

Exercices de révision

Centrale - Statistiques avancées

07 Mars 2011

A- Habitat et biodiversité

Selon l'une des (peu nombreuses) lois qui ont été énoncées en écologie, la relation entre le nombre N d'espèces présentes dans un habitat donné et la surface S de cet habitat est de la forme suivante :

$$N = a S^b, \quad (1)$$

où a et b désignent des constantes universelles. On dispose d'un relevé, pour une trentaine d'habitats, de la surface et du nombre d'espèces présentes. Expliquer comment le modèle linéaire peut être utilisé pour vérifier la validité de la loi (1), et pour estimer les constantes a et b .

B- Moindres carrés pondérés

Soit n un entier positif, et pour tout $i \in \{1, \dots, n\}$ on suppose que la variable aléatoire Y_i est égale à

$$Y_i = \sum_{k=1}^p \theta_k X_{i,k} + \sigma_i \epsilon_i,$$

où $(X_{i,1}, \dots, X_{i,p})$ désigne un vecteur de régresseurs, où $(\theta_1, \dots, \theta_p)$ désigne un paramètre inconnu de $\mathbb{R}^p$, où $\sigma_i > 0$ et où les variables $(\epsilon_i)_{1 \leq i \leq n}$ sont des variables gaussiennes centrées réduites indépendantes. On note Σ la matrice diagonale telle que, pour $1 \leq i \leq n$, $\Sigma_{i,i} = \sigma_i^2$.

1. Montrer que $\hat{\theta}_\Sigma = (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} Y$ est un estimateur sans biais de θ , et calculer sa variance.
2. Montrer que sa variance est minimale parmi tous les estimateurs linéaires sans biais de θ .
3. Proposer un test pour $H_0 : \theta_1 = 0$ contre $H_1 : \theta_1 \neq 0$.
4. Application : données agrégées. Dans le problème de régression de l'exercice ci-dessus, on suppose que les données brutes n'ont malencontreusement pas été conservées, mais que l'on dispose seulement de données agrégées par forêts : ainsi, pour la forêt numéro k , on dispose seulement de la surface moyenne S_k des n_k habitats étudiés et du nombre moyen N_k d'espèces par habitat. Expliquer, intuitivement, pourquoi dans ce cas une régression linéaire "classique" n'est pas optimale. Comment peut-on malgré cela estimer a et b le plus efficacement possible ?
5. Le logiciel R ne permet pas de spécifier toutes les valeurs $(\sigma_i)_{1 \leq i \leq n}$. Par contre, il sait résoudre le problème de minimisation des moindres carrés pondérés :

$$\min_{\theta \in \mathbb{R}^p} \sum_{i=1}^n p_i (Y_i - \theta_1 X_{i,1} - \dots - \theta_p X_{i,p})^2$$

pour toute famille de poids $(p_i)_{1 \leq i \leq n}$ positifs. Peut-on quand même s'en sortir avec cela ?

6. (bonus) Que faire si la matrice de covariance du vecteur aléatoire $(\epsilon_1, \dots, \epsilon_n)$ n'est pas diagonale ?

C- Test gaussien vs Mann-Whitney

Soient $X_1, \dots, X_m$ et $Y_1, \dots, Y_n$ deux échantillons de lois respectives $\mathcal{N}(\mu_1, \sigma^2)$ et $\mathcal{N}(\mu_2, \sigma^2)$. On suppose que $\sigma^2 = 1$ est connue, et on cherche à tester $H_0 : \mu_1 = \mu_2$ contre $H_1 : \mu_1 > \mu_2$. On suppose qu'en fait $\mu_1 = 1$ et $\mu_2 = 0$.

- Calculer, en fonction de m et de n , la probabilité de rejeter H_0 avec test de niveau de confiance $1 - \alpha$ basé sur la statistique suivante :

$$Z_{m,n} = \sqrt{\frac{mn}{m+n}} (\bar{X}_m - \bar{Y}_n)$$

- Calculer, quand m et n sont "grands", la probabilité de rejeter H_0 avec le test de Mann-Whitney.
- Comparer les deux tests.