M1 IMAT - Modèles linéaires — Université Paul Sabatier
Année 2010 – 2011

Examen final

8 décembre 2010
Durée 2H

Epreuve sans calculatrice. Les seuls documents autorisés sont la table des quantiles des lois de Student et Fisher et une feuille A4 manuscrite qui sera rendue avec la copie. Toutes les réponses doivent être justifiées. Pensez à numéroter les copies si vous en rendez plusieurs.

Exercice 1 (4 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 $\widehat{\theta} = (X'X)^{-1}X'Y$. En utilisant seulement les résultats de base sur les vecteurs gaussiens (que l'on rappellera), déterminer la loi de G, $-G$, Y et $\widehat{\theta}$.

Exercice 2 (12 points)**.** Les données analysées dans ce problème portent sur 110 avions de chasse américains construits entre 1940 et 1970, pour lesquels on cherche à étudier l'évolution de leurs caractéristiques au cours du temps. Les variables disponibles sont les suivantes :

- y : date du premier vol (en nombre de mois après janvier 1940),
- x^1 : puissance,
- x^2 : facteur de gamme de vol,
- x^3 : charge utile,
- x^4 : facteur de charge.

Le but de cette étude est donc de modéliser la date du premier vol en fonction des quatre autres variables quantitatives.

Dans un premier temps, on considère la régression de y en fontion de x_1 et x_2 :

$$y_i = \beta_0 + \beta_1 x_i^1 + \beta_2 x_i^2 + e_i$$

où e_i sont les réalisations indépendantes d'une v.a. de loi $N(0, \sigma^2)$.

A partir des données, on a calculé les quantités suivantes (on a posé $n = 110$) :

$$\sum_{i=1}^{n} x_i^1 = 433.89 \qquad \sum_{i=1}^{n} x_i^2 = 503.5 \qquad \sum_{i=1}^{n} y_i = 18290$$

$$\sum_{i=1}^{n} (x_i^1)^2 = 2299.85 \qquad \sum_{i=1}^{n} (x_i^2)^2 = 2364.19 \qquad \sum_{i=1}^{n} (y_i)^2 = 3381570$$

$$\sum_{i=1}^{n} x_i^1 x_i^2 = 2069.04 \qquad \sum_{i=1}^{n} x_i^1 y_i = 82408.86 \qquad \sum_{i=1}^{n} x_i^2 y_i = 86815.8$$

$$\sum_{i=1}^{n} (y_i)^2 - \frac{1}{n}\Big(\sum_{i=1}^{n} y_i\Big)^2 = 340442$$

1. Spécifier y, e, X et β pour écrire le modèle sous la forme $y = X\beta + e$.
   Calculer $X'X$ et $X'y$. Expliquer comment estimer les paramètres de la régression à partir des données fournies dans l'énoncé. On ne demande pas de faire l'application numérique finale, qui donne $\widehat{\beta_0} \approx -40,41$, $\widehat{\beta_1} \approx 12,15$ et $\widehat{\beta_2} \approx 34,57$.

   Note : on vous donne $(X'X)^{-1} = \begin{pmatrix} 0.37359 & 0.00516 & -0.08407 \\ 0.00516 & 0.00211 & -0.00295 \\ -0.08407 & -0.00295 & 0.02091 \end{pmatrix}$

2. Rappeller la définition de la somme des carrés résiduels (SSR) et expliquer en détail comment la calculer à partir des données numériques disponibles. On admettra que $SSR \approx 118000$. En déduire l'estimation de σ^2.

3. Tester l'hypothèse selon laquelle la date du premier vol ne dépend ni de la puissance ni du facteur de gamme de vol (préciser les hypothèses testées, la statistique utilisée et sa loi).

   Indication : on remarquera que les ordres de grandeur suffisent pour conclure.

4. Tester l'hypothèse d'absence d'effet de la puissance sur la date du premier vol.

5. Dans un deuxième temps, on a réalisé la régression linéaire de y en fonction des 4 variables explicatives disponibles : x^1, x^2 (déjà introduites dans le premier modèle), x^3 et x^4. La procédure `REG` de `SAS` a donné les résultats suivants

   ```
        Analysis of Variance
                                  Sum of
   Source                         Squares
   Model                          241395
   Error                           99046
   Corrected Total                340442
   ```

   Au vu des résultats obtenus pour les deux modèles, construire le test permettant de dire quel est le meilleur modèle.

Exercice 3 (6 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, six 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 y_{ijk} est une réalisation d'une v.a. Y_{ijk} satisfaisant, pour $i = 1, 2$ et $j = 1, 2$, le modèle suivant :

$$y_{ijk} = \mu_{ij} + e_{ijk} = \mu + \alpha_i^L + \alpha_j^C + \beta_{ij} + e_{ijk} \qquad k = 1, ..., 6$$

dans lequel e_{ijk} est la réalisation d'une v.a. $E_{ijk} \sim N(0, \sigma^2)$ et les E_{ijk} sont indépendantes.

La mise en œuvre de ce modèle a donné les estimations suivantes :

$$\widehat{\mu}(y) = 1184 \qquad \widehat{\alpha}_1^L(y) = 52 \qquad \widehat{\alpha}_1^C(y) = 220 \qquad \widehat{\beta}_{11}(y) = 75$$

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 estimations données ci-dessus, retrouver les estimations des paramètres μ_{ij}.
4. Représenter le diagramme d'interactions. Commenter.
5. Expliquer comment calculer les sommes des carrés expliquées par l'interaction (SSI) et par chaque facteur ($SS1$ et $SS2$). Comparer leurs ordres de grandeur et en tirer les conclusions.