Modèle linéaire gaussien

14 mars 2011 - v0 non débuguée ; 7 avril 2014 v0.1 non relue sérieusement

1 Théorème de Cochran

Théorème 1.1 *Soit X un vecteur de loi gaussienne centrée réduite dans l'espace euclidien $\mathbb{R}^n$. Soit L un sous-espace de $\mathbb{R}^n$ de dimension d où $1 \leq d \leq n-1$. On désigne par X_L la projection orthogonale de X sur L. On désigne par $X_{L^\perp}$ la projection orthogonale de X sur $L^\perp$. Alors :*

- *Les v.a. X_L et $X_{L^\perp}$ sont indépendantes.*
- *Les coordonnées de X_L dans une base orthonormale sont des gaussiennes centrée réduites indépendantes. La variable aléatoire $\|X_L\|^2$ suit donc une loi de Pearson à d degrés de liberté.*
- *Les coordonnées de $X_{L^\perp}$ dans une base orthonormale sont des gaussiennes centrée réduites indépendantes. La variable aléatoire $\|X_{L^\perp}\|^2$ suit donc une loi de Pearson à $n-d$ degrés de liberté.*

Idées de la preuve. L'image d'un vecteur de loi gaussienne centrée réduite par une isométrie est un vecteur de loi gaussienne centrée réduite. Cela permet de montrer que, dans toute base orthonormale, les coordonnées d'un vecteur de loi gaussienne centrée réduite sont des variables aléatoires réelles de loi gaussienne centrée réduite. □

Quelques utilisations classiques. Théorème de Fisher, test du χ^2, modèle linéaire gaussien, ...

2 Modèle linéaire gaussien

Modèle. Une quantité y dépend d'un paramètre x par une relation affine :

$$y = ax + b.$$

Les constantes a et b sont inconnues. On effectue plusieurs mesures de la quantité y en faisant varier x. On suppose que chaque mesure est affectée d'une erreur aléatoire. On suppose que les erreurs sont normales et centrées. On suppose enfin que les erreurs affectant les différentes mesures sont indépendantes et de même loi. Autrement dit, on dispose pour un ensemble de paramètres $x_1, \ldots, x_n$ de mesures Y_i vérifiant :

$$Y_i = ax_i + b + \varepsilon_i$$

où les ε_i sont i.i.d. normaux et centrés. L'objectif est de donner des intervalles de confiances pour a, b, $ax+b$ pour un x fixé etc.

Moindres carrés. L'idée de base est de déterminer des estimations de a et de b par la méthode des moindres carrés. On cherche les valeurs $\widehat{a}$ et $\widehat{b}$ minimisant la quantité

$$D^2 = \sum_{i=1}^{n} (Y_i - \widehat{y}_i)^2$$

avec

$$\widehat{y}_i = \widehat{a}x_i + \widehat{b}.$$

Considérons les vecteurs $x = (x_1, \ldots, x_n)$, $Y = (Y_1, \ldots, Y_n)$ et $\mathbb{1} = (1, \ldots, 1)$ de $\mathbb{R}^n$. On a

$$D^2 = \sum_{i=1}^{n} \left(Y_i - (\widehat{a}x_i + \widehat{b})\right)^2 = \|Y - (\widehat{a}x + \widehat{b}\mathbb{1})\|_2^2.$$

Le vecteur $\widehat{a}x + \widehat{b}\mathbb{1}$ minimisant la quantité précédente est donc le projeté orthogonal de Y sur $F = \mathbb{R}x + \mathbb{R}\mathbb{1}$. Pour expliciter cette projection, on considère la base orthonormale suivante de F :

$$e_1 = \frac{1}{\sqrt{n}}\mathbb{1},\ e_2 = \frac{1}{\sqrt{n}s}(x - \overline{x}\mathbb{1})$$

où

$$\overline{x} = \frac{1}{n}\sum_i x_i \text{ et } s^2 = \frac{1}{n}\sum_i (x_i - \overline{x})^2.$$

On obtient :

$$\begin{aligned} \widehat{a}x + \widehat{b}\mathbb{1} &= < Y, e_2 > e_2 + < Y, e_1 > e_1 \\ &= \frac{1}{\sqrt{n}s} < Y, e_2 > x + \left(-\frac{1}{\sqrt{n}s} < Y, e_2 > \overline{x} + \frac{1}{\sqrt{n}} < Y, e_1 > \right) \mathbb{1}. \end{aligned}$$

En identifiant, on obtient :

$$\widehat{a} = \frac{1}{\sqrt{n}s} < Y, e_2 > \text{ et } \widehat{b} = \overline{Y} - \widehat{a}\overline{x}$$

où

$$\overline{Y} = \frac{1}{n}\sum_i Y_i.$$

En explicitant, on obtient :

$$\widehat{a} = \frac{1}{ns^2} < Y, x - \overline{x}\mathbb{1} > = \frac{\sum_i Y_i(x_i - \overline{x})}{ns^2} = \frac{\sum_i (Y_i - \overline{Y})(x_i - \overline{x})}{ns^2}$$

Notons que, si les $x_i - \overline{x}$ sont tous non nuls, on peut écrire :

$$\widehat{a} = \frac{\sum_i \frac{Y_i - \overline{Y}}{x_i - \overline{x}}(x_i - \overline{x})^2}{\sum_i (x_i - \overline{x})^2}.$$

Intervalle de confiance pour a. Rappelons que $\widehat{a}x + \widehat{b}\mathbb{1}$ est le projeté orthogonal de Y sur F. Notons E le projeté orthogonal de Y sur $F^\perp$. On a donc :

$$Y = \widehat{a}x + \widehat{b}\mathbb{1} + E$$

et $E = (E_1, \ldots, E_n)$ où $E_i = Y_i - \widehat{a}x_i - \widehat{b}$. On appelle E le vecteur des résidus.

On a par ailleurs :

$$Y = ax + b\mathbb{1} + \varepsilon$$

où $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$ est un vecteur gaussien centrée de matrice de variance σI pour un certain $\sigma > 0$. On a donc :

$$\begin{aligned} \varepsilon &= (\widehat{a} - a)x + (\widehat{b} - b)\mathbb{1} + E \\ &= (\widehat{a} - a)\sqrt{n}se_2 + \big((\widehat{a} - a)\sqrt{n}\overline{x} + (\widehat{b} - b)\sqrt{n}\big)e_1 + E. \end{aligned}$$

On constate que le premier vecteur est dans F tandis que E est dans $F^\perp$. Par conséquent, le premier vecteur est la projection orthogonale de ε sur F tandis que E est la projection orthogonale de ε sur $F^\perp$.

Par le théorème de Cochran, on en déduit donc :

- Le premier vecteur aléatoire est indépendant de E. Par conséquent $\sqrt{n}s(\widehat{a} - a)$ est indépendant de E.
- La variable aléatoire $\|E\|^2/\sigma^2$ suit la loi $\chi^2(n-2)$.
- La variable aléatoire$\sqrt{n}s(\widehat{a} - a)/\sigma$ suit la loi normale centrée réduite.

Posons :

$$\widehat{\sigma}^2 = \frac{1}{n-2}\|E\|^2.$$

Par ce qui précède on a :

Théorème 2.1 *La variable aléatoire*

$$\sqrt{n}\, s\frac{\widehat{a} - a}{\widehat{\sigma}}$$

suit une loi de Student à $n-2$ degrés de liberté.

Ce résultat permet d'obtenir des intervalles de confiance pour a. On pourrait aussi obtenir des intervalles de confiance pour b ou pour $ax + b$.