Condensé de cours

Régression linéaire multiple

Définition

Soient p données statistiques $x_k = (x_{1,k}, \dots, x_{n,k})$ et $y = (y_1, \dots, y_n)$ des données statistiques.
Les variables x_k sont appelées **variables exogènes** ou *expliquées*.
La variable y est appelée **variable endogène** ou *à expliquer*.
Une **modélisation linéaire multiple** consiste à considérer les variables aléatoires

$$Y_i = a_0 + a_1 x_{i,1} + a_2 x_{i,2} + \cdots + a_p x_{i,p} + \varepsilon_i$$

où es ε_i sont des variables aléatoires i.i.d. appelés **termes d'erreurs** et suivent une loi normale $\mathcal{N}(0, \sigma)$.

Proposition

Avec les notations de la définition précédente, un modèle linéaire multiple

$$Y_i = a_0 + a_1 x_{i,1} + a_2 x_{i,2} + \cdots + a_p x_{i,p} + \varepsilon_i$$

est équivalent à

$$Y = xm + e$$

Où x est une matrice à n lignes et $p+1$ colonnes, m le vecteur a estimer de dimension $p+1$ et e est un vecteur gaussien de dimension n.

Théorème

Avec les notations précédentes, $e = Y - xm$ et le minimum de ${}^t ee$ est atteint lorsque

$$m = \hat{m} \overset{\text{def}}{=} ({}^t xx)^{-1\,t} xy$$

Définition

Avec les notations précédentes, on appel **résidu du modèle** le vecteur $\hat{e} = y - \hat{y}$ où $\hat{y} = x\hat{m}$

Définition

Avec les notations précédentes on définit le **coefficient de détermination du modèle**, noté R^2, par :

$$R^2_{x,y} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2}$$

Proposition

Avec les notations précédentes

$$R^2_{x,y} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2} = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2}$$

En particulier $R^2_{x,y} \in [0, 1]$.

Proposition

Avec les notations précédentes, $\mathbb{V}(M_n) = \sigma^2({}^t x x)^{-1}$

Corollaire

Si la matrice $({}^t x x)^{-1}$ tend vers la matrice nulle alors l'estimateur M_n est convergent.

Théorème

Soit $\mathbf{\varepsilon} = Y - \hat{y}$, alors

$$\mathbb{E}(\mathbf{\varepsilon}) = (n - (p+1))\sigma^2$$

Corollaire

La variable aléatoire

$$S_n = \frac{{}^t\mathbf{\varepsilon}\mathbf{\varepsilon}}{n - (p+1)}$$

est un estimateur convergent et sans biais de σ^2. De plus :

1. $S_n^M = S_n({}^txx)^{-1}$ est un estimateur convergent et sans biais de $\mathbb{V}(M_n)$.
2. En particulier, les coefficient diagonaux de S_n^M sont des estimateurs convergent et sans biais de la variance des coordonnées de M_n.
3. $(n - (p+1))\dfrac{S_n}{\sigma^2} \sim \chi^2(n - (p+1))$

Corollaire

Soit $(M_n)_i$ la i-ieme coordonnée de M_n et S_i^M le coefficient diagonale de la i-ième ligne de S_n^m
La variable aléatoire $\dfrac{M_{n,i} - m_i}{\sqrt{S_i^M}}$ suit une loi de Student à $n - (p+1)$ degrés de libertés.

Corollaire

Soient $0 < \beta < \alpha < 1$, $t_1 = Q_{\mathcal{T}(n-(p+1))}(\beta)$ et $t_2 = Q_{\mathcal{T}(n-(p+1))}(1 - \alpha + \beta)$ alors

$$\left[(M_n)_i - t_2\sqrt{S_i^M}; (M_n)_i - t_1\sqrt{S_i^M}\right]$$

est un intervalle de confiance $1 - \alpha$ de m_i.