Corrigé examen 29 janvier 2020

Exercice 1

Y : Conc et trois facteurs : F1 : LAB à 6 niveaux
F2 : SPC à 7 niveaux
F3 : BAT à 3 niveaux

1) On a un modèle d'analyse de variance à 3 facteurs avec une interaction :

$$(M1) : Y_{ijkl} = \mu + (F1)_i + (F2)_j + (F3)_k + (F1*F3)_{ik} + \varepsilon_{ijkl}$$

$\boxed{n_{ijk} = 2}$ $\quad i = 1, \dots, 6 \quad j = 1, \dots, 7 \quad k = 1, 2, 3 \quad l = 1, \dots, n_{ijk}$

$F1_i$ c'est l'effet du niveau i de F1 sur Y

$F2_j$ . . . j de F2 sur Y

$F3_k$ . . . k F3 . . .

$(F1*F3)_{ik}$ . . . du niveau i de F1 et de k de F3 sur Y

Contraintes :
$$\sum_{i=1}^{6} F1_i = 0, \quad \sum_{j=1}^{7} F2_j = 0, \quad \sum_{k=1}^{3} F3_k = 0$$

$$\forall i = 1, \dots, 6, \quad \sum_{k=1}^{3} (F1*F3)_{ik} = 0$$

$$\forall k = 1, 2, 3 \quad \sum_{i=1}^{6} (F1*F3)_{ik} = 0$$

$\varepsilon_{ijkl} \sim N(0, \sigma^2)$ indép.

2) H_0 : (M1) non significatif $\Leftrightarrow$ aucun des 3 facteurs et interaction n'influent Y.

$$\begin{cases} F1_i = 0, \; F2_j = 0, \; F3_k = 0, \; (F1*F3)_{ik} = 0 \\ \forall i = 1, \dots, 6 \text{ et } \forall j = 1, \dots, 7 \text{ et } \forall k = 1, 2, 3 \end{cases}$$

Modèle réduit : $Y_{ijkl} = \mu + \varepsilon_{ijkl}$

(2)

H_1 : (M1) signif $\Leftrightarrow$ Y influencé par au moins un des facteurs ou par l'interaction entre F1 et F3

$\exists F1_i \neq 0$ ou $\exists F2_j \neq 0$ ou $\exists F3_k \neq 0$ ou $\exists (F1*F3)_{ik} \neq 0$

Modèle complet : (M1)

Statistique de test : $Z = \dfrac{SM/23}{SR/228} \underset{H_0}{\sim} F(23,228)$

Valeur de la stat. de test $z = 541.4$ avec la pvalue $< 10^{-16}$

$\Rightarrow H_0$ rejetée $\Rightarrow$ (M1) significatif

3) On teste F1*F3 :

H_0 : l'interaction F1*F3 n'influe pas Y si F1, F2, F3 sont dans le modèle

$\Leftrightarrow (F1*F3)_{ik} = 0 \quad \forall i = 1,..,6$ et $\forall k = 1,2,3$ | F1, F2, F3 dans le modèle

Modèle réduit : $Y_{ijk\ell} = \mu + (F1)_i + F2_j + F3_k + \varepsilon_{ijk\ell}$

H_1 : (F1*F3) influe Y si F1, F2, F3 sont dans le modèle :

$\Leftrightarrow \exists (F1*F3)_{ik} \neq 0$ si F1, F2, F3 sont dans le modèle

Modèle : (M1)

Stat de test : $Z = \dfrac{S_{F1*F3}/10}{SR/228} \underset{H_0}{\sim} F(10,228)$

Valeur stat de test : $z = 2{,}6226$ avec pvalue $= 0{,}004$

$\Rightarrow H_0$ rejetée $\Rightarrow$ F1*F3 influe Y si F1, F2, F3 sont dans le modèle

Facteurs significatifs F1 et F2, pendant que F3 n'est pas significatif. Donc il faut enlever le facteur BAT du modèle.

4) $\hat{F1}_1 = -0,32$ (3) $\hat{F1}_2 = 0,04$ . . . $\hat{F1}_5 = 0.01$, $\hat{F1}_6 = -\sum_{i=1}^{5} F1_i$

$\hat{F2}_1 = -1,41$, . . . $\hat{F2}_6 = -0,13$, $\hat{F2}_7 = -\sum_{j=1}^{6} F2_j$

$\hat{F3}_1 = -0.03$ $\hat{F3}_2 = 0.05$ $\hat{F3}_3 = -(0.05 - 0.03) = -0.02$

$(F1 * F3)_{11} = -0.02$ $(F1 * F3)_{12} = 0.02$, $(F1 * F3)_{13} = 0$

$\hat{\mu} = 1.92$

Par rapport à la moyenne la concentration de Y pour le laboratoire 1 est plus petite de 0.32 ($\hat{F1}_1 = -0.32$)

$\hat{F3}_2 = 0.05$ => par rapport à la moyenne, l'échantillon 2 a une concentration plus grande de 0.05

$(F1 \hat{*} F3)_{12} = 0,02$: par rapport à la moyenne, l'interaction du laboratoire 1 et du lot 3 augmente la valeur de Y de 0.02.

