Coefficient de corrélation et régression linéaires

Jean-Guillaume.Dumas@imag.fr

22 mars 2012

1 Droite de régression, méthode des moindres carrés

Plusieurs droites peuvent s'ajuster à un nuage de points mais parmi toutes ces droites on peut retenir celle qui jouit d'une propriété remarquable : celle qui minimise la somme des carrés des écarts des ordonnées observées. Soit un échantillon de n couples (x_i, y_i). Nous voulons donc calculer la droite $\boxed{Y = aX + b}$ minimisant la fonction de a et b suivante :

$$F(a,b) = \sum_{i=1}^{n} (y_i - (ax_i + b))^2.$$

Une condition nécessaire pour atteindre un extrémum d'une fonction différentiable est d'annuler ses dérivées partielles. $F(a,b)$ est différentiable comme somme et produits de fonctions classiquement différentiables. Or

$$\frac{\delta F}{\delta a} = \sum_{i=1}^{n} 2x_i(ax_i + b - y_i) \quad \text{et} \quad \frac{\delta F}{\delta b} = \sum_{i=1}^{n} 2(ax_i + b - y_i).$$

D'où l'on tire les conditions nécessaires suivantes :

$$\sum_{i=1}^{n} y_i = a\sum_{i=1}^{n} x_i + nb \quad \text{et} \quad \sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i.$$

Ainsi, si l'on note

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad \text{et} \quad \overline{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

les moyennes respectives en x et y, alors on a une valeur à l'origine de la droite : $\boxed{b_0 = \overline{y} - a_0\overline{x}}$ et une pente :

$$\boxed{a_0 = \frac{\left(\sum_{i=1}^{n} x_i y_i\right) - \overline{x} n \overline{y}}{\left(\sum_{i=1}^{n} x_i^2\right) - n\overline{x}^2}.}$$

Seules ces valeurs de a et b peuvent donner un minimum de la fonction F. En fait, l'étude des dérivées partielles secondes avec un développement de Taylor montre que les valeurs trouvées correspondent bien à un extrémum et que celui-ci est un minimum : On a

$$r = \frac{\delta^2 F}{\delta a^2} = \sum_{i=1}^{n} 2x_i^2, \quad s = \frac{\delta^2 F}{\delta a \delta b} = \sum_{i=1}^{n} 2x_i \quad \text{et} \quad t = \frac{\delta^2 F}{\delta b^2} = 2n.$$

Ainsi, on considère le polynôme en h (ou en k) définit par le développement de Taylor suivant : $F(a_0 + h, b_0 + k) - F(a_0, b_0) \approx rh^2 + shk + tk^2$, de discriminant $k^2(s^2 - rt)$. Or, en (a_0, b_0), on a $s^2 - rt < 0$ dès que $n > 1$ et donc les valeurs correspondent à un extrémum. En outre $r(a_0, b_0)$ est positif prouvant que (a_0, b_0) est un minimum de F.

2 Coefficient de corrélation linéaire

2.1 Définition

En notant

$$S(z) = \sqrt{(\sum_{i=1}^{n} z_i^2) - n\overline{z}^2}$$

pour un échantillon de n valeurs z_i, le **coefficient de corrélation linéaire** d'un échantillon est défini par la relation suivante :

$$\boxed{r = \frac{\sum_{i=1}^{n} x_i y_i - \overline{x} n \overline{y}}{S(x)S(y)}.}$$

On voit alors que $\boxed{a_0 = r\frac{S(y)}{S(x)}}$.

2.2 Interprétation

Tout d'abord,

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}.$$

Si l'on note $\overrightarrow{z} = [z_1 - \overline{z}, \ldots, z_n - \overline{z}]$, on peut alors écrire r sous la forme suivante :

$$r = \frac{(\overrightarrow{x}, \overrightarrow{y})}{||\overrightarrow{x}||_2 \; ||\overrightarrow{y}||_2} = cos(\overrightarrow{x}, \overrightarrow{y}).$$

On peut alors remarquer que r est compris entre -1 et 1. D'autre part, dans les cas où $r = \pm 1$, les deux vecteurs ayant un cosinus nul doivent être colinéaires. Donc $\exists a, \overrightarrow{y} = a\overrightarrow{x}$. C'est à dire $\forall i, y_i = ax_i + (\overline{y} - a\overline{x})$, exactement. On retrouve les valeurs de la section 1 et tous les points sont parfaitement alignés. Enfin dans les cas où $|r| \neq 1$, la seule conclusion possible est que les points ne sont pas alignés.

2.3 Ecart moyen résiduel

2.3.1 Comparaison de deux échantillons d'un même modèle

On suppose ici que l'on compare deux séries d'expériences sur un *même* phénomène linéaire (i.e. les points devraient être alignés) et que l'on veut en comparer les résultats. En utilisant l'interprétation de la section 2.2, on peut alors dire que l'échantillon dont le coefficient de corrélation est le plus proche de un reflète le mieux la linéarité du problème.

2.3.2 Comparaison de deux modèles pour un même échantillon

Dans ce cas on dispose d'une seule série d'expériences et l'on voudrait comparer la linéarité de différentes fonctions appliquées aux valeurs de l'échantillon. Malheureusement, dans ce cas il est très difficile de conclure. En effet l'action de différentes fonctions modifie les échelles respectives d'observation. Le coefficient de corrélation est aussi affecté par ces modifications ; ainsi pour en comparer les valeurs il faudrait les comparer relativement à une même échelle de mesure et non dans l'absolu ! Empiriquement il est clair que d'un résultat avoisinant zéro où d'un résultat très proche de un, ce dernier peut sembler meilleur (Remarquer l'usage très flou de : "très proche", "avoisinant", "sembler" ...).

Dans ce cas, pour comparer les modèles, il faut comparer les erreurs résiduelles $F(a_1, b_1)$ et $F(a_2, b_2)$. La plus faible donne le modèle le mieux adapté (il est aussi possible de regarder les moyennes de ces résidus : $\sqrt{\frac{1}{n}\sum[y_i - M_1(x_i)]^2}$ et $\sqrt{\frac{1}{n}\sum[y_i - M_2(x_i)]^2}$; on parle alors d'*écart-moyen résiduel*).