Statistiques

Regression linéaire

Julian Tugaut

Télécom Saint-Étienne

Sommaire

Plan

Introduction

Dans ce chapitre, on étudiera la régression linéaire simple et uniquement la régression linéaire simple. Il convient néanmoins de savoir que l'on peut être amené à effectuer de la régression linéaire multiple.

Introduction

Dans ce chapitre, on étudiera la régression linéaire simple et uniquement la régression linéaire simple. Il convient néanmoins de savoir que l'on peut être amené à effectuer de la régression linéaire multiple.
On suppose que l'on a un n-échantillon de valeurs $(x[i], y[i])$ où $x[i]$ est la valeur du caractère quantitatif X pour l'individu i tandis que $y[i]$ est celle du caractère Y.

Plan

Équation de la droite de régression - 1

On souhaite obtenir une droite qui s'ajuste "au mieux" au nuage de points. Il s'avère que les mots "au mieux" peuvent varier selon le modèle que l'on utilise.

Équation de la droite de régression - 1

On souhaite obtenir une droite qui s'ajuste "au mieux" au nuage de points. Il s'avère que les mots "au mieux" peuvent varier selon le modèle que l'on utilise.
On suppose ici que l'on a

$$y[i] = \alpha x[i] + \beta + \varepsilon[i]\,,$$

où ε est un Bruit Blanc de variance σ^2. En d'autres termes, pour tout i, on a $\mathbb{E}[\varepsilon[i]] = 0$, $\mathrm{Var}[\varepsilon[i]] = \sigma^2$ et de plus, $\mathbb{E}[\varepsilon[i]\varepsilon[j]] = 0$ si $i \neq j$.

Équation de la droite de régression - 2

De fait, si $\mathbb{E}[\varepsilon[i]^4] < +\infty$, on dispose de la limite

$$\lim_{n \to +\infty} \frac{1}{n} \sum_{i=1}^{n} \varepsilon[i]^2 = \sigma^2 \,,$$

d'après la loi forte des grands nombres.

Équation de la droite de régression - 2

De fait, si $\mathbb{E}[\varepsilon[i]^4] < +\infty$, on dispose de la limite

$$\lim_{n \to +\infty} \frac{1}{n} \sum_{i=1}^{n} \varepsilon[i]^2 = \sigma^2 \,,$$

d'après la loi forte des grands nombres.
Ainsi, minimiser le bruit revient à minimiser la moyenne des distances au carré :

$$D(\alpha, \beta) := \frac{1}{n} \sum_{i=1}^{n} |y[i] - \alpha x[i] - \beta|^2 \,.$$

Équation de la droite de régression - 3

On dispose ici d'une fonction de $\mathbb{R}^2$ dans $\mathbb{R}_+$. De plus, la fonction D est de la forme :

$$D(\alpha, \beta) = \beta^2 + \alpha^2 \frac{1}{n} \sum_{i=1}^{n} x[i]^2 + 2\alpha\beta \frac{1}{n} \sum_{i=1}^{n} x[i] + L(\alpha, \beta)\,,$$

où L est une fonction affine. On regarde donc la forme quadratique $Q(\alpha, \beta) := \beta^2 + \alpha^2 \frac{1}{n} \sum_{i=1}^{n} x[i]^2 + 2\alpha\beta \frac{1}{n} \sum_{i=1}^{n} [i]$. La trace et le déterminant de la matrice associée à cette forme quadratique sont positifs. De fait, $D(\alpha, \beta)$ tend vers l'infini dès que $\alpha^2 + \beta^2$ tend vers l'infini.

Équation de la droite de régression - 4

On en déduit que le minimiseur de D est atteint en un point critique. On regarde maintenant les dérivées partielles de D par rapport à α et à β :

$$\frac{\partial}{\partial \alpha} D(\alpha, \beta) = -\frac{2}{n} \sum_{i=1}^{n} x[i] \left(y[i] - \alpha x[i] - \beta\right)$$

ainsi que

$$\frac{\partial}{\partial \beta} D(\alpha, \beta) = -\frac{2}{n} \sum_{i=1}^{n} \left(y[i] - \alpha x[i] - \beta\right) .$$

Équation de la droite de régression - 5

Ainsi, on trouve $\beta_0 = \frac{1}{n}\sum_{i=1}^{n} y[i] - \alpha_0 \frac{1}{n}\sum_{i=1}^{n} x[i]$. Cette égalité implique $\beta_0 = \overline{y} - \alpha_0 \overline{x}$, ce qui traduit l'absence de direction privilégiée dans le Bruit Blanc.

