Régression, aspects déterministes

Fredon Daniel
IREM de Limoges

L'atelier a pour but d'étudier quelques aspects de la régression simple dans sa première approche déterministe.

I. Choix des variables

La recherche d'une droite de régression $y = a + bx$ suppose a priori que les caractères statistiques X et Y ont des statuts différents. Y est la variable à expliquer et X la variable potentiellement explicative (ou susceptible d'expliquer Y). Il peut aussi arriver que X soit plus facile à mesurer et que la recherche d'un modèle ait pour but d'obtenir Y par le calcul.
Le calcul simultané des droites de régression de Y par rapport à X et de X par rapport à Y n'a donc guère de sens ni sur le plan des applications concrètes, ni sur le plan des aspects plus développés de la régression.

II. Ajustement affine par la méthode des moindres carrés : décomposition de la variance

Après obtention de la droite de régression de Y par rapport à X, cherchons à décomposer $V(Y)$.

$$V(Y) = \frac{1}{n}\sum_{i=1}^{k} n_i\,(y_i - \overline{y})^2$$

$$= \frac{1}{n}\sum_{i=1}^{k} n_i\,[(y_i - a - bx_i) + (bx_i - b\overline{x})]^2 \qquad \text{car} \qquad \overline{y} = a + b\overline{x}$$

On a : $\dfrac{1}{n}\sum_{i=1}^{k} n_i\,(bx_i - b\overline{x})^2 = \dfrac{b^2}{n}\sum_{i=1}^{k} n_i\,(x_i - \overline{x})^2 = b^2 V(X) = V(a + bX)$

et $\dfrac{2}{n}\sum_{i=1}^{k} n_i(y_i - a - bx_i)\,(bx_i - b\overline{x}) = \dfrac{2b}{n}\sum_{i=1}^{k} n_i\,[y_i - \overline{y} + b(\overline{x} - x_i)]\,[x_i - \overline{x}]$

$$= 2b\,[Cov(X,Y) - bV(X)] = 0.$$

Pour les valeurs de a et de b correspondant à la droite de régression, on a donc :

$$V(Y) = V(a + bX) + \frac{1}{n}\sum_{i=1}^{k} n_i(y_i - a - bx_i)^2$$

égalité que l'on interprète par :
variance de Y = variance expliquée (par l'ajustement affine) + variance résiduelle

On constate donc que :

$$\frac{\text{variance expliquée}}{\text{variance totale}} = \frac{V(a+bX)}{V(Y)} = b^2\frac{V(X)}{V(Y)}$$
$$= \frac{[Cov(X,Y)]^2}{V(X)\,V(Y)} \quad \text{car} \quad b = \frac{Cov(X,Y)}{V(X)}$$
$$= r^2$$

r^2 apparaît donc comme mesure de la qualité de l'ajustement affine.

III. Ajustement linéaire

Le modèle affine, même après avoir appliqué des transformations aux variables, n'est pas le seul modèle possible. Dans le but de donner un autre exemple, considérons le modèle linéaire $Y = aX$ et conservons le choix classique de mesure de la distance entre les points expérimentaux et une courbe de la famille par la somme des carrés des écarts verticaux. Cette variante peut faire l'objet d'un problème en STS. En voici un énoncé possible.

Soit $(x_1; y_1), \ldots, (x_n; y_n)$ une série statistique à deux dimensions. On se propose d'ajuster sur ces données une relation linéaire $y = ax$.

1. Déterminez a tel que $S(a) = \sum_{i=1}^{n}(y_i - ax_i)^2$ soit minimum.

2. a étant égal à la valeur obtenue dans la première question, vérifiez qu'après ajustement on a :

$$\left(\sum_{i=1}^{n} x_i^2\right)\left(\sum_{i=1}^{n} y_i^2\right) = \left(\sum_{i=1}^{n} x_i y_i\right)^2 + \left(\sum_{i=1}^{n}(y_i - ax_i)^2\right)\left(\sum_{i=1}^{n} x_i^2\right)$$

3. Pour mesurer la qualité de l'ajustement linéaire, on calcule :

$$d = \frac{\left(\sum_{i=1}^{n} x_i y_i\right)^2}{\left(\sum_{i=1}^{n} x_i^2\right)\left(\sum_{i=1}^{n} y_i^2\right)}$$

Vérifiez que $0 \leq d \leq 1$ et que $(d = 1) \Longleftrightarrow$ (tous les points sont alignés).

