Partiel de Statistique

Mercredi 14 janvier 2009, 8h00-10h00

Partiel sans documents (une feuille manuscrite A4 est autorisée)

Exercice 1 : Estimation

On considère n couples de données $(x_i, y_i), i = 1, ..., n$ et on cherche la droite d'équation $y = ax$ la plus proche de ces couples de données. Pour cela, on introduit un modèle appelé modèle de régression linéaire défini par

$$Y_i = ax_i + e_i,$$

dans lequel a est une constante et e_i modélise l'erreur entre Y_i et ax_i. On suppose que les erreurs e_i sont des variables aléatoires indépendantes et de même loi normale $\mathcal{N}(0, \sigma^2)$ et que $x = (x_1, ..., x_n)$ n'est pas le vecteur nul. On fera attention de noter que dans le modèle ci-dessus, x_i est une valeur déterministe (non aléatoire) observée tandis que Y_i est une variable aléatoire de loi normale $\mathcal{N}(ax_i, \sigma^2)$.

Partie 1 : variance σ^2 connue

1) Déterminer la vraisemblance de $(y_1, ..., y_n)$ et en déduire l'estimateur du maximum de vraisemblance du paramètre a noté $\widehat{a}_{\text{MV}}$ (on pourra pour simplifier l'expression de $\widehat{a}_{\text{MV}}$ utiliser la notation $\|x\|^2 = \sum_{i=1}^n x_i^2$).
2) Montrer que l'estimateur $\widehat{a}_{\text{MV}}$ est un estimateur non biaisé de a.
3) Déterminer la borne de Cramer-Rao pour les estimateurs non biaisés du paramètre a. L'estimateur $\widehat{a}_{\text{MV}}$ est-il l'estimateur efficace du paramètre a ?
4) Quelle est la loi de l'estimateur $\widehat{a}_{\text{MV}}$? En déduire un intervalle de confiance pour le paramètre a construit à partir de $\widehat{a}_{\text{MV}}$ et de

$$G(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) du.$$

5) On suppose désormais qu'on dispose d'une information a priori sur le paramètre a résumée dans la loi normale $\mathcal{N}(m_a, \sigma_a^2)$. Montrer que la loi a posteriori de $a|\, y_1, ..., y_n$ est une loi normale dont on précisera la moyenne et la variance. En déduire l'estimateur du maximum a posteriori du paramètre a noté $\widehat{a}_{\text{MAP}}$. Expliquer le comportement de $\widehat{a}_{\text{MAP}}$ lorsque $\sigma_a^2 \to 0$ et lorsque $\sigma_a^2 \to \infty$.

Partie 2 : variance σ^2 inconnue

1) Déterminer l'estimateur du maximum de vraisemblance du vecteur $\theta = (a, \sigma^2)$ noté $\widehat{\theta}_{\text{MV}} = \left(\widehat{a}_{\text{MV}}, \widehat{\sigma^2}_{\text{MV}}\right)$.
2) Déterminer $E\left[\widehat{\theta}_{\text{MV}}\right]$ et en déduire un estimateur non biaisé de θ.
3) Déterminer les variances minimales d'estimateurs non biaisés de a et σ^2.

Exercice 2 : Test de Neyman-Pearson

On dispose toujours de n couples de données $(x_i, y_i), i = 1, ..., n$, où y_i est une réalisation de la variable aléatoire

$$Y_i = ax_i + e_i, \text{ avec } e_i \sim \mathcal{N}(0, \sigma^2)$$

Comme dans l'exercice précédent, on suppose que les erreurs e_i sont des variables aléatoires indépendantes et de même loi et que $x = (x_1, ..., x_n)$ n'est pas le vecteur nul.On désire tester les deux hypothèses

$$\begin{aligned} H_0 &: a = a_0 > 0, \\ H_1 &: a = a_1 \text{ avec } a_1 > a_0, \end{aligned}$$

où a_0, a_1 et σ^2 sont des paramètres connus.
1) Montrer que la statistique du test de Neyman-Pearson s'écrit

$$T_n = \sum_{i=1}^{n} x_i Y_i$$

et indiquer la région critique de ce test.
2) Exprimer les risques α et β en fonction du seuil du test noté K_α et de la fonction $G(x)$ définie dans l'exercice 1.
3) Déterminer les courbes caractéristiques opérationnelles du récepteur (courbes COR) pour le test considéré ci-dessus. Comment la performance du test dépend-elle des paramètres a_0 et a_1, de σ et de $\|x\|$? Représenter la forme des courbes COR pour différentes valeurs de $\|x\|$.

Exercice 3 : Test d'ajustement

On désire effectuer des tests d'ajustement permettant de vérifier si les variables aléatoires $Y_i = ax_i + e_i$ suivent une loi normale $\mathcal{N}(ax_i, \sigma^2)$. On suppose dans cet exercice que $x_i > 0$ pour tout $i = 1, ..., n$ et que les paramètres a et σ^2 sont connus. On introduit alors les deux hypothèses

$$\begin{aligned} H_0 &: Y_i \sim \mathcal{N}(ax_i, \sigma^2), \ i = 1, ..., n \\ H_1 &: \text{non } H_0 \end{aligned}$$

1) On introduit la variable U_i définie par

$$U_i = \frac{Y_i - ax_i}{\sigma}.$$

Expliquer comment effectuer un test du χ^2 (à quatre classes équiprobables) à l'aide des variables aléatoires $U_1, ..., U_n$.
2) On suppose dans cette partie que $x_i > 0$, pour $i = 1, ..., n$. On introduit la variable aléatoire Z_i définie par

$$Z_i = \begin{cases} 1 \text{ si } \left|\frac{Y_i}{x_i}\right| > 2a \\ 0 \text{ sinon} \end{cases},$$

et $Z = \sum_{i=1}^{n} Z_i$ qui est le nombre de données telles que $\left|\frac{Y_i}{x_i}\right| > 2a$. Déterminer

$$p_i = P\left[Z_i = 1 | H_0 \text{ vraie}\right],$$

puis la moyenne et la variance de la variable aléatoire Z sous l'hypothèse H_0 (notées respectivement $m_Z = E\left[Z | H_0\right]$ et $\sigma_Z^2 = \text{var}[Z | H_0]$) . On décide d'approcher la loi de Z sous H_0 par une loi normale $\mathcal{N}(m_Z, \sigma_Z^2)$ et on considère la stratégie de test

$$\text{Rejet de } H_0 \text{ si } Z > \frac{n}{10}.$$

Déterminer le risque α associé à cette stratégie.