Corrigé examen 14 décembre 2016
M2 "Modèles de régression"

Exercice 1

1) On modélise une variable qualitative fonction de variables numériques. On a donc un modèle logistique : y = rand, 2 valeurs $\begin{cases} R : \text{absence} \\ N : \text{présence} \end{cases}$

$\pi(x) = \mathbb{P}[Y = "R" \mid X = x]$ avec $x = (km91, km118, \dots, km338) = (x_1, x_2, \dots, x_7)$

Alors, le modèle logistique est :

$$(1) \quad \underbrace{\mathbb{P}[Y = "R" \mid X = x]}_{\text{proba d'absence de nids}} = \frac{\exp(b_0 + \sum_{j=1}^{7} b_j x_j)}{1 + \exp(b_0 + \sum_{j=1}^{7} b_j x_j)}$$

2) Le nombre de sites est 60. Le nombres de sites où il y a des nids de chouettes = 30.

3) Pour étudier la présence de nids de chouettes on regarde les sorties SAS :

H_0 : les 7 variables n'ont aucune influence sur l'absence des nids de chouettes

$b_1 = b_2 = \dots = b_7 = 0$

Modèle : $$(2) \quad \pi(x) = \mathbb{P}[Y = "R" \mid X = x] = \frac{\exp(b_0)}{1 + \exp(b_0)}$$

en fait, $\pi(x)$ est constante, ne dépend pas de x.

H_1 : $\exists j \in \{1, \dots, 7\}$ t.q. $b_j \neq 0 \Leftrightarrow \pi(x)$ influencé par au moins une des sept variables.

Modèle : (1)

Par les trois statistiques de test : rapport de vraisemblance, score, Wald, l'hypothèse H_0 est rejetée (pval < 0.05) et donc le modèle (1) significatif.

Indicateurs quantitatifs :

Pour (1) AIC = 68.134 BIC = 84.888

(2) AIC = 85.178 BIC = 87.272

Les deux critères diminuent pour le modèle (1).

4) On étudie les sorties R. Les sorties SAS donnent les mêmes résultats.

H_0 : X_1 n'influe pas $\mathbb{P}[Y = "R"]$ si $X_2, \dots, X_7$ sont dans le modèle

$H_0 : \beta_1 = 0 \mid X_2, \dots, X_7$ dans le modèle

Modèle :

$$(3)\quad \mathbb{P}[Y = "R" \mid X = x] = \frac{\exp\left(\beta_0 + \sum_{j=2}^{7} \beta_j x_j\right)}{1 + \exp\left(\beta_0 + \sum_{j=2}^{7} \beta_j x_j\right)}$$

$H_1 : \beta_1 \neq 0 \mid X_2, \dots, X_7$ dans le modèle.

Modèle : (1).

Stat de test : $Z_1 = \frac{B_1}{\hat{\sigma}_{B_1}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$

La réalisation : $z_1 = -1{,}535 \Rightarrow$ p-value = 0.12

$\Rightarrow H_0$ acceptée.

Conclusion : les var. qui influent $\mathbb{P}[Y = "R"]$ sont : km118, km140

Rq : $\beta_0 \neq 0$. On garde aussi km91 et km241

5) Le Modèle (m2) est :

$$(4)\quad P[Y="R" \mid X=x) = \frac{\exp(b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + b_6 x_6}{1+\exp(b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + b_6 x_6)}$$

Les estimations des paramètres :

$\hat{b}_0 = 10,2758$ $\quad \hat{b}_2 = 0,1199$ $\quad \hat{b}_6 = -0,08$

$\hat{b}_1 = -0,0643$ $\quad \hat{b}_3 = -0,12$

La var x_2 = km118 a une influence positive sur la probabilité d'absence de nids pendant que km91, km140 et km241 ont une influence négative.

Prévisions par (M2) :

- parmi les 30 sites avec nids, 24 sont bien prévus
- — " sans — " —

Taux d'erreur : $\frac{1}{2}\left(\frac{6}{30} + \frac{6}{30}\right) = \frac{1}{5} = 20\%$

Exercice 2

1) Les variables explicatives "site" et "faam" sont de type qualitatives, pendant que la variable expliquée "biomass" est numérique. Donc, on a un modèle d'analyse de variance à 2 facteurs

2) Le nb d'obs totales : 612
Nb d'obs utilisées : 450
Le modèle (M3) est :

$$Y_{ijk} = \mu + (site)_i + (faam)_j + (site * faam)_{ij} + \varepsilon_{ijk}$$

avec : $\varepsilon_{ijk} \sim N(0, \sigma^2)$ $\quad i = 1, \dots, 6$ $\quad j = 1,2,3,4$ $\quad k = 1, \dots, n_{ij}$

$\varepsilon_{ijk} \perp\!\!\!\perp \varepsilon_{i'j'k'}$

$(site)_i$: l'effet du site i sur la biomasse

Y : biomasse

$(faam)_j$: l'effet de la famille j sur la biomasse
$(site * faam)_{ij}$ l'effet de l'interaction entre le site i et la famille j sur Y.

