Examen terminal

Durée 2 heures. Epreuve sans calculatrice. Toutes les réponses doivent être justifiées. Pensez à numéroter les copies si vous en rendez plusieurs.

Exercice 1 (7 points)**.** On considère le modèle linéaire gaussien $Y = X\theta + \sigma G$ avec $\theta \in \mathbb{R}^p$, X de taille $n \times p$ et de rang p et G de loi $\mathcal{N}_n(0, I_n)$. Soit $v \in \mathbb{R}^p$ un vecteur non nul. On rappelle que les estimateurs des paramètres inconnus sont

$$\widehat{\theta} := (X'X)^{-1}X'Y \quad \text{et} \quad \widehat{\sigma}^2 := \frac{\left|P_{\mathrm{Im}(X)^\perp} Y\right|^2}{n-p}\,.$$

1. Exprimer $\widehat{\theta}$ et $\widehat{\sigma}^2$ en fonction de G.
2. En déduire la loi de $\widehat{\theta}$, la loi de $\dfrac{1}{\sigma}\langle\widehat{\theta} - \theta, v\rangle$ et celle de $(n-p)\dfrac{\widehat{\sigma}^2}{\sigma^2}$. On précisera les propriétés des vecteurs gaussiens qui sont utilisées.
3. Déterminer un nombre $z > 0$ tel que $\dfrac{1}{z\,\widehat{\sigma}}\langle\widehat{\theta} - \theta, v\rangle$ suive la loi de Student $s(n-p)$.
4. En déduire un intervalle de confiance de niveau de sûreté $1-\alpha$ pour $\langle\theta, v\rangle$.
5. Construire un test de niveau α pour H_0 : "$\langle\theta, v\rangle = c$" contre H_1 : "$\langle\theta, v\rangle \neq c$".
6. Expliciter le test obtenu pour l'hypothèse nulle "$\theta_1 = \theta_2$".

Exercice 2 (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, déterminer celui qui modélise le mieux 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 3 (8 points)**.** Un psychologue a voulu étudier la vitesse de lecture d'enfants entre 6 et 8 ans. Pour cela, il a mesuré le temps écoulé (en centièmes de seconde) entre le moment où le mot est présenté à l'enfant et le moment où il est correctement épelé. On cherche maintenant à étudier cette variable selon deux facteurs :

- le sexe (1 = garçon ; 2 = fille)
- l'âge (1 = fin de C.P. ; 2 = C.E.1)

Pour chacune des 4 configurations définies par ces 2 facteurs, 6 observations de la variable réponse ont été réalisées. Sur les 24 observations, on a estimé un temps moyen de lecture de 1184 cs, avec une variance de 80727.
On désigne par y_{ijk} le temps réalisé par le k^e enfant de sexe i et d'âge j. On suppose que pour $i \in \{1,2\}$, $j \in \{1,2\}$ et $k \in \{1,\ldots,6\}$,

$$y_{ijk} = m_{ij} + e_{ijk} = \mu + a_i + b_j + c_{ij} + e_{ijk}$$

où e_{ijk} est la réalisation d'une v.a. $E_{ijk} \sim \mathcal{N}(0,\sigma^2)$ et les E_{ijk} sont indépendantes.
La mise en œuvre de ce modèle a donné les estimations suivantes :

$$\widehat{\mu}(y) = 1184 \qquad \widehat{a}_1(y) = 52 \qquad \widehat{b}_1(y) = 220 \qquad \widehat{c}_{11}(y) = 75$$

1. Quelles sont les caractéristiques du plan d'expérience ?
2. Préciser les contraintes associées à cette paramétrisation. En déduire les estimations des autres paramètres intervenant dans cette paramétrisation.
3. A partir des données ci-dessus, trouver les estimations des paramètres m_{ij}.
4. Représenter le diagramme d'interactions. Commenter.
5. Calculer les sommes des carrés expliquées par l'interaction (SSI).
6. Expliquer en détail comment calculer, à partir des données disponibles, les quantités $SS1$, $SS2$ et SSR. On ne demande pas le calcul final de l'application numérique, qui donne $SS1 = 64896$, $SS2 = 1161600$ et $SSR = 575952$.
7. Testez l'absence d'effet principal de chacun des facteurs (les ordres de grandeur suffisent pour conclure).
8. Tester l'hypothèse d'absence d'intéraction entre les deux facteurs.
9. Proposez un sous-modèle pertinent et calculer approximativement son coefficient de détermination.