TP 3 : Régression linéaire

1 Rappels de théorie (cf. poly)

On suppose que l'on dispose de données $x_1,\dots,x_n \in \mathbb{R}^p$ et $y_1,\dots,y_n \in \mathbb{R}$ liées par une relation de la forme

$$y_k = f(x_k) + \varepsilon_k,$$

où $f(x) = {}^t\!ax + b$ est une fonction affine $\mathbb{R}^p \to \mathbb{R}$ inconnue et $\varepsilon_1,\dots,\varepsilon_n$ sont des erreurs de mesures, inconnues elles aussi. L'objectif est de déterminer la fonction f à partir des données.
Sous forme matricielle, ceci s'écrit

$$Y = X\Theta + \varepsilon,$$

où

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,p} \\ 1 & x_{2,1} & \cdots & x_{2,p} \\ & & \cdots & \\ 1 & x_{n,1} & \cdots & x_{n,p} \end{pmatrix}, \quad \Theta = \begin{pmatrix} b \\ a_1 \\ \vdots \\ a_p \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Pour que Θ soit défini de façon unique, il faut que X soit une matrice injective, donc de rang maximal $\mathrm{rg}(X) = p+1$. En particulier, il faut bien sûr $n \geq p+1$.
On peut considérer ce modèle de façon probabiliste : ε est aléatoire, et par suite Y aussi, tandis que X et Θ sont fixés. Le modèle a notamment des propriétés intéressantes lorsque $\varepsilon_1,\dots,\varepsilon_n$ sont indépendantes et de même loi $\mathcal{N}(0,\sigma^2)$.

On cherche le vecteur $\widehat{\Theta}$ qui minimise $\|Y - X\widehat{\Theta}\|_2$, autrement dit la fonction affine $\widehat{f}$ qui minimise $\sum_k |y_k - \widehat{f}(x_k)|^2$. C'est l'« estimateur des moindres carrés », ou « régression linéaire ».

Calcul de $\widehat{\Theta}$. Par définition, $X\widehat{\Theta}$ est la projection orthogonale de Y sur $E = \mathrm{Im}(X)$. Autrement dit, $Y = X\widehat{\Theta} + Z$ où $Z \perp E$. On a donc ${}^t\!XZ = 0$ d'où

$${}^t\!XY = {}^t\!XX\widehat{\Theta},$$

et l'hypothèse $\ker X = \{0\}$ implique que ${}^t\!XX$ est inversible (si ${}^t\!XXu = 0$, alors $0 = {}^t\!u{}^t\!XXu = \|Xu\|^2$ d'où $Xu = 0$ puis $u = 0$), donc

$$\widehat{\Theta} = ({}^t\!XX)^{-1}({}^t\!X)Y.$$

Le cas où les x_k sont réels ($p = 1$) admet une expression simple : alors $\widehat{\Theta} = \begin{pmatrix} \widehat{b} \\ \widehat{a} \end{pmatrix}$ avec

$$\widehat{a} = \frac{\mathrm{Cov}(x,y)}{\mathrm{Var}(x)} = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2}$$

où $\overline{z} = \frac{1}{n}\sum_{k=1}^n z_k$ (avec $z_k = x_k y_k$, x_k, etc.), et $\widehat{b}$ se déduit de $\overline{y} = \widehat{a}\overline{x} + \widehat{b}$.

Cas gaussien. Si $\varepsilon \sim \mathcal{N}(0,\sigma^2 I_n)$, alors $Y \sim \mathcal{N}(X\Theta,\sigma^2 I_n)$. Ainsi, pour tout $y \in \mathbb{R}^n$, la densité de Y en y vaut $f_\Theta(y) = (2\pi\sigma^2)^{-n/2} e^{-\|y - X\Theta\|^2/(2\sigma^2)}$ et est maximale (comme fonction de Θ) quand $\|y - X\Theta\|^2$ est minimale : l'estimateur des moindres carrés est aussi ici l'estimateur du maximum de vraisemblance.
On pose $\widehat{Y} = X\widehat{\Theta} = P_E(Y)$. On a $Y - \widehat{Y} = P_{E^\perp}(Y)$, donc le théorème de Cochran montre que $Y - \widehat{Y}$ est indépendant de $\widehat{Y}$ (et donc de $\widehat{\Theta}$), et que $\|Y - \widehat{Y}\|_2^2$ suit la loi $\chi^2_{n-(p+1)}$. En particulier, la variable aléatoire

$$\widehat{\sigma}^2 = \frac{\|Y - \widehat{Y}\|^2}{n - (p+1)}$$

est un estimateur sans biais de σ^2. On pourrait aussi déduire des régions de confiance pour Θ (voir poly).

2 Dans Scilab

Si `X` est une matrice de taille (p,n) dont les *colonnes* sont les différentes données x_k, et `Y` est un vecteur colonne de taille n, `[a,b]=reglin(X,Y)` renvoie le vecteur ligne `a` et le réel `b` tels que `a*X+b` est l'estimateur de `Y` au moindres carrés.
De plus, `[a,b,s]=reglin(x,y)` permet d'obtenir également $\widehat{\sigma}$.

NB. L'exposé précédent utilise la présentation usuelle en statistique $Y = X\Theta + \varepsilon$, tandis que Scilab considère le modèle équivalent $Y = aX + b + \varepsilon$, ce qui revient à transposer X, sans ajouter une série de 1, et à avoir ${}^t\Theta = \begin{pmatrix} b & a \end{pmatrix}$.

1) Pour $n = 50$ et $p = 1$: prendre `X=1:n; Y=2*X-7+grand(1,n,"nor",0,0.3)`, et retrouver les coefficients 2 et -7 par `reglin`, puis par les formules précédentes. Calculer $\widehat{\sigma}$ et comparer avec la valeur fournie par `reglin`.

2) Pour $n = 50$ et $p = 2$: partir de `X=[(1:n);(1:n)^2]; Y=[2 -1]*X+5+grand(1,n,"nor",0,0.3)`, et adapter les questions précédentes.

3) Calculer les 100 premiers termes de la suite $(u_n)_{n\geq 0}$ définie par récurrence par $u_0 = 1$ puis $u_{n+1} = \sin(u_n)$. Représenter graphiquement $\log(u_n)$ en fonction de $\log n$; qu'est-ce que ceci suggère ? Utiliser `reglin` pour deviner l'ordre de grandeur de u_n. Comparer graphiquement cette estimation avec la suite u_n.