Modèles linéaires

1 Régression linéaire.

On considère la régression linéaire gaussienne simple donnée pour $n > 2$ et $i = 1, \cdots, n$, par

$$Y_i = a + bx_i + \varepsilon_i$$

où (x_i) est une suite de nombres réels connus non tous égaux et où (ε_i) est une suite de variables aléatoires indépendantes et de même loi $\mathcal{N}(0, \sigma^2)$. Soit $\widehat{\theta} = (\widehat{a}, \widehat{b})$ et $\widehat{\sigma}^2$ les estimateurs des moindres carrés de $\theta = (a, b)$ et σ^2. Si $X = (\mathbb{1}_n \; x)$, on a $\widehat{\theta} \sim \mathcal{N}(\theta, \sigma^2(X^tX)^{-1})$ donc

$$\widehat{a} \sim \mathcal{N}\Big(a, \frac{\sigma^2\overline{x^2}}{n\mathrm{Var}(x)}\Big) \qquad \text{et} \qquad \widehat{b} \sim \mathcal{N}\Big(b, \frac{\sigma^2}{n\mathrm{Var}(x)}\Big).$$

Par le théorème de Cochran, les estimateurs $\widehat{\theta}$ et $\widehat{\sigma^2}$ sont indépendants et $(n-2)\widehat{\sigma^2} \sim \sigma^2\chi^2(n-2)$. De plus, on a

$$\sqrt{\frac{n\mathrm{Var}(x)}{\overline{x^2}}}\Big(\frac{\widehat{a}-a}{\widehat{\sigma}}\Big) \sim t(n-2) \qquad \text{et} \qquad \sqrt{n\mathrm{Var}(x)}\Big(\frac{\widehat{b}-b}{\widehat{\sigma}}\Big) \sim t(n-2).$$

On peut donc effectuer des tests sur a et b et obtenir des intervalles de confiance pour a et b. De plus, si $a = 0$, on peut montrer que $\sum_{i=1}^n (\widehat{a} + \widehat{b}x_i - \tilde{b}x_i)^2/\widehat{\sigma}^2 \sim F(1, n-2)$ avec $\tilde{b} = \overline{xY}/\overline{x^2}$. Enfin, si $b = 0$, on peut montrer que $\sum_{i=1}^n (\widehat{a} + \widehat{b}x_i - \overline{Y})^2/\widehat{\sigma}^2 \sim F(1, n-2)$. On peut ainsi tester H_0 :"$a = 0$" contre H_1 :"$a \neq 0$" ou bien H_0 :"$b = 0$" contre H_1 :"$b \neq 0$".

Exercice 1. Créer un code Matlab permettant de générer une régression linéaire gaussienne simple où les valeurs n, a, b et σ^2 sont affectées par l'utilisateur et où (x_i) est une réalisation d'un n-échantillon de loi uniforme sur $[0, 1]$. Calculer les estimateurs des moindres carrés $\widehat{\theta} = (\widehat{a}, \widehat{b})$ et $\widehat{\sigma}^2$. Représenter graphiquement le nuage de points formé par les couples (x_i, y_i) et tracer la droite des moindres carrés $y = \widehat{a} + \widehat{b}x$. Donner pour chaque paramètre a, b et σ^2 un intervalle de confiance de risque $\alpha = 5\%$. Tracer l'ellipsoïde de confiance à l'intérieur duquel se trouve $\theta = (a, b)$. Reprendre cet exercice en faisant varier n, a, b et σ^2 ainsi que la loi associée à (x_i).

Exercice 2. On appelle fréquence seuil d'un sportif amateur, sa fréquence cardiaque obtenue après trois quarts d'heure d'un effort soutenu de course à pied. Elle est mesurée à l'aide d'un cardio-fréquence-mètre. On cherche à savoir si l'âge d'un sportif a une influence sur sa fréquence seuil. On dispose des résultats suivants où x_i représente l'âge du sportif et y_i sa fréquence seuil.

x_i	30	54	29	51	36	41	40	23	49	30
y_i	175	165	169	172	170	170	167	170	166	167

x_i	32	22	22	32	44	34	32	20	46	45
y_i	177	169	172	173	168	169	170	172	175	168

Créer un code Matlab permettant d'étudier cette régression linéaire simple que l'on supposera gaussienne puis comparer vos résultats avec ceux obtenus grace à *linreg* de Matlab.

2 Analyse de la variance.

On considère le modèle d'analyse de la variance donné pour $p \geq 2$, $i = 1, \cdots, p$ et $j = 1, \cdots, n_i$, par

$$Y_{ij} = m_i + \varepsilon_{ij}$$

où $m_i \in \mathbb{R}$ et (ε_i) est une suite de variables aléatoires indépendantes et de même loi $\mathcal{N}(0, \sigma^2)$. Le nombre d'observations de chaque classe $n_i \geq 1$ et le nombre total d'observations $n > p + 1$. Soit $\widehat{m}$ et $\widehat{\sigma}^2$ les estimateurs des moindres carrés de m et σ^2. Si X est la matrice des effectifs associés au modèle, on a $\widehat{m} \sim \mathcal{N}(m, \sigma^2(X^tX)^{-1})$ et $(n-p)\widehat{\sigma^2} \sim \sigma^2\chi^2(n-p)$.

Exercice 3. Proposer un test d'égalité des moyennes basé sur la somme des carrés intra-groupe et la somme des carrés intergroupe.

Exercice 4. On cherche à savoir si le mode de fabrication d'une ampoule électrique est différent selon la marque en terme de la durée de vie de l'ampoule. On dispose des durées de vie suivantes obtenues pour six marques différentes notées $A_1, A_2, \cdots, A_6$ sur 36 ampoules.

A_1	1602	1615	1617	1624				
A_2	1480	1482	1485	1493	1500	1507	1510	
A_3	1548	1555	1559	1563	1575			
A_4	1435	1438	1448	1449	1454	1458	1467	1475
A_5	1493	1498	1500	1502	1509	1510		
A_6	1596	1599	1602	1604	1612	1620		

Créer un code Matlab permettant d'estimer les paramètres inconnus m et σ^2 du modèle. Tester dans le cadre gaussien l'hypothèse d'égalité de la durée moyenne de vie pour les six marques en dressant le tableau d'analyse de la variance. Comparer vos résultats avec ceux obtenus grace à *anova*1 de Matlab.