Modéle de régression linéaire - Feuille 7

Analyse de variance

Exercice 1

	Nombre	Moyenne	écart-type
Mâle	94	31.70	2.62
Femelle	83	25.23	2.00

Table 1 – Statistiques résumées pour le poids (g) des souris issue de la famille 141G6 (transgéniques ou non).

Le tableau 1 contient les moyennes, écart-types pour le poids des souris mâles et femelles issues de la famille 141G6. Utilisez ces statistiques résumées pour construire un tableau ANOVA. Conclusions ?

Exercice 2 On dispose de k n-échantillons $(Y_{i1}, ..., Y_{in})$, $i = 1, ..., k$, les n-échantillons étant indépendants les uns des autres. Pour l'échantillon $i = 1, ..., k$, les variables $Y_{i1}, ..., Y_{in} \sim \mathcal{N}(m_i, \sigma^2)$. On veut tester l'homogénéité des moyennes :

$$H_0 : m_1 = ... = m_k \qquad \text{contre} \qquad H_1 : \exists i \neq j, \quad m_i \neq m_j.$$

On utilise les notations suivantes :

- pour la moyenne empirique : $\bar{Y} = \frac{1}{nk} \sum_{i=1}^{k} \sum_{j=1}^{n} Y_{ij}$,
- dans l'échantillon $i = 1, ..., k$, la moyenne empirique des variables est notée $\bar{Y}_{i.} = \frac{1}{n} \sum_{j=1}^{n} Y_{ij}$,
- pour la variabilité totale de l'échantillon : $(nk-1)S^2 = \sum_{i=1}^{k} \sum_{j=1}^{n} (Y_{ij} - \bar{Y})^2$.
- la variabilité intra-groupe est $\sum_{i=1}^{k} \sum_{j=1}^{n} (Y_{ij} - \bar{Y}_{i.})^2$.
- la variabilité inter-groupe est $n \sum_{i=1}^{k} (\bar{Y}_{i.} - \bar{Y})^2$.

1. Montrer que la variabilité de l'échantillon s'écrit comme la somme des variabilités intra et inter-groupes.
2. Considérons les vecteurs
$$\begin{aligned} Y &= (Y_{11}, ..., Y_{1n}, ..., Y_{k1}, ..., Y_{kn}), \\ \Upsilon &= (\bar{Y}_{1.}, ..., \bar{Y}_{1.}, ..., \bar{Y}_{k.}, ..., \bar{Y}_{k.}). \end{aligned}$$
Montrer que Υ est la projection orthogonale de Y sur le sous espace vectoriel E (de $\mathbb{N}^{nk}$) de dimension k engendré par les vecteurs indicateurs de bloc
$$v_1 = (1, ..., 1, 0, ..., 0, 0, ..., 0)^T, \quad \ldots, \quad v_k = (0, ..., 0, 0, ..., 0, 1, ..., 1)^T.$$
3. Soit la statistique $Z = \dfrac{n \sum_{i=1}^{k} (\bar{Y}_{i.} - \bar{Y})^2}{\sum_{i=1}^{k} \sum_{j=1}^{n} (Y_{ij} - \bar{Y}_{i.})^2}$. Démontrer que sous l'hypothèse H_0, la v.a. $\dfrac{nk-k}{k-1} Z$ suit une loi de Fisher $\mathcal{F}(k-1, nk-k)$.
Indication : on pourra utiliser le théorème des 3 perpendiculaires.

4. Application numérique : On a relevé les scores des étudiants de 4 écoles á un concours. Comparer les performances des écoles. Les différences observées sont-elles significatives au risque 5 % ? Comparer deux à deux les échantillons.

E_1	E_2	E_3	E_4
73	84	69	65
57	95	80	58
95	96	73	82
78	62	62	86
86	80	50	35
61	87	71	52
80	100	84	70
98	74	66	79
64	85	52	43
78	77	73	60

$\bar{Y}_{1.} = 77, \bar{Y}_{2.} = 84, \bar{Y}_{3.} = 68, \bar{Y}_{4.} = 63.$

Exercice 3 Considérons le modèle

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, 1 \leq i \leq I, 1 \leq j \leq n_i,$$

où les ε_{ij} sont indépendantes de loi $\mathcal{N}(0, \sigma^2)$, $\sum_{i=1}^{I} n_i \alpha_i = 0$ et $\sum_{i=1}^{I} n_i = n$.

1. Montrer que $Y = A\theta + \varepsilon$, où $A = [\mathbb{1}_n a_1 ... a_I]$ avec $\mathbb{1}_n, a_1, ..., a_I$, éléments de $\mathbb{N}^{IJ}$ que l'on précisera.
2. Soit $F_0 = \{\mu \mathbb{1}_n, \mu \in \mathbb{N}\}, F_1 = \left\{ \sum_{i=1}^{I} \alpha_i a_i, \alpha_i \in \mathbb{N}, \sum_{i=1}^{I} n_i \alpha_i = 0 \right\}$.
   Montrer que F_0 et F_1 sont des sous-espaces vectoriels deux-à-deux orthogonaux. En déduire qu'il existe un espace G tel que
   $$\mathbb{N}^{IJ} = F_0 \oplus F_1 \oplus G.$$
3. Calculer $P_0, P_0 + P_1$ (P_0 étant la projection orthogonale sur F_0). En déduire P_1. Calculer P_G.
4. Construire la table d'analyse de la variance.
5. Tester l'hypothèse H_0 : "tous les α_i sont nuls" contre H_1 : "tous les α_i ne sont pas nuls".
6. **Application.** Les mesures de teneur en octane sur des échantillons de carburant prélevés dans quatre régions du nord-est des États Unis durant l'été 1953, sont reproduites dans le tableau ci-après.
   Notant Y_{ij} la jème mesure effectuée dans la région i, on donne les quantités suivantes :

Région	A	B	C	D
n_i	16	13	18	22
$\bar{Y}_{i.}$	83.875	82.846	83.22	83.009
$\sum_j (Y_{ij} - \bar{Y}_{i.})^2$	17.81	2.67	28.97	33.04

   Peut-on conclure que la teneur en octane est différente suivant les régions ?

