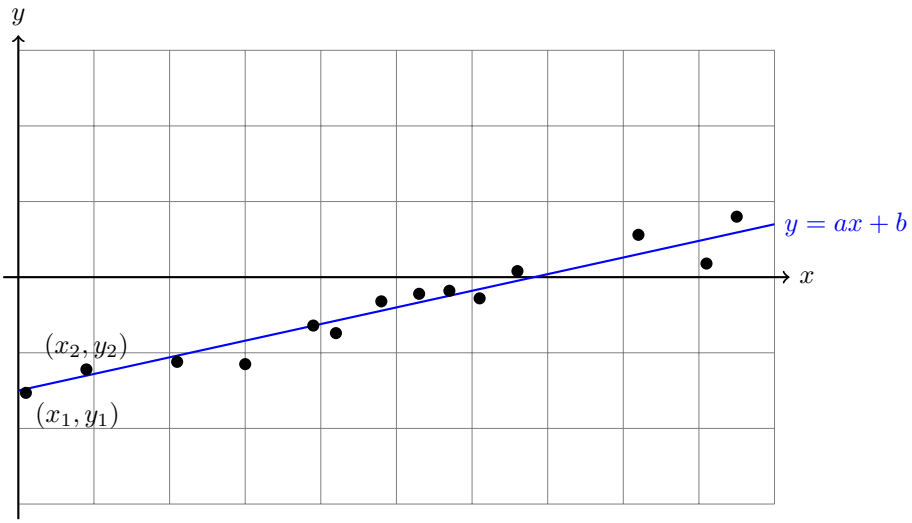
Régressions linéaires



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.

> $(y_i - (ax_i + b))^2$ est une mesure la distance verticale entre le point (x_i, y_i) et la droite. Minimiser la somme de ces distances pour tous les points semble donc être une bonne façon d'obtenir une droite approchant bien les données.

2. Le résoudre «à la main» dans le cas où l'on n'a que 2 points expérimentaux : $(0.5, 1)$ et $(1, 2)$.

> On cherche la droite affine passant par ces 2 points. On obtient $a = 2$ et $b = 0$.
> Pour ces valeurs de a et de b, la quantité à minimiser est alors $(1 - 2 \times 0.5)^2 + (2 - 2 \times 1)^2 = 0$.
> La quantité à miniser étant par ailleurs toujours positive, on a bien trouvé un minimum.

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).

> $\|u\| = 1$. $u.x = \frac{1}{n}\sum_{i=1}^{n} x_i = m(x)$.

4. Justifier que $v(x) > 0$ (on pourra utiliser l'inégalité de Cauchy-Schwarz, ou des probabilités).

D'après l'inégalité de Cauchy-Schwarz, $|u.x| \leqslant \|u\| \, \|x\|$, i.e. $|m(x)| \leqslant \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}$.
En passant au carré (les 2 membres étant positifs), $(m(x))^2 \leqslant \frac{1}{n}\sum_{i=1}^{n} x_i^2$, d'où $v(x) \geqslant 0$.
(On aurait également pu voir $v(x)$ comme la variance d'une certaine variable aléatoire – à définir – et en déduire donc qu'elle est positive.)

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.

$\sum_{i=1}^{n}(y_i - (ax_i + b))^2 = \sum_{i=1}^{n}(y_i - (ax_i + b1))^2 = \|y - (ax + bu)\|^2$.
$ax + bu$ parcours F lorsque a et b parcourrent $\mathbb{R}$.
Minimiser la somme est donc équivalent à trouver le projeté orthogonal de y sur F, qui minimise la distance entre y et tout point de F. D'après le cours, il y a un unique point solution dans F, et comme (x, u) est libre, 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 $\sum$).

(u, x) est génératrice de F par définition, et libre car les x_i ne sont pas tous égaux. C'est donc une base de F.
$\|u\| = \frac{1}{n}\sum_{i=1}^{n} 1^2 = 1$.
$\tilde{x} = x - (u.x)u = x - m(x)u$.

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 $\tilde{x}$.

$P_F(y) = (y.u)u + (y.\tilde{x})\tilde{x}$.

8. En déduire les solutions a et b du problème (P). On exprimera les résultats en fonction de $x, y, m(x), m(y), v(x), \tilde{x}$.

