UPS TOULOUSE III **Le 26/01/2005**
MAGISTÈRE

EXAMEN
MODÉLISATION STATISTIQUE

Durée 3 heures

PROBLÈME I

5 points

Un particulier cherche à acquérir un appartement aux alentours immédiats de la place du Capitole. Il a sélectionné les 24 offres de vente suivantes où X représente la surface en mètres carrés et Y correspond au prix en milliers d'Euros.

x_i	28	50	196	55	190	110	60	48	90	35	86	65
y_i	130	280	800	268	790	500	320	250	378	250	350	300

x_i	32	52	40	70	28	30	105	52	80	60	20	100
y_i	155	245	200	325	85	78	375	200	270	295	85	495

On suppose que chaque y_i est une réalisation d'une variable aléatoire Y_i satisfaisant

$$Y_i = a + bx_i + \varepsilon_i, \qquad i = 1, 2, \cdots, n$$

où (ε_i) est une suite de variables aléatoires indépendantes et de même loi $\mathcal{N}(0, \sigma^2)$.

1) Estimer les paramètres inconnus a, b et σ^2.

2) Proposer un intervalle de confiance à 95% pour σ^2.

3) Tester la significativité de la surface sur le prix des appartements.

4) Pour une nouvelle valeur $x^* = 100$, calculer la prédiction naturelle $\widehat{Y}^*$ de Y^* et trouver un intervalle de prévision à 95% pour Y^*.

PROBLÈME II

3 points

On considère le modèle de régression polynômiale défini pour $n \geq 3$ par

$$Y_i = a + bx_i + cx_i^2 + \varepsilon_i, \qquad i = 1, 2, \cdots, n$$

où (ε_i) est une suite de variables aléatoires indépendantes, centrées et de variance $\sigma^2 z_i^2$ et (z_i) est une suite connue, strictement positive.

1) Proposer une condition simple pour que cette régression soit identifiable.
2) Trouver l'expression des estimateurs des moindres carrés pondérés de a, b, c et σ^2 sous l'hypothèse

$$\sum_{i=1}^{n} \frac{x_i}{z_i^2} = 0 \qquad \text{et} \qquad \sum_{i=1}^{n} \frac{x_i^3}{z_i^2} = 0.$$

PROBLÈME III

7 points

On souhaite étudier l'efficacité de trois types d'aliment A_1, A_2, A_3 et de trois conditions d'élevage B_1, B_2, B_3 sur la gavage des canards. Pour chaque type d'aliment et chaque condition d'élevage, on dispose d'un échantillon de quatre canards et on mesure le poids en grammes des foies gras obtenus après gavage.

Poids des foies	B_1	B_2	B_3
A_1	620 612 615 627	670 631 628 614	704 679 687 707
A_2	586 618 604 614	662 678 597 619	711 678 682 649
A_3	591 622 579 574	623 647 580 607	671 623 597 615

On désigne par y_{ijk} le poids du foie du k^e canard ayant reçu l'aliment A_i sous la condition d'élevage B_j. On suppose que y_{ijk} est une réalisation d'une variable aléatoire Y_{ijk} satisfaisant pour $i = 1, 2, 3$ et $j = 1, 2, 3$

$$Y_{ijk} = \mu + a_i + b_j + c_{ij} + \varepsilon_{ijk} \qquad k = 1, 2, 3, 4$$

où (ε_{ijk}) est une suite de variables aléatoires indépendantes et de même loi $\mathcal{N}(0, \sigma^2)$. On suppose également que les contraintes standards sur la décomposition de la moyenne sont réalisées.

1) Estimer les paramètres inconnus μ, a, b, c et σ^2.

2) Dresser le tableau d'analyse de la variance à deux facteurs sachant que

$$\begin{aligned} S_R &= \sum_{i=1}^{3}\sum_{j=1}^{3}\sum_{k=1}^{4}(Y_{ijk} - Y_{ij*})^2 = 15\,918.25, \\ S_T &= \sum_{i=1}^{3}\sum_{j=1}^{3}\sum_{k=1}^{4}(Y_{ijk} - Y_{***})^2 = 52\,234.75. \end{aligned}$$

3) Tester l'hypothèse selon laquelle il n'y a pas d'interaction.

4) Tester l'hypothèse d'absence d'effet de chacun des 2 facteurs.

PROBLÈME IV

5 points

On considère le modèle de régression linéaire simple défini, pour $n \geq 2$, par

$$Y_i = a + bx_i + \varepsilon_i, \qquad i = 1, 2, \cdots, n$$

où (x_i) est une suite déterministe de nombres réels non tous égaux et (ε_i) est une suite de variables aléatoires indépendantes, de même loi, centrées et de variance $\sigma^2 > 0$. Le modèle dépend du nombre n d'observations.

1) Rappeler la condition du cours sous laquelle $\widehat{Y}_i$ converge en probabilité, quand n tend vers l'infini, vers $a + bx_i$ avec $i = 1, 2 \cdots, n$.

2) Montrer que, pour étudier cette convergence, on peut remplacer le régresseur x_i par $z_i^n = x_i - \bar{x}_n$ avec
$$\overline{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i.$$
On pose alors, pour $n \geq 2$, $Y_i = m + bz_i^n + \varepsilon_i$ avec m à préciser.

3) Ecrire la matrice X du modèle linéaire en utilisant la paramétrisation ci-dessus.

4 Montrer que, pour $i = 1, 2 \cdots, n$,

$$\widehat{Y}_i = \widehat{m} + \widehat{b}\, z_i^n = \overline{Y}_n + \left(\frac{\sum_{j=1}^n z_j^n Y_j}{\sum_{j=1}^n (z_j^n)^2} \right) z_i^n.$$

5) Finalement, montrer que $\widehat{Y}_i$ converge en probabilité, quand n tend vers l'infini, vers $a + bx_i$ avec $i = 1, 2 \cdots, n$ si

$$\lim_{n \to \infty} \frac{(z_i^n)^2}{\sum_{j=1}^n (z_j^n)^2} = 0.$$