4. Les rayons γ pur émis par une substance radioactive sont en partie absorbés par les écrans de plomb. Si N désigne le nombre d'atomes radioactifs exprimé dans une unité qui correspond à la mesure au compteur Geiger, on a la loi théorique : $N = N_0 e^{-ax}$ où N_0 correspond à l'absence d'écran, x désigne l'épaisseur des écrans et a une constante.

On a obtenu les mesures ci-dessous où n est le nombre d'écrans de 2 mm d'épaisseur.

n	0	1	2	3	4	5	6	7
N	8 623	7 333	6 273	5 347	4 679	4 031	3 384	2 956

Déterminez la valeur de la constante a en m^{-1} en utilisant un ajustement linéaire. Calculez la valeur du coefficient d.

Pour une solution, voir page 32
F.Couty, J.Debord, D.Fredon
Probabilités et statistiques pour les biologistes
collection Flash U A.Colin

IV. Régression orthogonale

Si on choisit le modèle affine alors que Y et X jouent des rôles symétriques, au lieu de retenir comme mesure de la distance entre une droite et les points expérimentaux la somme des carrés des écarts verticaux, il est plus logique de considérer la somme des carrés des distances des points à la droite. C'est la régression orthogonale.
Comme la droite obtenue est le premier axe factoriel de l'analyse en composantes principales des données, ce thème peut aussi servir d'introduction à l'analyse des données.
Voici un énoncé possible sous forme de problème; seules les deux premières questions sont faciles pour un élève de STS.

Soit (X,Y) une série double représentée dans le plan rapporté à un repère $(O;\vec{i};\vec{j})$ par les points $M_k(x_k,y_k)$, $k \in \{1,2,\ldots,n\}$. On considère le point moyen $G(\bar{x},\bar{y})$ et le changement de variables : $x' = x - \bar{x}$, $y' = y - \bar{y}$.
Etant donné une droite Δ dont une équation dans le repère $(G;\vec{i};\vec{j})$ est :

$$x' \cos\alpha + y' \sin\alpha - \rho = 0,$$

on note H_k la projection orthogonale de M_k sur Δ.
On se propose de déterminer Δ de manière à minimiser $S = \sum_{k=1}^{n} H_k M_k{}^2$.

1. Montrez que $S = \sum_{k=1}^{n} (x'_k \cos\alpha + y'_k \sin\alpha - \rho)^2$.

2. Montrez que, s'il existe une droite Δ qui minimise S, cette droite passe par G.

3. Déterminez le coefficient directeur de Δ qui rend S minimum. Précisez dans quel cas la solution est unique. La droite Δ ajustée est appelée *droite de régression orthogonale.*

4. On suppose que la droite Δ de régression orthogonale est unique. Soit D et D' les droites de régression de y par rapport à x et de x par rapport à y. Etudiez la position relative des droites D, D' et Δ.

5. Soit $T = \sum_{k=1}^{n} GM_k^2$ et $S' = \sum_{k=1}^{n} GH_k^2$. Montrez que le critère de l'ajustement revient à maximiser S'.

On pose $q = \frac{S'}{T}$. Montrez que q est un indice de qualité de l'ajustement compris entre $\frac{1}{2}$ et 1.

Quelle est la signification des valeurs $q = 1$ et $q = \frac{1}{2}$?

6. On donne les notes d'un groupe de 8 étudiants dans deux disciplines : mathématiques (note x sur 20) et informatique (note y sur 20).

x	12	15	8	6	10	12	11	9
y	10	16	11	4	9	14	17	8

Ajustez sur ces données la droite Δ de régression orthogonale et calculez le coefficient q de qualité de l'ajustement. Interprétez l'ajustement de Δ.

Pour une solution, voir page 73

C.Raffin
Statistiques et probabilités DEUG A 2ème année
collection Flash U A.Colin

Daniel FREDON
IREM de LIMOGES