TP X - Modèle linéaire gaussien (III)

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 $\epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$.

1. Montrer que ces deux écritures précédentes sont parfaitement équivalentes lorsque A est injective.
2. Préciser le modèle obtenu dans le cadre de la régression linéaire. Que vaut A ? Dans quel espace vit la moyenne m de X ?
3. Modèle d'analyse de la variance : 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.

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

   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 ?
5. 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 ?
6. 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.
7. 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.
8. 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.

9. 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

   (a) Quel test faire pour s'assurer que ces données proviennent d'un modèle linéaire ?

   (b) Quels sont les cables dont les variances sont identiques ?(au seuil de 5%)

   (c) Sur de tels cables, quels sont les cables dont les moyennes d'élasticité sont identiques ? (au seuil de 5%)

10. Le revenu net par action de la compagnie Gillette pour les années 57 à 64 est

Année (z_j)	1957	1958	1959	1960	1961	1962	1963	1964
Revenu (y_j)	0.93	0.99	1.11	1.33	1.52	1.60	1.47	1.33

   On pose $x_j = z_j - 1960.5$, puis $\psi_0(x) = 1$, $\psi_1(x) = 2x$ et enfin $\psi_2(x) = x^2 - 21/4$.

   (a) On suppose que $y_j = \lambda_0\psi_0(x_j) + \lambda_1\psi_1(x_j) + \lambda_2\psi_2(x_j) + \epsilon_j$ avec ϵ_j indépendants de même loi $\mathcal{N}(0, \sigma^2)$. Estimer $\lambda_0, \lambda_1, \lambda_0$ et σ^2.

   (b) Tester l'hypothèse $\lambda_2 = 0$ au seuil de 5%.

   (c) Faire une prévision pour le revenu net par action en 1965.