Examen terminal

Durée : 2 heures. Epreuve sans calculatrice. Toutes les réponses doivent être justifiées.

Exercice 1 (4 points)**.** On considère un modèle M de régression linéaire simple :

$$y_i = \theta_0 + \theta_1 x_i + e_i, \quad i = 1, \dots, n,$$

ainsi que le sous modèle M_0 : $\quad y_i = \beta_0 + \varepsilon_i, \quad i = 1, \dots, n.$

1. Exprimer $\widehat{\beta}_0$ et $\widehat{\theta}_0$ en fonction des observations x et y.
2. Donner une condition nécessaire et suffisante pour que $\widehat{\beta}_0 = \widehat{\theta}_0$.
3. Proposer une condition, portant seulement sur le prédicteur x, qui assure l'égalité $\widehat{\beta}_0 = \widehat{\theta}_0$, quelle que soit la valeur de y.
4. Généraliser le résultat de la question précédente au cas où le modèle M est une régression linéaire multiple faisant intervenir des prédicteurs $x^1, x^2, \dots, x^{p-1}$.

Exercice 2 (5 points)**.** On considère un modèle linéaire gaussien de la forme

$$Y_i = \sum_{j=1}^{p} x_i^j \theta_j + \sigma G_i, \quad i = 1, \dots, n$$

où les variables aléatoires G_i sont indépendantes et de loi gaussienne centrée réduite. Ce système peut s'écrire $Y = X\theta + \sigma G$ avec $\theta \in \mathbb{R}^p$, $X \in M_{n,p}(\mathbb{R})$ et Y, G vecteurs aléatoires à valeurs dans $\mathbb{R}^n$. On suppose que X est de rang p. L'estimateur des moindres carrés est donné par la formule $\widehat{\theta} = (X'X)^{-1}X'Y$.

1. En utilisant seulement les résultats de base sur les vecteurs gaussiens (que l'on rappellera), déterminer la loi de chacune des variables aléatoires suivantes

$$Y, \ \widehat{Y} - X\theta, \ \frac{|\widehat{Y} - X\theta|^2}{\sigma^2}.$$

2. En déduire une région de confiance pour $X\theta$, avec un niveau de sûreté de 95%.

Exercice 3 (5 points)**.** La fédération des offices du tourisme a effectué une enquête sur la répartition des hôtels. Cette étude a porté sur 120 offices du tourime. Chaque observation correspond à une zone géographique définie autour de chaque office de tourisme participant. Pour chaque zone, les variables quantitatives suivantes ont été relevées :

- le nombre d'hôtels, noté y ;
- l'indice d'attractivité touristique, noté x^1, correspondant à une note entre 0 et 10 ;
- le taux d'occupation annuel moyen, noté x^2 ;
- le nombre moyen d'étoiles pour les hôtels de la zone, noté x^3.

On cherche à modéliser le nombre d'hôtels en fonction des autres variables disponibles. Trois modèles de régressions linéaires ont été estimés :

- M1 : y en fonction de x^1 de la forme $y_i = \beta_0 + \beta_1 x_i^1 + e_i$,
- M2 : y en fonction de x^1, x^3 de la forme $y_i = \beta_0 + \beta_1 x_i^1 + \beta_3 x_i^3 + e_i$,
- M3 : y en fonction de x^1, x^2, x^3 de la forme $y_i = \beta_0 + \beta_1 x_i^1 + \beta_2 x_i^2 + \beta_3 x_i^3 + e_i$.

On vous donne ci-dessous les sommes des carrés des résidus obtenus pour chaque modèle :

Modèle	SSR
$M1 : y \sim x^1$	279507
$M2 : y \sim x^1, x^3$	227364
$M3 : y \sim x^1, x^2, x^3$	224467

En comparant ces 3 modèles à l'aide de tests, montrer que le modèle $M2$ est la meilleure régression pour modéliser le nombre d'hôtels. On vous demande d'expliquer précisément les tests utilisés (hypothèses, statistiques et leurs lois, valeurs lues dans les tables). Indication : les calculs numériques exacts ne sont pas indispensables pour conclure.

Exercice 4 (6 points)**.** On injecte à 24 lapins de l'insuline en leur donnant des doses notées D_1, D_2 et D_3, préparées suivant deux protocoles différents notés P_1 et P_2. La réduction de sucre dans leur sang a été mesurée et elle a donné les résultats suivants :

Réduction	P_1	P_2	tous P_j
D_1	17 21 49 54 (moyenne : $35,25$) (variance : $269,188$)	33 37 40 16 (variance : $86,25$)	moyenne : $33,375$
D_2	64 48 34 63 (moyenne : $52,25$) (variance : $151,188$)	41 64 34 64 (moyenne : $50,75$) (variance : $181,688$)	moyenne : $51,5$
D_3	62 72 61 91 (moyenne : $71,5$) (variance : $145,25$)	56 62 57 72 (moyenne : $61,75$) (variance : $40,188$)	moyenne : $66,625$
toutes D_i	moyenne : 53	moyenne : 48	moyenne : $50,5$

On note $y_{i,j,k}$ la réduction de sucre dans le sang du k^e lapin ayant reçu une dose D_i d'insuline préparée selon le protocole P_j, avec $i = 1, 2, 3$ et $j = 1, 2$. On considère le modèle à deux facteurs

$$y_{i,j,k} = m_{i,j} + e_{i,j,k} \text{ pour } i = 1, 2, 3,\ j = 1, 2 \text{ et } k = 1, 2, 3, 4$$

où les $e_{i,j,k}$ sont les réalisations indépendantes d'une v.a. distribuée selon une loi $N(0, \sigma^2)$.

1. Estimer les paramètres inconnus $m_{i,j}$.
2. Tracer le diagramme d'interactions. Commenter.
3. Expliquer en détail comment calculer la somme des carrés des résidus (SSR) à partir des données fournies. On ne demande pas de faire l'application numérique finale, qui donne SSR= 3495.
4. Dans la suite on s'intéresse à la paramétrisation centrée

$$y_{i,j,k} = \mu + a_i + b_j + c_{i,j} + e_{i,j,k} \text{ pour } i = 1, 2, 3,\ j = 1, 2 \text{ et } k = 1, 2, 3, 4.$$

   (a) Expliciter les relations qui existent entre les coefficients $c_{i,j}$ dans ce modèle.

   (b) Expliquer en détail comment calculer les $\widehat{c}_{i,j}$ et la somme des carrés liés à l'intéraction (SSI) à partir des données fournies. On ne demande pas de faire le calcul numérique final. On admettra que SSI= $75,25$.

   (c) Tester l'hypothèse d'absence d'interactions entre les deux facteurs.