Région A	Région B	Région C	Région D
84.0	82.4	83.2	80.2
83.5	82.4	82.8	82.9
84.0	83.4	83.4	84.6
85.0	83.3	80.2	84.2
83.1	83.1	82.7	82.8
83.5	83.3	83.0	83.0
81.7	82.4	85.0	82.9
85.4	83.3	83.0	83.4
84.1	82.6	85.0	83.1
83.0	82.0	83.7	83.5
85.8	83.2	83.6	83.6
84.0	83.1	83.3	86.7
84.2	82.5	83.8	82.6
82.2		85.1	82.4
83.6		83.1	83.4
84.9		84.2	82.7
		80.6	82.9
		82.3	83.7
			81.5
			81.9
			81.7
			82.5

Data from O.C. Blade "National motor-gasoline survey" Bureau of Mines Report of Investigation 5041.

Exercice 4 Dans le cadre de l'analyse de variance à 2 facteurs, le modèle peut être réécrit sous la forme suivante

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk},$$

pour $1 \leq i \leq I$, $1 \leq j \leq J$, et $1 \leq k \leq n_{ij}$.

1. Quel est le nombre de paramètres à identifier ?
2. Le modèle est-il identifiable ?
3. Commenter les conditions d'identifiabilité, dites d'orthogonalité.

Exercice 5 Considérons le modèle

$$Y_{ij} = \mu + \alpha_i + \beta_j + cP_{ij} + \varepsilon_{ij},$$

pour $1 \leq i \leq I$, $1 \leq j \leq J$, où les P_{ij} sont connus, les ε_{ij} sont indépendants de loi $\mathcal{N}(0, \sigma^2)$ et $\sum_{i=1}^{I} \alpha_i = \sum_{j=1}^{J} \beta_j = 0$.

1. Montrer que $Y = \tilde{X}\theta + \varepsilon$, où $A = [\mathbb{1}|a_1| \cdots |a_I|b_1| \cdots |b_J|P]$ avec $\mathbb{1}, a_1, ..., a_I, b_1, ..., b_J, P$ éléments de $\mathbb{N}^{IJ}$ que l'on précisera.
2. Soit $F_0 = \{\mu 1, \mu \in \mathbb{N}\}, F_1 = \Big\{ \sum_{i=1}^{I} \alpha_i a_i, \alpha_i \in \mathbb{N}, \sum_{i=1}^{I} \alpha_i = 0 \Big\}$ et $F_2 = \Big\{ \sum_{j=1}^{J} \beta_j b_j,\ \beta_j \in \mathbb{N}, \sum_{j=1}^{J} \beta_j = 0 \Big\}$.
   Montrer que F_0, F_1 et F_2 sont des sous-espaces vectoriels deux-à-deux orthogonaux. En déduire que
   $$\mathbb{R}^{IJ} = F_0 \oplus F_1 \oplus F_2 \oplus G.$$

3. Calculer P_0, P_0+P_1. (P_0 étant la projection orthogonale sur F_0). En déduire P_1. Calculer P_G.
4. En déduire que l'on a la décomposition

$$\sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij}-\mu-\alpha_i-\beta_j-cP_{ij})^2 = IJ(\bar{Y}_{..}-\mu-c\bar{P}_{..})^2 + J\sum_{i=1}^{I}(\bar{Y}_{i.}-\bar{Y}_{..}-\alpha_i-c(\bar{P}_{i.}-\bar{P}_{..}))^2$$

$$+I\sum_{j=1}^{J}(\bar{Y}_{.j}-\bar{Y}_{..}-\beta_j-c(\bar{P}_{.j}-\bar{P}_{..}))^2$$

$$+\sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij}-\bar{Y}_{i.}-\bar{Y}_{.j}+\bar{Y}_{..}-c(P_{ij}-\bar{P}_{i.}-\bar{P}_{.j}+\bar{P}_{..}))^2.$$

5. En déduire que les ESBVM de $\mu, \alpha_1, ..., \alpha_I, \beta_1, ..., \beta_J$ vérifient $\hat{\mu} = \bar{Y}_{..} - \hat{c}\bar{P}_{..}$, $\hat{\alpha}_i = \bar{Y}_{i.} - \bar{Y}_{..} - \hat{c}(\bar{P}_{i.} - \bar{P}_{..})$ et $\hat{\beta}_j = \bar{Y}_{.j} - \bar{Y}_{..} - \hat{c}(\bar{P}_{.j} - \bar{P}_{..})$.
6. Déterminer $\hat{c}$ et $\hat{\sigma}^2$.
7. On suppose que $c = 0$, construire la table d'analyse de la variance dans ce cas.
8. On suppose que $c = 0$, déterminer un intervalle de confiance de β_j au seuil $1-\alpha$ de type Student pour j tel que $1 \leq j \leq J$.
9. On suppose que $c = 0$, tester au seuil $1-\alpha$ l'hypothèse que $\beta_j = 0$ pour $1 \leq j \leq J$.
10. Proposer un test pour l'hypothèse $c = 0$.
11. On suppose avoir rejeté l'hypothèse $c = 0$, tester alors l'hypothèse H_0 : "tous les β_j sont nuls" contre H_1 : "tous les β_j ne sont pas nuls".
12. Donner la table d'analyse de la variance dans le cas $c \neq 0$.