Corrigé examen RGD, 14/02/2024

Exercice 1

$Y = \log(\text{temps})$
$X1 = \text{age}, \dots, X7 = SCALC$

1) Regression linéaire multiple :

(M1) $Y_i = b_0 + b_1 X_{1i} + \dots + b_7 X_{7i} + \varepsilon_i$, $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ i.i.d.
$i = 1, \dots, n$ $n = 65$

2) Coefficients : $b_0, b_1, \dots, b_7$ estimés par moindres carrés
σ^2 par maximum de vraisemblance.

3) H_0 : (M1) non significatif $\Leftrightarrow$ aucune var $X_1, \dots, X_7$ n'influe Y
$\Leftrightarrow b_1 = \dots = b_7 = 0$
Modèle réduit : $Y_i = b_0 + \varepsilon_i$

H_1 : (M1) significatif $\Leftrightarrow \exists j \in \{1, \dots, 7\}$ t.q. X_j influe Y
$\Leftrightarrow \exists j$ t.q. $b_j \neq 0$
Modèle complet : (M1)

Statistique de test : $Z = \dfrac{SM/7}{SR/57} \underset{H_0}{\sim} F(7, 57)$

Valeur statistique de test : $z = 3,939 \Rightarrow$ pvalue $= 0,0014 < 0.05$
$\Rightarrow H_0$ rejetée avec un risque $\alpha = 0.05$.

4) On teste X_1 : l'âge.
H_0 : X_1 n'influe pas $Y \mid X_2, \dots, X_7$ dans le modèle
$\Leftrightarrow b_1 = 0 \mid X_2, \dots, X_7$ dans le modèle
$\Leftrightarrow$ Modèle réduit : $Y_i = b_0 + b_2 X_{2i} + \dots + b_7 X_{7i} + \varepsilon_i$

H_1 : X_1 influe $Y \mid X_2, \dots, X_7$ dans le modèle
$\Leftrightarrow b_1 \neq 0 \mid$ —"—
$\Leftrightarrow$ Modèle complet : (M1)

Statistique de test $Z = \dfrac{B_1}{\sqrt{\widehat{\text{var}}(B_1)}} \underset{H_0}{\sim} t(n-p-1) = t(57)$
$p = 7$.

Valeur de la stat de test : $z = 0,321 \Rightarrow$ pvalue $= 0,75 > $
$\Rightarrow$ l'âge n'influe pas Y si $X_2, \dots, X_7$ dans le modèle.
Les variables significatives : X_2

Les variables : X_1, X_4, X_5, X_6, X_7 non signif. => à enlever.
La var X_3 a une pvalue $= 0,09$, on va la garder quand même dans le modèle, pour une étude plus poussée.

5) $\hat{b}_0 = 6,80$ => c'est la valeur prévue pour Y si $X_1 = X_2 = \dots = X_7 = 0$

$\hat{b}_1 = 0,004$ => avec chaque année de plus la valeur de Y augmente de $0,004$

$\hat{b}_2 = -1,487$ => si X_2 augmente alors Y diminue

$\vdots$

$\hat{b}_7 = -0.09$

$\hat{\sigma} = 0,93$

6) $R^2 adj = 0.24$ => (M1) de qualité médiocre.

7) $\hat{y}_i = \hat{b}_0 + \hat{b}_1 X_{1i} + \dots + \hat{b}_7 X_{7i}$ la prévision de Y_i

Le résidus correspondant : $e_i = y_i - \hat{y}_i$.

Pour tester la normalité on utilise le test de Shapiro.

$H_0 : e_i \sim N$ $\qquad H_1 : e_i \nsim N$.

Valeur stat de test : $W = 0,978$, pvalue $= 0,31 > 0.05$

=> H_0 acceptée, donc $e_i \sim N$ $\quad i = 1, \dots, n$.

Donc, on a eu raison de considérer comme méthode d'estimation celle des moindres carrés.

8) (M2) $\quad y_i = b_0 + b_2 X_{2i} + b_3 X_{3i} + \varepsilon_i \qquad i = 1, \dots, n$
$\varepsilon_i \sim N(0, \sigma^2)$ i.i.d.

9) Modèle (M2) significatif avec seulement la var X_2 significative. La var X_3 a une pvalue associée $0,13$ donc elle n'influe pas Y si X_2 dans le modèle.

$\hat{\sigma} = 0.939$ à peu près pareil que pour (M1).

$R^2 adj = 0,23$ pareil que pour (M1).

3

10) Les modèle (M3) et (M4) sont identiques :

(M3), (M4) $\quad Y_i = b_1 X_{1i} + \dots + b_7 X_{7i} + \varepsilon_i \quad , \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ i.i.d.

$i = 1, \dots, n$

Les coef de (M3) sont estimés par moindres carrés.

