Ajustements d'une série statistique

par Daniel SAADA, Lycée de Rambouillet

Ajuster des données (x_i, y_i), c'est trouver une fonction (une loi) F telle que chaque $F(x_i)$ soit "proche" de y_i.

On peut alors effectuer des interpolations (évaluer une population entre deux recensements par exemple) ou des extrapolations (pour dégager une prévision). De plus, une formule compacte représentant correctement un grand nombre d'observations libère une place précieuse dans la mémoire d'un ordinateur.

Cet article examine les problèmes numériques parfois ardus qu'entraîne le choix rationnel des paramètres du modèle F.

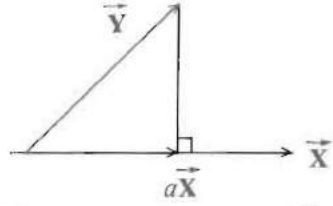
A. L'ajustement linéaire

C'est le modèle le plus simple. On choisit de rendre minimale la somme S des carrés des erreurs : $S = \sum_{i=1}^{n} (y_i - ax_i - b)^2$, a et b étant les paramètres inconnus.

Une fois établi que le point moyen $(\bar{x}, \bar{y})$ est sur la droite cherchée, on écrit : $S = \sum_{1}^{n} [(y_i - \bar{y} - a(x_i - \bar{x})]^2$.

Introduisons alors les deux vecteurs de $\mathbf{R}^n$:

$$\vec{Y} = (y_i - \bar{y}) \qquad \text{et} \qquad \vec{X} = (x_i - \bar{x}) \qquad i = 1,2,\ldots,n.$$



S valant $\|\vec{Y} - a\vec{X}\|^2$ sera minimale quand $\vec{Y} - a\vec{X}$ devient orthogonal à $\vec{X}$. D'où

$$a = \frac{\vec{X}.\vec{Y}}{\|\vec{X}\|^2} = \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{\Sigma (x_i - \bar{x})^2}.$$

Après développement, on retrouve bien :

$$a = \frac{n \Sigma x_i y_i - \Sigma x_i \Sigma y_i}{n \Sigma x_i^2 - (\Sigma x_i)^2}$$

que la machine calcule rapidement au moyen des registres statistiques.

Le coefficient de corrélation ϱ, qui mesure traditionnellement la qualité de l'ajustement, n'est autre que le cosinus de l'angle $(\vec{X},\vec{Y})$:

$$\varrho = \frac{\vec{X}.\vec{Y}}{\|\vec{X}\|.\|\vec{Y}\|} = a.\frac{\|\vec{X}\|}{\|\vec{Y}\|}.$$

Enfin, $S = \|\vec{Y}\|^2 - \|a\vec{X}\|^2 = \|\vec{Y}\|^2(1-\Gamma^2)$.

Exemple :

Les trois points

x	0	1	25
y	1	10	100

conduisent à la droite $y = 3{,}86x + 3{,}52$, équation qui donne pour les y respectivement :

3,5 7,4 100,1

L'ajustement est médiocre malgré un "ϱ" très élevé : 0,999.

L'erreur relative énorme commise pour $x=0$ montre que ce modèle n'est pas sans défaut.

On peut alors lui préférer celui minimisant la somme S' des carrés des erreurs relatives : $S' = \Sigma\left(\frac{y_i - ax_i - b}{y_i}\right)^2$ (en supposant aucun y_i nul).

On trouve : $a' = \frac{1}{D} \times [(\Sigma u_i)\,\Sigma v_i^2 - (\Sigma u_i v_i)\,\Sigma v_i]$

$$b' = \frac{1}{D} \times [\Sigma u_i^2 . \Sigma v_i - \Sigma u_i v_i \,\Sigma u_i]$$

où $D = \Sigma u_i^2 . \Sigma v_i^2 - (\Sigma u_i v_i)^2$; $u_i = x_i/y_i$; $v_i = 1/y_i$.

Ici : $a' = 4{,}65$ et $b' = 1{,}04$ qui fournissent

1,04 5,7 117,2

Les résultats, très différents des précédents, montrent la fragilité des évaluations basées sur un seul modèle d'ajustement.

B. Les ajustements exponentiels

A la verticale d'un même lieu, on a mesuré plusieurs fois la pression à diverses altitudes :

Altitudes en km	0	1	2	5	10	15	20	30
Pressions moyennes en mm	760	674	596	403	198	90	41	9

Pour représenter la pression P, nous choisissons le modèle ae^{-bh}.

La démarche classique est d'écrire : $Log\,P = Log\,a - b\,h$ et d'ajuster linéairement les couples $(h, Log\,P)$.

Le calcul donne :

$$a = 809{,}020\,686\,3$$

et $$b = 0{,}148\,578\,274\,4$$

avec un ϱ exceptionnel : $-0{,}999\,54$!

