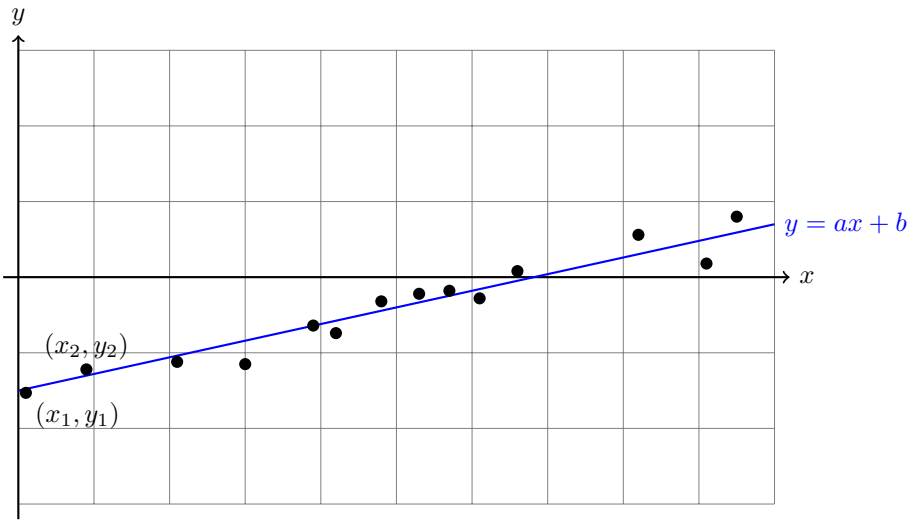
Régressions linéaires



On se donne $n \geqslant 2$, et une série de valeurs expérimentales (x_i, y_i). On souhaite trouver la droite affine $y = ax + b$ «expliquant» le mieux ces résultats – i.e. faire une *régression linéaire*. On souhaiterait également savoir à quel point les données sont «bien expliquées» par cette droite.
On peut formuler le problème de la façon suivante :
«Trouver $(a, b) \in \mathbb{R}^2$ qui minimisent $\sum_{i=1}^{n}(y_i - (ax_i + b))^2$» (P).
On parle de problème de «moindres carrés».

1. Se convaincre que (P) est une façon «raisonnable» de modéliser ce problème.
2. Le résoudre «à la main» dans le cas où l'on n'a que 2 points expérimentaux : $(0.5, 1)$ et $(1, 2)$.

Mathématiquement, on se place dans $E = \mathbb{R}^n$, muni du produit scalaire $(x_1, \dots x_n).(y_1, \dots y_n) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$ (qui est à une constante multiplicative près le produit scalaire usuel).
On se donne $x = (x_1 \dots x_n) \in E$, et $y = (y_1 \dots y_n) \in E$ correspondant au jeu de données.
On notera par ailleurs $u = (1, \dots 1) \in E$, et pour tout $t = (t_1 \dots t_n) \in \mathbb{R}^n$, $m(t) = \frac{1}{n}\sum_{i=1}^{n} t_i$ et $v(t) = \left(\frac{1}{n}\sum_{i=1}^{n} t_i^2\right) - m(t)^2$.
Enfin, on note $F = \text{Vect}(x, u)$.

Dans la suite, on supposera que les x_i ne sont pas tous égaux.

3. Déterminer $\|u\|$ et $u.x$ pour le produit scalaire donné (on exprimera ce dernier résultat en fonction des grandeurs données dans l'énoncé).
4. Justifier que $v(x) > 0$ (on pourra utiliser l'inégalité de Cauchy-Schwarz, ou des probabilités).
5. En interprétant $\sum_{i=1}^{n}(y_i - (ax_i + b))^2$ à l'aide d'une norme, montrer que le problème (P) est équivalent au fait de trouver le projeté orthogonal de y sur F – en déduire que ce problème admet un unique couple (a, b) solution.
6. Justifier que (u, x) est une base de F. L'orthonormaliser pour le produit scalaire donné. Exprimer le résultat obtenu à l'aide de u, x, $m(x)$, $v(x)$, sans symbole $\sum$).
7. On note $(u, \tilde{x})$ la b.o.n. obtenue à la question précédente. Donner une expression du projeté orthogonal de y sur F – $P_F(y)$ – en fonction de y, u et $\tilde{x}$.
8. En déduire les solutions a et b du problème (P). On exprimera les résultats en fonction de $x, y, m(x), m(y), v(x), \tilde{x}$.
9. Vérifier que votre résultat est correct dans les cas particuliers suivants :
   (a) $y = 0$
   (b) $y = (1, \dots 1)$

(c) $y = x$

10. Vérifier que, si l'on multiplie x et y par une même constante $\lambda \in \mathbb{R}^*$, a n'est pas modifié. Que se passe-t-il si l'un des deux vecteurs seulement est multiplié par λ ?
11. Déterminer la valeur minimale d_{min} atteinte pour $\sum_{i=1}^{n}(y_i - (ax_i + b))^2$ en fonction de y et $\tilde{x}$. Indication : on utilisera le théorème de Pythagore. Quel sens donner à cette quantité ?
12. À partir de la question précédente, en posant $\tilde{y} = \dfrac{y - m(y)u}{\sqrt{v(y)}}$ (dans le cas où $v(y) \neq 0$), montrer que la quantité $(\tilde{x}.\tilde{y})^2$ est un «bon indicateur» de la «corrélation linéaire» entre x et y.

Quelques questions plus ouvertes :

13. On change le carré en une valeur absolue (uniquement dans cette question). On doit donc mimimiser $\sum_{i=1}^{n}|y_i - (ax_i + b)|$.
    (a) Résoudre ce problème dans le cas particulier où l'on a les 3 points expérimentaux $(0,0), (1,1), (2,0)$.
    (b) Résoudre ce problème dans le cas particulier où l'on a les 4 points expérimentaux $(0,0), (1,1), (0,1)$ et $(1,0)$.
14. Que se passe-t-il si les (x_i) sont tous égaux ?
15. Les résultats obtenus ne sont pas symétriques en x et en y. Discutez.
16. Généraliser la méthode obtenue :
    (a) si l'on veut obtenir une relation du type $ax^2 + b$;
    (b) si l'on veut obtenir une relation du type $ax^2 + bx + c$;
    (c) si l'on a des jeux de données (x_i, y_i, z_i), et que l'on veut trouver les «meilleurs» coefficients (a, b, c) tels que $z = ax + by$.