Contraintes : $\sum_{i=1}^{6} (site)_i = 0$, $\sum_{j=1}^{4} (faam)_j = 0$

$\forall i = 1, \dots, 6$, $\sum_{j=1}^{4} (site * faam)_{ij} = 0$

$\forall j = 1, 2, 3, 4$ $\sum_{i=1}^{6} (site * faam)_{ij} = 0$

Les valeurs des variables :

"site" : $1, 2, \dots, 6$

"faam" : $1, 2, 3, 4$

3) H_0 : le modèle (M3) n'est pas significatif
$\Leftrightarrow$ le site et la famille des pins et leur interaction n'ont aucune influence sur la biomasse
$\Leftrightarrow \forall site_i = 0$, $\forall faam_j = 0$, $\forall (site * faam)_{ij} = 0$
$\Leftrightarrow$ Modèle : $Y_{ijk} = \mu + \varepsilon_{ijk}$

H_1 : (M3) significatif
$\Leftrightarrow$ Y influencé par le site ou par la famille ou par leur interaction.
$\Leftrightarrow \exists i$, $site_i \neq 0$ ou $\exists faam_j \neq 0$ ou $\exists i, j$ $(site * faam)_{ij} \neq 0$
$\Leftrightarrow$ Modèle (M3)

Statistique de test :

$$Z = \frac{SM/17}{SR/432} \underset{H_0}{\sim} F(17, 432)$$

Valeur stat de test : $z = 16,78$ et p-value $< 10^{-4}$
$\Rightarrow$ H_0 rejetée $\Rightarrow$ (M3) significatif.

Tests des 2 facteurs et de leur interaction :

H_0 : le site n'influe pas la biomasse | la famille et l'interaction site*famille sont dans le modèle

$\Leftrightarrow \forall i = 1, \dots, 6 \quad site_i = 0$ (...

Modèle : $Y_{ijk} = \mu + (faam)_j + (site * faam)_{ij} + \varepsilon_{ijk}$

H_1 : le site influe | fam et site*faam dans le modèle

$\Leftrightarrow \exists i, \ site_i = 0$

Modèle : (M3)

Stat de test : $Z = \dfrac{S_{site}/5}{SR/432} \underset{H_0}{\sim} F(5, 432)$

Valeur stat de test $z = 54,97$, pvalue $< 10^{-4}$

$\Rightarrow$ le site signif.

Pareil, on obtient que "faam" et "site*faam" n'influent pas Y, si on garde les autres var dans le modèle.

4) (M4) : $Y_{ij} = \mu + site_i + \varepsilon_{ij}$ $\quad i = 1, \dots, 6$, $j = 1, \dots, n_i$

$\sum_{i=1}^{6} site_i = 0$ $\quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$

$\varepsilon_{ij} \perp\!\!\!\perp \varepsilon_{i'j'}$

5) H_0 : pour (M4) le site n'influe pas Y

$\Leftrightarrow site_i = 0 \quad \forall i = 1, \dots, 6$

$\Leftrightarrow$ Modèle : $Y_{ij} = \mu + \varepsilon_{ij}$

H_1 : le site influe Y dans (M4)

$\Leftrightarrow$ Modèle (M4)

Stat de test : $Z = \frac{SM/5}{SR/444} \underset{H_0}{\sim} F(5, 444)$

La valeur de la stat de test $Z = 55.79$

p value $< 10^{-4}$ $\Rightarrow$ le site influe $Y \Rightarrow$ (M4) signif

6) $\hat{\sigma}^2 = 129807,23$
$\hat{\mu} = 773,83$
$\widehat{site}_1 = 301,5$
$\vdots$
$\widehat{site}_5 = 30,57$
$\widehat{site}_6 = -\sum_{j=1}^{5} \widehat{site}_j$

} sorties R.

Par rapport à la moyenne, un pin du site = 1 a une biomass plus grande avec 301,5, le site = 2 a une biomasse inférieure de 479,7

7) On utilise les sorties SAS : on a que l'hypothese $site_4 = 0$ est acceptée. Mais la contrainte est $site_6 = 0$. Donc les sites 4 et 6 ont la même biomasse.

8) Pour (M3) $R^2_{adj} = 0,374$ donc on a les
(M4) : $R^2_{adj} = 0,379$ mêmes R^2 ajustés

parce que (M4) est obtenu à partir de (M3) en supprimant les variables non significatives.
Donc par (M4) et (M3) on obtient la même qualité de prévision : assez bonne.