5) $R^2 adj = 0.98$ donc proche de 1 => (M1) de très bonne qualité d'ajustement.

Exercice 2

Y : class , var qualitative avec les valeurs 1 et 2

V1, V2, V3 sont des var numériques

1) On a un modèle de régression logistique, qui a la forme :

(M2) $\mathbb{P}[Y_i = 2 | V1_i, V2_i, V3_i] = \dfrac{\exp(\mu + b_1 V1_i + b_2 V2_i + b_3 V3_i)}{1 + \exp(\mu + \sum_{j=1}^{3} b_j Vj_i)}$

$i = 1, .., n$ $\boxed{n = 699}$

2) H0 : (M2) non signif $\Leftrightarrow b_1 = b_2 = b_3 = 0 \Leftrightarrow$ aucune des trois var n'influe la proba $\mathbb{P}[Y = 2]$.

Modèle : $\mathbb{P}[Y_i = 2] = \dfrac{\exp(\mu)}{1 + \exp(\mu)}$

H_1 : (M2) significatif $\Leftrightarrow \exists\, b_j \neq 0 \quad j = 1, 2, 3 \Leftrightarrow$ au moins une des trois var influe $\mathbb{P}[Y=2]$.

Modèle : (M2)

Tests utilisés : LR, SCORE, Wald : pvalue $< 10^{-4} \Rightarrow H_0$ rejetée.
Donc (M2) significatif.

3) On étudie la proba que le cancer soit malin : $\mathbb{P}[Y=2]$.

H_0 : V1 n'influe pas $\mathbb{P}[Y=2]$ si V2 et V3 sont dans le modèle

$\Leftrightarrow b_1 = 0 \Rightarrow$ Modèle réduit : $\mathbb{P}[Y_i = 2 \mid V2_i, V2_i) = \dfrac{\exp(\mu + b_2 V2_i + b_3}{1 + \exp(\mu + b_2 V2_i + b_3}$

H_1 : V1 influe $\mathbb{P}[Y=2]$ si V2 et V3 dans le modèle.

$\Leftrightarrow b_1 \neq 0 \Rightarrow$ Modèle (M2)

Stat de test :

$$Z = \left(\frac{\hat{b}_1}{\hat{\sigma}_{b_1}}\right)^2 \xrightarrow[H_0]{\mathcal{L}} \chi^2(1) \qquad \text{(avec SAS)}$$

Valeur stat de test : $z = 36.54 \Rightarrow$ p value $= 10^{-4}$
$\Rightarrow H_0$ rejetée

Remarque : avec R, la stat de test est :

$$\cancel{Z}_2 = \frac{\hat{b}_1}{\hat{\sigma}_{b_1}} \xrightarrow[H_0]{\mathcal{L}} \mathcal{N}(0,1)$$

$z_2 = 6.04$ p value $< 10^{-9} \Rightarrow H_0$ rejetée.

Pareil : V2 et V3 influent aussi $\mathbb{P}[Y=2]$ donc aussi $\mathbb{P}[Y=1]$ c'est à dire la proba que le cancer soit bénin. Pour rappel : $\mathbb{P}[Y=1] + \mathbb{P}[Y=2] = 1$.

4) $\hat{b}_1 = 0{,}59$ $\hat{b}_2 = 0.55$ $\hat{b}_3 = 0.71$ $\hat{\mu} = -7.59$

5) $\mathbb{P}[Y=2 \mid V1=3, V2=1, V3=1] = \dfrac{\exp(\hat{\mu} + \hat{b}_1 \cdot 3 + \hat{b}_2 \cdot 1 + \hat{b}_3 \cdot 1)}{1 + \exp(\hat{\mu} + \hat{b}_1 3 + \hat{b}_2 \cdot 1 + \hat{b}_3 \cdot 1)}$

6) $\mathbb{P}[Y=2]$ augment avec V3 car $\hat{b}_3 > 0$. Donc la première patiente a une proba plus grande pour que le cancer soit malin.

7) On utilise SAS :

Sur 458 patientes qui ont un cancer bénin, le modèle (M2) prévoit bien 446 cas.

Sur 241 patientes avec cancer malin, (M2) prévoit bien 220 cas.

Taux de mauvais prévision : $\frac{1}{2}\left(\frac{12}{458} + \frac{21}{241}\right) = 0.05$

Donc très bonne prévision par le modèle.

Exercice 3.

1) C'est un modèle de régression multiple : avec $Y = V1$

$$Y_i = \cancel{b_0} + \sum_{j=2}^{9} b_j V_{j_i} + \varepsilon_i \quad i = 1 \ldots, n \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \text{ indip.}$$

2) Param μ et $b_2, \ldots, b_9$

Pour (M3) les param $b_2, \ldots, b_9$ sont estimés par la méthode des moindres carrés pendant que par (M4) ils sont estimés par LASSO adaptatif.

3) Par (M3) aucune estimation de b_j n'est égale à 0.

Par (M4) $\hat{b}_2, \hat{b}_4, \hat{b}_5, \hat{b}_6, \hat{b}_8, \hat{b}_9$ ~~sont~~ sont 0 (ou proches)

Par (M3) et (M4) : $\hat{b}_7$ et $\hat{b}_3$ sont $\neq 0$.