11) Les coef de (M3) sont estimés par la méthode LASSO adaptative :

$$(\hat{b}_{1n}^{aLASSO}, \dots, \hat{b}_{7n}^{aLASSO}) = \underset{(b_1, \dots, b_7) \in \mathbb{R}^7}{\arg\min} \left(\sum_{i=1}^{n} (y_i - b_1 X_{1i} - \dots - b_7 X_{7i})^2 + \lambda_n \sum_{j=1}^{7} \hat{\omega}_j |\beta_j| \right)$$

avec $\lambda_n \to \infty \quad , \quad \lambda_n = n^{-\frac{3}{5}+1}$

$\hat{\omega}_j = \frac{1}{|\hat{b}_{jn}^{LS}|^{\gamma}} \quad , \quad \gamma = \frac{2}{5} \quad , \quad \hat{b}_j^{LS}$ estim par moindres carrés.

$$(\hat{b}_{1n}^{LS}, \dots, \hat{b}_{7n}^{LS}) = \underset{b_1, \dots, b_7 \in \mathbb{R}^7}{\arg\min} \sum_{i=1}^{n} (y_i - b_1 X_{1i} - \dots - b_7 X_{7i})^2.$$

12) Les valeurs de $\hat{b}_{1n}^{LS}, \dots, \hat{b}_{7n}^{LS}$ ne sont égales à 0.

Les valeurs de $\begin{cases} \hat{b}_{jn}^{aLASSO} = 0 & \forall j \in \{1, 2, 3, 5, 6, 7\} \\ \hat{b}_{4n}^{aLASSO} \neq 0 \end{cases}$

13) $A1 = \{4\}$ contient l'indice j de la var X_j qui a des estimations non nulles pour sont coefficient par la méthode LASSO adaptative.

14) (M5) $\quad y_i = b_4 X_{4i} + \varepsilon_i \quad , \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ i.i.d, $i = 1, \dots, n$.

pvalue $= 10^{-16} \Rightarrow$ (M5) significatif

15) R^2adj par (M2) $= 0{,}23$

(M5) $= 0{,}848$

Le R^2 adj de (M2) est plus faible que celui de (M5). Une explication possible. l'élimination des var X_1, X_4, X_5, X_6, X_7 de (M1) n'est pas conseillée.

16) (M6) $Y_i = b_1 X_{1i} + \cdots + b_7 X_{7i} + \varepsilon_i$, ε_i i.i.d $i = 1, \dots, n$.
médiane $(\varepsilon_i) = 0$.

Les coef sont estimés par la méthode médiane ~~ou~~ moindres déviations :

$$(\hat{b}_{1n}^{LAD}, \dots, \hat{b}_{7n}^{LAD}) = \underset{b_1, \dots, b_7 \in \mathbb{R}^7}{\arg\min} \sum_{i=1}^{n} | Y_i - b_1 X_{1i} - \dots - b_7 X_{7i} |$$

17) H_0 : X_1 n'influe pas Y | $X_2, \dots, X_7$ sont dans le modèle
$\Leftrightarrow b_1 = 0$ | — " —
$\Leftrightarrow$ Modèle réduit : $Y_i = b_2 X_{2i} + \dots + b_7 X_{7i} + \varepsilon_i$.

H_1 : X_1 influe Y | $X_2, \dots, X_7$ dans le modèle
$\Leftrightarrow b_1 \neq 0$ | — " —
$\Leftrightarrow$ Modèle (M6)

Statistique de test : $Z = \dfrac{\hat{b}_{1n}^{LAD}}{\sqrt{\widehat{Var}(\hat{b}_{1n}^{LAD})}} \xrightarrow[\substack{n \to \infty \\ H_0}]{\mathcal{L}} \mathcal{N}(0,1)$

Valeur stat de test $z = 2{,}29$ p value $= 0{,}02 < 0{,}05$
$\Rightarrow H_0$ rejetée $\Rightarrow X_1$ influe.

Variables
- influentes : X_1, X_2
- non influentes : X_3, X_4, X_5, X_6, X_7. A enlever.

18) (M7) a été obtenu de (M6) en gardant les var. significatives
Pour X_1 : pvalue $= 4 \cdot 10^{-5}$ $\Rightarrow X_1$ influe Y si X_2 dans le modèle
X_2 : pvalue $= 0.33$ $\Rightarrow X_2$ n'influe pas Y si X_1 dans le modèle.

19) Par la méthode LASSO adaptative quantile ($\tau = \frac{1}{2}$) :

$$(\hat{b}_{1n}^{aq}, \dots, \hat{b}_{7n}^{aq}) = \underset{(b_1, \dots, b_7) \in \mathbb{R}^7}{\arg\min} \left(\sum_{i=1}^{n} |Y_i - b_1 X_{1i} - \dots - b_7 X_{7i}| + \lambda_n \sum_{j=1}^{7} \hat{\omega}_j |\beta_j| \right)$$

avec $\lambda_n = n^{2/5}$, $\hat{\omega}_j = \dfrac{1}{|\hat{b}_j^{LAD}|^{\gamma}}$, $\gamma = 1{,}225$

20) $A2=\{4\} \Rightarrow$ (M9) $y_i = b_4 X_{4i} + \varepsilon_i$, ε_i i.i.d , $i=1,\dots,n$
médiane$(\varepsilon_i)=0$.

