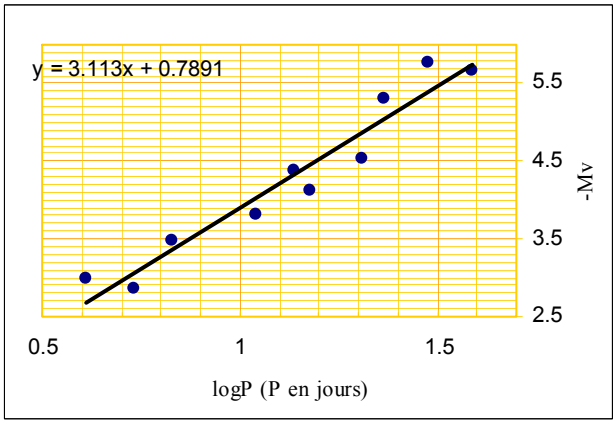
Étoile	logP (P en jours)	$-M_v$ (mag. V)
BF Oph	0.6094	3.00
CV Mon	0.7307	2.85
U Sgr	0.8290	3.49
XX Cen	1.0395	3.81
BN Pup	1.1358	4.38
VW Cen	1.1771	4.13
RY Sco	1.3079	4.54
VZ Pup	1.3650	5.30
AQ Pup	1.4787	5.78
U Car	1.5889	5.67

On trouve :

SX=11.262	*SY*=42.95	
X2=13.6336825	*Y2*=194.3189	*XY*= 51.328942
β = 0.95041806	γ = 9.84865	α = 2.958652
A = 3.113	*B* = 0.7891	
$\overline{\sigma}_A$ =0.289	$\overline{\sigma}_B$ =0.338	
A' = 3.329	*B'* = 0.5461	ρ = 0.967

Dans ce cas la *régression directe* est bien justifiée, l'erreur sur *logP* étant faible par rapport à l'erreur sur la magnitude absolue.

On vérifie que la pente n'est pas significativement différente de 3 (t = 0,39)[3] et que le coefficient de corrélation n'est pas nul ($t \approx 47$). Ceci peut aussi se voir en vérifiant que la pente ne peut pas être considérée comme nulle (t = 10,8).



■

[3] $t_{0,01} \approx 3{,}2$ d'après la note 1. Ici, $t << t_{0,01}$