Pourtant, on obtient pour les pressions :

809,0 697,3 601,0 384,9 183,1 87,1 41,4 9,4

ce qui est décevant. L'indicateur $S = \sum_{1}^{8} (P_i - ae^{-bh_i})^2$ vaut 3527,24. Nous allons réduire au maximum S en résolvant le problème :

$$\inf_{(x,y)\in\mathbf{R}^2} \Sigma (P_i - xe^{yh_i})^2$$

L'annulation des dérivées partielles en x et en y conduit au système :

$$\begin{cases} x \Sigma e^{2yh_i} = \Sigma P_i e^{yh_i} \\ x \Sigma h_i e^{2yh_i} = \Sigma h_i P_i e^{yh_i}, \end{cases}$$

puis à l'équation en y :

$$\Sigma e^{2yh_i} \cdot \Sigma P_i h_i e^{yh_i} = \Sigma h_i e^{2yh_i} \cdot \Sigma P_i e^{yh_i}$$

dont la résolution est manifestement inconcevable sans outil de calcul.

On trouve cette fois-ci : $y = -0{,}135\,220\,286\,7$

puis : $x = 770{,}3279\,59$.

Les valeurs ajustées deviennent :

770,3 672,9 587,8 391,8 199,3 101,3 51,6 13,3

S vaut maintenant 558,1 : le gain est donc considérable.

Toutefois, l'erreur relative sur $P=9$ étant importante, minimisons $\Sigma \left(\dfrac{P_i - xe^{yh_i}}{P_i} \right)^2$. L'équation en y, d'ailleurs plus régulière, devient :

$$\Sigma \alpha_i^2 \cdot \Sigma h_i \alpha_i = \Sigma \alpha_i \cdot \Sigma h_i \alpha_i^2 \qquad \text{avec} \qquad \alpha_i = e^{yh_i}/P_i.$$

En résolvant : $y = 0{,}148\,531\,8$ et $x = 806{,}202\,55$.

Curieusement, on obtient des valeurs voisines de celles de l'ajustement linéaire, mais on peut montrer que ce n'est pas fortuit :

806,2 694,9 599,0 383,6 182,5 86,9 41,3 9,4

Des calculs analogues sont possibles avec les modèles :

$y = ax^b$ et $y = a + be^{cx}$.

Le lecteur formera les équations correspondantes.

Il n'en est pas de même pour la loi dite logistique : $y = \dfrac{a}{1 + be^{-cx}}$ qui,

entre autres utilisations, est connue pour représenter fidèlement la population des Etats-Unis (en millions d'habitants) en fonction du temps :

Années t_i	Populations P_i	Années t_i	Populations P_i
1790	3,9	1880	50,2
1800	5,3	1890	62,9
1810	7,2	1900	76,0
1820	9,6	1910	92,0
1830	12,9	1920	105,7
1840	17,1	1930	122,8
1850	23,2	1940	132,0
1860	31,4	1950	150,0
1870	38,6		

En effet, la minimisation de $\sum_i \left(P_i - \frac{a}{1+be^{-ct_i}}\right)^2$ aboutit à un système en a, b, c, dont il paraît impossible d'isoler une inconnue :

$$\begin{cases} \Sigma\, P_i/D_i = a\,\Sigma\, 1/D_i^2 \\ \Sigma\, P_i e^{-ct_i}/D_i^2 = a\,\Sigma\, e^{-ct_i}/D_i^3 \\ \Sigma\, P_i t_i e^{-ct_i}/D_i^2 = a\,\Sigma\, t_i e^{-ct_i}/D_i^3 \end{cases} \qquad \text{avec} \quad D_i = 1 + be^{-ct_i}$$

Pour résoudre le système $f(b,c) = g(b,c) - 0$ obtenu en éliminant a, j'ai adopté la méthode suivante :

0. Choisir un c initial,
1. Calculer b, solution de $f(b,c) = 0$,
2. Evaluer alors $g(b,c)$,
3. Rechoisir c en vue d'annuler g.

En d'autres termes, il s'agit de résoudre l'équation $g(c;h(c))$ où $b = h(c)$.

La programmation de cet algorithme, malaisée et d'exécution lente (on regrette que "SOLVE" des HP 34C et 15C ne soit pas récurrent), aboutit à $c = 0{,}3092$ puis à $b = 570{,}95$ et $a = 198{,}53$, les années t_i ayant été numérotées pour des raisons techniques de 8 à 24. On a donc

$$P(t) = \frac{198{,}53}{1 + 570{,}95\, e^{-0{,}030\,92(t-1710)}}$$

On obtient pour la population des U.S.A. :

Années t_i	Populations P_i	Années t_i	Populations P_i
1790	4,0	1880	49,9
1800	5,5	1890	62,3
1810	7,4	1900	76,2
1820	9,9	1910	91,2
1830	13,3	1920	106,5
1840	17,6	1930	121,5
1850	23,3	1940	135,4
1860	30,4	1950	148,0
1870	39,3		

Reste à expliquer le choix initial de c. Le tâtonnement peut échouer ou rendre le temps de calcul prohibitif. On peut prendre les valeurs prônées par divers auteurs... ou faire passer la courbe logistique par trois points équidistants et résoudre le système en a, b, c. La meilleure méthode, ici, est d'ajuster linéairement $\left(t_i, Log\left(\frac{a}{P_i} - 1\right)\right)$, a étant choisi pour rendre ρ maximal. On trouve : $a = 198$ $b = 602{,}3$ et $c = 0{,}3122$.

