TP 11 - Modèle linéaire gaussien

1 Modèle de régression simple

On considère tout d'abord dans ce TP le modèle de régression linéaire simple

$$Y_i = a + bX_i + \epsilon_i \qquad i = 1 \dots n$$

où a et b sont les paramètres inconnus du modèle et les $(\epsilon_i)_{i=1\dots n}$ sont des variables aléatoires indépendantes gaussiennes centrées de variance σ^2 inconnue. L'objectif consiste à construire un ensemble de confiance $A_\alpha(X, Y)$ de niveau α à partir des données (X_i, Y_i), ensemble tel que

$$P\left(a + bx \notin A_\alpha(X, Y)\right) \leq \alpha$$

1. Rappeler le principe de l'algorithme de Box-Muller et écrire une fonction BoxMuller.m simulant une loi gaussienne de moyenne et variance à préciser en paramètres.
2. Générer les données $(X_i, Y_i)_{i=1\dots n}$ avec $n = 50$, $a = 1.5$, $b = -1$ et $\sigma = 0.5$ pour des X_i uniformément distribués sur $[0; 1]$ et les Y_i construits via l'algorithme de Box-Muller.
3. Représenter le nuage de points grâce à l'option 's' de plot et en ayant préalablement classé par ordre croissant les X_i.
4. Identifier pour votre modèle, en fonction de vos données, $\hat{a}, \hat{b}$ et $\hat{\sigma^2}$.
5. Donner un intervalle de prédiction pour la loi conditionnelle $\mathcal{L}(Y|X = x)$, de niveau α quelconque, en fonction du quantile symétrique h_α d'ordre α pour une loi de Student (à combien de degrés de liberté ?).
6. Donner un intervalle de confiance de niveau α quelconque, qui contiendra $a + bx$, en fonction des mêmes quantités précédemment évoquées.
7. Tracer sur le même graphique les données, la fonction à estimer, l'ensemble de prédiction et la droite de régression obtenus.
8. Écrire alors un programme.m automatisant ces procédures, prenant en argument a, b, σ, n et lancer le programme plusieurs fois pour vérifier que la région de confiance couvre la régression.

2 Test d'hypothèse linéaire gaussienne

Soit X le vecteur des observations, on dit que X satisfait un modèle linéaire gaussien ssi

$$X = A\theta + \epsilon$$

où
- θ est un paramètre du modèle de taille p (à estimer)
- A est une matrice $n \times p$ connue
- ϵ est un vecteur gaussien centré de variance $\sigma^2 I_n$.

Une autre écriture possible est

$$X = m + \epsilon$$

où $m \in V = ImA$ et $\epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$.

Question 1 Montrer que ces deux écritures précédentes sont parfaitement équivalentes lorsque A est injective.

2.1 Régression linéaire

Question 2 Préciser le modèle obtenu dans le cadre de la régression linéaire. Que vaut A ? Dans quel espace vit la moyenne m de X ?

2.2 Analyse de variance à un facteur

Ce modèle est utilisé pour étudier une variable d'intérêt en fonction d'une autre variable (ou facteur). La donnée de base est la donnée de I échantillons indépendants (un échantillon par valeurs possibles du facteur) : $(X_{ij})_{i,j}$ où $i \in \{1 \dots I\}, j \in \{1 \dots n_i\}$. On suppose de plus que chacun des I échantillons est gaussien et que tous les échantillons ont la même variance.

Question 3 Écrire le modèle précédent sous la forme de modèle linéaire gaussien. Que vaut A ? Dans quel espace vit la moyenne m ?

2.3 Analyse de variance à deux facteurs

A noter que le modèle précédent peut être étendu à deux facteurs d'influence.

Question 4 Écrire le modèle à deux facteurs sous la forme de modèle linéaire gaussien. Que vaut A ? Dans quel espace vit la moyenne m ?

2.4 Estimateurs

Question 5 Montrer l'équivalence entre l'estimateur des moindres carrés de (θ, σ^2) et l'estimateur du maximum de vraisemblance dans le modèle linéaire gaussien.