Équation de la droite de régression - 5

Ainsi, on trouve $\beta_0 = \frac{1}{n}\sum_{i=1}^{n} y[i] - \alpha_0 \frac{1}{n}\sum_{i=1}^{n} x[i]$. Cette égalité implique $\beta_0 = \overline{y} - \alpha_0 \overline{x}$, ce qui traduit l'absence de direction privilégiée dans le Bruit Blanc.
Également, on a

$$\overline{x^2}\alpha_0 = \overline{xy} - \beta_0 \overline{x} = \overline{xy} - \overline{y}\ \overline{x} + \alpha_0 \overline{x}^2 \,.$$

Équation de la droite de régression - 6

Il s'ensuit

$$\alpha_0 = \frac{s_{xy}}{s_x^2} \text{ et } \beta_0 = \overline{y} - \overline{x}\frac{s_{xy}}{s_x^2},$$

où s_{xy} est la covariance des séries marginales tandis que s_x^2 est la variance empirique de la série x.

Plan

Lien avec $r_{x,y}$ - 1

On peut maintenant se demander la valeur du bruit moyen σ^2. On estime celle-ci par

$$D(\alpha_0, \beta_0) = \frac{1}{n} \sum_{i=1}^{n} |y[i] - \alpha_0 x[i] - \beta_0|^2 \ .$$

Or, $\beta_0 = \overline{y} - \alpha_0 \overline{x}$ donc on en déduit :

Lien avec $r_{x,y}$ - 2

$$
\begin{aligned}
& D(\alpha_0, \beta_0) \\
&= \frac{1}{n} \sum_{i=1}^{n} |(y[i] - \overline{y}) - \alpha_0 (x[i] - \overline{x})|^2 \\
&= \frac{1}{n} \sum_{i=1}^{n} |y[i] - \overline{y}|^2 + \alpha_0^2 \frac{1}{n} \sum_{i=1}^{n} |x[i] - \overline{x}|^2 - 2\alpha_0 \frac{1}{n} \sum_{i=1}^{n} (y[i] - \overline{y})(x[i] - \overline{x}) \\
&= s_y^2 + \left(\frac{s_{xy}}{s_x^2} \right)^2 s_x^2 - 2 \frac{s_{xy}}{s_x^2} s_{xy} \\
&= s_y^2 \left(1 + \frac{s_{xy}^2}{s_x^2 s_y^2} - 2 \frac{s_{xy}^2}{s_x^2 s_y^2} \right) \\
&= s_y^2 \left(1 - r_{xy}^2 \right) ,
\end{aligned}
$$

où r_{xy} est le coefficient de corrélation linéaire.

Lien avec $r_{x,y}$ - 3

On utilise aussi la valeur R^2 (sur excel par exemple) qui correspond au carré du coefficient de corrélation linéaire entre y et $\hat{y}$ avec $\hat{y}[i] := \alpha_0 x[i] + \beta_0 = \frac{s_{xy}}{s_x^2} x[i] + \overline{y} - \overline{x}\frac{s_{xy}}{s_x^2}$. Ce coefficient de corrélation est donc égal à

$$\begin{aligned} r_{y\hat{y}} &= \frac{1}{n}\sum_{i=1}^{n} y[i]\left(\frac{s_{xy}}{s_x^2}x[i] + \overline{y} - \overline{x}\frac{s_{xy}}{s_x^2}\right) - \overline{y}^2 \\ &= \frac{s_{xy}\overline{xy}}{s_x^2} + \overline{y}^2 - \overline{x}\ \overline{y}\frac{s_{xy}}{s_x^2} - \overline{y}^2 \\ &= \frac{s_{xy}^2}{s_x^2}\ . \end{aligned}$$

Lien avec $r_{x,y}$ - 3

On utilise aussi la valeur R^2 (sur excel par exemple) qui correspond au carré du coefficient de corrélation linéaire entre y et $\hat{y}$ avec $\hat{y}[i] := \alpha_0 x[i] + \beta_0 = \frac{s_{xy}}{s_x^2} x[i] + \overline{y} - \overline{x}\frac{s_{xy}}{s_x^2}$. Ce coefficient de corrélation est donc égal à

$$\begin{aligned} r_{y\hat{y}} &= \frac{1}{n}\sum_{i=1}^{n} y[i]\left(\frac{s_{xy}}{s_x^2} x[i] + \overline{y} - \overline{x}\frac{s_{xy}}{s_x^2}\right) - \overline{y}^2 \\ &= \frac{s_{xy}\overline{xy}}{s_x^2} + \overline{y}^2 - \overline{x}\ \overline{y}\frac{s_{xy}}{s_x^2} - \overline{y}^2 \\ &= \frac{s_{xy}^2}{s_x^2}\ . \end{aligned}$$

Ce coefficient correspond à la valeur résiduelle de la variance de y : $s_y^2 - \sigma^2$. Il s'agit en d'autres termes de la variance expliquée.