C. **Autres critères d'ajustement**

Nous en avons étudié deux :

$$(1) \quad \inf \sum_i (y_i - F(x_i))^2$$

et $$(2) \quad \inf \sum_i \left(\frac{y_i - F(x_i)}{y_i} \right)^2$$

En voici quatre autres, aussi légitimes :

$$(3) \quad \inf \sum_i | y_i - F(x_i) |$$

$$(4) \quad \inf \sum_i \left| \frac{y_i - F(x_i)}{y_i} \right|$$

$$(5) \quad \inf \sup_i | y_i - F(x_i) |$$

$$(6) \quad \inf \sup_i \left| \frac{y_i - F(x_i)}{y_i} \right|$$

La présence des valeurs absolues empêche leur examen par dérivation. Nous emploierons donc une méthode purement numérique.

Supposons qu'il s'agisse de minimiser $S(a,b) = \sum_i |y_i - F(a,b,x_i)|$.

On fixe a et on cherche b réalisant $\inf_b S(a,b)$: S est devenue fonction d'une variable, b est déterminé à 10^{-p} près. Puis on calcule a satisfaisant à $\inf_a S(a,b)$, et ainsi de suite. Cet algorithme, simple mais lent, diminue S à chaque étape.

J'ai mené ce type de calcul pour la série Pression/Altitude mentionnée au début du B. Les résultats sont consignés dans le tableau ci-après ; pour une bonne lecture les critères 1, 2, 3 et 4 ont été "normalisés". J'ai été aidé dans ma tâche par les possibilités vectorielles de la HP 15 C. En effet, posons $d_i = P_i - x e^{yh_i}$ et stockons au préalable les huit P_i dans la matrice-colonne A. La matrice-colonne B, variable, contiendra les e^{yh_i} puis on calculera $C = A - xB$. L'instruction MATRIX 8 fournit $\sqrt{\Sigma d_i^2}$, tandis que MATRIX 7 donne $sup\ |d_i|$; MATRIX 4 transpose la matrice C et MATRIX 7 affichera alors $\Sigma |d_i|$! Le calcul est fulgurant pour y fixe ; lorsque y varie, il est nécessaire de recomposer la matrice B.

Critères n° / Solutions n°	① $\sqrt{\frac{1}{n}\Sigma d_i^2}$	② $\sqrt{\frac{1}{n}\Sigma \frac{d_i^2}{P_i^2}}$	③ $\frac{1}{n}\Sigma \lvert d_i \rvert$	④ $\frac{1}{n}\Sigma \lvert d_i/P_i \rvert$	⑤ $sup\ \lvert d_i \rvert$	⑥ $sup\ \lvert d_i/P_i \rvert$
a = 770,328 ① b = −0,135 22	8,4	19,8 %	7,3	11,6 %	11,3	48,1 %
806,20255 ② −0,148 532	20,0	4,5 %	13,6	3,8 %	46,2	7,8 %
772,3 ③ −0,1361	8,4	18,4 %	7,0	10,7 %	12,3	44,7 %
800,0 ④ −0,148 55	18,2	4,5 %	12,6	3,7 %	40,0	8,5 %
770,13 ⑤ −0,135 22	8,4	19,8 %	7,3	11,6 %	11,3	48,1 %
810,8 ⑥ −0,1479	21,5	4,7 %	14,5	4,2 %	50,8	6,7 %

Voici maintenant les valeurs ajustées engendrées par les six variations du modèle (à gauche les valeurs réelles) :

	1	2	3	4	5	6
760	770,3	806,2	772,3	800,0	770,1	810,8
674	672,9	694,9	674,0	689,6	672,7	699,3
596	587,8	599,0	588,3	594,4	587,6	603,2
403	391,8	383,6	391,1	380,6	391,7	387,0
198	199,3	182,5	198,0	181,1	199,2	184,8
90	101,3	86,9	100,3	86,2	101,3	88,2
41	51,5	41,3	50,8	41,0	51,5	42,1
9	13,3	9,4	13,0	9,3	13,3	9,6

Enfin, comparons les estimations pour $h = 8$, 25, 35, et 40 km :

h	Valeurs extrêmes de P	
8	243,8	261,1
25	19,5	26,2
35	4,4	6,8
40	2,1	3,4

Une grande prudence est donc de mise.