Question 6 En terme de projection, préciser les estimateurs obtenus pour $\hat{m}$ et $\hat{\sigma^2}$.

Question 7 Donner les formules analytiques obtenues pour $\hat{\theta}$ et $\hat{\sigma^2}$ en fonction de A, X et n.

Question 8 Montrer que $\hat{m}$ et $\hat{\sigma^2}$ sont indépendantes et que

$$\hat{m} \sim \mathcal{N}(m, \sigma^2 \Pi_V) \qquad \text{et} \qquad n\hat{\sigma^2}/\sigma^2 \sim \chi^2_{n-\dim V}$$

3 Tests statistiques et simulations

3.1 Tests et intervalles de confiance

L'objectif des tests sur le modèle linéaire gaussien est de tester l'hypothèse : $\mathcal{H}_0 =$ "$m \in W$" contre $\mathcal{H}_1 =$ "$m \notin W$" où W est un s.e.v. de V. Ce test est destiné à diminuer le nombre de paramètres dans le modèle linéaire gaussien.

Question 9 Dans le cas du modèle de régression, on souhaite tester si $\theta_1 = 0$, ce qui signifie que la première des variables explicatives est superflue dans le modèle. Que vaut W ?

Question 10 Dans le cas d'analyse de variance à un facteur, on veut tester par exemple $\mathcal{H}_0$: "pour chaque facteur, la moyenne est identique" contre l'hypothèse opposée $\mathcal{H}_1$. Que vaut W ?

Question 11 On décide de former un test d'acceptation ou de rejet de $\mathcal{H}_0$. Quel test vous semble-t-il naturel d'utiliser ? Donner, sous $\mathcal{H}_0$, la statistique de

$$F := \frac{\frac{1}{\dim V - \dim W}\|\Pi_V(X) - \Pi_W(X)\|^2}{\frac{1}{n - \dim V}\|X - \Pi_V(X)\|^2}$$

En déduire une région de rejet de $\mathcal{H}_0$ de niveau α en fonction de n, V et W.

Question 12

1. Écrire une fonction $[p, f] = homvar(y)$ qui prend comme argument une matrice $n \times nech$ contenant $nech$ échantillons de taille n et qui renvoie
   - dans p une matrice $p(i, j)$, i et j variant de 1 à $nech$ contenant la P-valeur du test d'égalité des variances pour les échantillons i et j.
   - dans f une matrice $f(i, j)$ contenant la statistique de Fisher du test d'égalité des variances.
2. Écrire une fonction $[p, t, tsub, f] = testlin(y, A, Asub)$ qui prend comme arguments un vecteur y contenant l'échantillon de données, A la matrice $n \times p$ du modèle linéaire, $Asub$ la matrice $n \times q$ du sous-modèle à tester et qui renvoie :
   - p la P-valeur du test d'hypothèse linéaire
   - f la statistique de Fisher du test d'hypothèse linéaire
   - t l'estimation des paramètres du modèle
   - tsub l'estimation des paramètres du sous-modèle.

Question 13 Exemple issu du Bickel & Doksum On donne les limites d'élasticité des fils électriques de neufs cables utilisés sur un réseau haute tension. Chaque cable est fait de 12 fils.

Cable 1	5	-13	-5	-2	-10	-6	-5	0	-3	2	-7	-5
Cable 2	-11	-13	-8	8	-3	-12	-12	-10	5	-6	-12	-10
Cable 3	0	-10	-15	-12	-2	-8	-5	0	-4	-1	-5	-11
Cable 4	-12	4	2	10	-5	-8	-12	0	-5	-3	-3	0
Cable 5	7	1	5	0	10	6	5	2	0	-1	-10	-2
Cable 6	1	0	-5	-4	-1	0	2	5	1	-2	6	7
Cable 7	-1	0	2	1	-4	2	7	5	1	0	-4	2
Cable 8	-1	0	7	5	10	8	1	2	-3	6	0	5
Cable 9	2	6	7	8	15	11	-7	7	10	7	8	1

Quel test faire pour s'assurer que ces données proviennent d'un modèle linéaire ? Construire un test pour savoir si les limites d'élasticité moyenne.