On a donc $b = y.u - \frac{(y.\tilde{x})m(x)}{\sqrt{v(x)}}$.
$a = \frac{(y.\tilde{x})}{v(x)} = \frac{y.x - m(x)m(y)}{v(x)}$.

9. Vérifier que votre résultat est correct dans les cas particuliers suivants :

   (a) $y = 0$

   Si $y = 0$, $a = b = 0$.
   Tous les points expérimentaux sont sur la droite d'équation $y = 0$, on on obtient bien la droite de régression $y = ax + b = 0$.

   (b) $y = (1, \ldots 1)$

   Si $y = u$, $a = \frac{u.x - m(x)m(u)}{v(x)} = \frac{m(x) - 1m(x)}{v(x)} = 0$, et $b = u.u - \frac{(u.(x - m(x)u))m(x)}{\sqrt{v(x)}} = 1 - \frac{(m(x) - m(x))m(x)}{\sqrt{v(x)}} = 1$
   Tous les points expérimentaux sont sur la droite d'équation $y = 1$, on on obtient bien la droite de régression $y = ax + b = 1$.

(c) $y = x$

> Si $y = x$, $a = \dfrac{\|x\|^2 - m(x)^2}{v(x)} = 1$, et $b = m(x) - \dfrac{(x.\tilde{x}m(x))}{v(x)} = m(x) - \dfrac{\|x\|^2 - m(x)^2}{v(x)} m(x) = 0$.
> Tous les points expérimentaux sont sur la droite d'équation $y = x$, on on obtient bien la droite de régression $y = ax + b = x$.

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 ?

> - $\dfrac{(\lambda y).(\lambda x) - m(\lambda x)m\lambda(y)}{v(\lambda x)} = \dfrac{\lambda^2}{\lambda^2} \dfrac{y.x - m(x)m(y)}{v(x)} = \dfrac{y.x - m(x)m(y)}{v(x)}$.
>   Donc a n'est pas modifié par la multiplication simultanée de x et y par λ (le nuage de points de départ est la même à un facteur d'échelle près, il est normal que la pente de la droite de régression ne change pas).
> - Si x est seul multiplié par λ, a est multiplié par λ (le nuage de point est modifié par un facteur d'échelle λ dans la direction des abscisses)
> - Si x est seul multiplié par λ, a est multiplié par $1/\lambda$ (au facteur d'échelle près, le nuage de point est modifié par un facteur d'échelle $1/\lambda$ dans la direction des ordonnées)

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 ?

> La distance au carré entre y et son projeté est $\|y\|^2 - \|P_F(y)\|^2 = \|y\|^2 - (y.u)^2 - (y.\tilde{x})^2 = v(y) - (y.\tilde{x})^2$.
> On a donc $d_{min} = \sqrt{v(y) - (y.\tilde{x})^2}$.
> Cette quantité est une indication de «à quel point la droite de régression est proche des données expérimentales».

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.

> $\tilde{x}.\tilde{y} = \tilde{x}.\dfrac{y - m(y)u}{\sqrt{v(y)}} = \dfrac{1}{\sqrt{v(y)}} (\tilde{x}.y)$ (car $\tilde{x}.u = 0$).
> Donc
> $(\tilde{x}.\tilde{y})^2 = \dfrac{(\tilde{x}.y)^2}{v(y)} = \dfrac{v(y) - d_{min}^2}{v(y)} = 1 - \dfrac{d_{min}^2}{v(y)}$.
> - Lorsque $d_{min} = 0$, cette quantité vaut 1 (les points expérimentaux sont sur une même droite).
> - D'après la question précédente, on sait que $d_{min}^2 = v(y) - (y.\tilde{x})^2 \leqslant v(y)$. La valeur maximale possible de d_{min}^2 est $v(y)$ (on peut montrer qu'elle peut être atteinte – en trouvant $y \perp \tilde{x}$ tel que $v(y) \neq 0$). Dans ce cas, $(\tilde{x}.\tilde{y})^2 = 0$.
>
> $(\tilde{x}.\tilde{y})^2$ est le *coefficient de corrélation* de la régression linéaire : il est proche de 1 lorsque les points expérimentaux sont presque alignés, et proche de 0 lorsqu'il n'y a pas de «bonne» droite affine permettant d'en rendre compte.

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 $z = ax + by$.