Méthode des moindres déviations.

Exercice 2

Y_v = VSTATUS qui prend 2 valeurs 0, 1

1) On modélise $\pi(x) = \mathbb{P}[Y=1 \mid x]$, avec $x=(x_1,\dots,x_7)$
Par une régression logistique.

$$(M10): \quad \pi(x) = \frac{\exp(\beta^T x)}{1+\exp(\beta^T x)}$$

avec $\beta = (b_0, b_1, \dots, b_7)$

2) H_0 : X_1 n'influe pas $\pi(x) \mid X_2, X_3, \dots, X_7$ dans le modèle
$b_1 = 0$ | — II —
Modèle : $\pi(x) = \dfrac{\exp(b_0 + b_2 x_2 + \dots + b_7 X_7)}{1+\exp(b_0 + b_2 X_2 + \dots + b_7 X_7)}$

H_1 : X_1 influe $\pi(X) \mid X_2, \dots, X_7$ dans le modèle
$b_1 \neq 0$ | — II —
Modèle (M10)

Statistique de test : $Z = \dfrac{\hat{b}_1}{\sqrt{\widehat{var}(\hat{b}_1)}} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1)$ sous H_0

Valeur de la stat de test : $z = -0{,}85$, pvalue $= 0{,}39 > 0.05$
$\Rightarrow X_1$ n'influe pas $\pi(X)$.

Toutes les $X_1, \dots, X_7$ sont non significatives.

3) Sur 17 valeurs de $Y_v = 0$ on prévoit bien seulement 2 observations. $\Rightarrow$ taux erreur $15/17$
Sur les 48 obs $Y_v = 1$ on prévoit correctement 45 obs.
Taux d'erreur $\dfrac{3}{48}$

La mauvaise prévision de Y_v par (M10) est due au fait qu'aucune var $(X_j)_{1\leq j\leq n}$ n'influe $\Pi(*)$.

Exercice 3

1) On a un modèle d'analyse de variance à 2 facteurs F1 et F2 avec interaction.

F1 : PLATELET $<^0_1$ Chaque facteur a 2 valeurs.

F2 : FRAC $<^0_1$

(M11) $Y_{ijk} = \mu + F1_i + F2_j + (F1*F2)_{ij} + \varepsilon_{ijk}$

$i=1,2 \quad j=1,2 \quad k=1,\dots,n_{ij}$

$\varepsilon_{ijk} \sim N(0,\sigma^2)$ i.i.d.

Contraintes $\begin{cases} F1_1 + F1_2 = 0 \\ F2_1 + F2_2 = 0 \\ \forall j\in\{1,2\},\ \sum_{i=1}^{2}(F1*F2)_{ij} = 0 \\ \forall i\in\{1,2\},\ \sum_{j=1}^{2}(F1*F2)_{ij} = 0 \end{cases}$

$F1_1$ l'effet d'une plaquette anormale sur Y

$F1_2$ —"— normale —"—

$F2_1$ l'effet qu'il n'y a pas de fracture sur Y

$F2_2$ —"— il y a de fracture sur Y

$(F1*F2)_{12}$ l'effet de l'interaction entre une plaquette anormale et la présence d'une fracture, sur Y.

2) H_0 : (M11) non significatif

$\Leftrightarrow$ F1, F2, F1*F2 n'influe pas Y

$\Leftrightarrow F1_i = F2_j = (F1*F2)_{ij}$, $\forall i,j\in\{1,2\}$

Modèle : $Y_{ijk} = \mu + \varepsilon_{ijk}$.

H_1 : (M11) significatif $\Leftrightarrow$ F1 ou F2 ou $\cancel{F1}$F1*F2 influe Y
$\Leftrightarrow \exists i$ ou j t.q. $F1_i \neq 0$ ou $F2_j \neq 0$ ou $(F1*F2)_{ij} \neq 0$
Modèle (M11).

Statistique de test :

$$Z = \frac{SM/3}{SR/61} \underset{H_0}{\sim} F(3,61)$$

Valeur stat de test $z = 0{,}56$, pvaleur $= 0{,}64 \Rightarrow H_0$ acceptée
$\Rightarrow$ Modèle (M11) non significatif.

3) La prévision de y_i :

- par (M2) $\hat{y}_i = \hat{b}_0^{LS} + \hat{b}_2^{LS} X_{2i} + \hat{b}_3^{LS} X_{3i}$ $\quad i = 1,..,n$

avec $\hat{b}_j^{LS}$ estimateurs par moindres carrés.

- par (M11). $\hat{Y}_{ijk} = \hat{\mu} + \widehat{F1}_i + \widehat{F2}_j + \widehat{(F1*F2)}_{ij}$ $\quad i = 1,2$; $j = 1,2$; $k = 1,..,n_{ij}$

R^2adj par (M2) $= 0{,}23$
R^2adj par (M11) $= -0{,}02$

$\Rightarrow$ (M2) donne des meilleurs résultats d'ajustement car son R^2adj est plus grand.