Université de Montpellier Année 2016-2017
HMMA304 - Modèles linéaires généralisés

Examen final - 7 novembre 2016
Durée 2h - Documents interdits

Exercice 1 (8 pts)

1) (2 pts) Expliquer l'objectif d'une modélisation log-linéaire graphique ?

2) (2 pts) Représenter le modèle log-linéaire graphique associé à l'équation suivante

$$x_1 * x_2 * x_3 + x_1 * x_4 + x_4 * x_5 * x_6 \, .$$

Donner deux relations d'indépendance conditionnelle entre les variables qui le composent.

3) (2 pts) Nous considérons 100 réalisations suivant deux variables qualitatives ayant chacune deux modalités. Donner la vraisemblance associée au modèle log-linéaire graphique poissonien d'indépendance. Montrer qu'il s'agit bien d'un modèle linéaire généralisé.

4) (2 pts) Nous disposons d'un jeu de données R (`data.frame`) nommée `vars` contenant 8 variables qualitatives. Donner le code R permettant de mettre en oeuvre le modèle saturé, le modèle d'indépendance et de déterminer le meilleur modèle log-linéaire graphique au sens du critère AIC par une méthode descendante.

Exercice 2 (7 pts)

On considère un n-échantillons $(Y_1, x_1), \ldots, (Y_n, x_n)$ du couple (Y, x) où Y est une variable aléatoire à valeurs dans $\{0, 1\}$ et x est une variable fixée prenant également les valeurs $\{0, 1\}$. On suppose qu'il existe une variable latente Y_i^* telle que $Y_i^* | x_i \sim \mathcal{N}(a + bx_i, 1)$ et

$$Y_i = \begin{cases} 1 & \text{si} \quad Y_i^* \leq 0 \\ 0 & \text{si} \quad Y_i^* > 0 \end{cases} .$$

On note $\Phi(\cdot)$ la fonction de répartition de la loi normale centrée réduite ($\Phi(u) = \mathbb{P}(\mathcal{N}(0,1) \leq u)$).

1 (2 pts) Montrer qu'il s'agit d'un modèle linéaire généralisée. En suivant les notations du cours quant aux familles exponentielles à un paramètre de nuisance, on explicitera les paramètres ϕ, θ, les fonctions r et b ainsi que les régresseurs à considérer.

2 (1 pt) Montrer que ce n'est pas le lien canonique qui a été choisi.

3 (2 pts) Nous disposons de 100 observations (y_i, x_i) décrites dans la table de contingence suivante

	$y = 0$	$y = 1$
$x = 0$	26	26
$x = 1$	32	16

.

On note $\alpha = \Phi(a)$ et $\beta = \Phi(a + b)$ donner la log-vraisemblance en fonction de α et β.

4 (2 pts) Calculer l'estimateur du maximum de vraisemblance du couple (α, β).

Exercice 3 (7 pts)

Commenter en détails le fichier R Markdown fourni en annexes. Il s'agit de données recueillies par l'Insee de février à avril 2003 dans le cadre de l'étude "Histoire de vie".

Annexes Examen HMMA304

Jean-Michel Marin

7 novembre 2016

```
library(questionr)
data(hdv2003)
d <- hdv2003
d <- d[,-c(1,5,7,9,11,12,13)]
summary(d)
```

```
##       age           sexe
##  Min.   :18.00   Homme: 899
##  1st Qu.:35.00   Femme:1101
##  Median :48.00
##  Mean   :48.16
##  3rd Qu.:60.00
##  Max.   :97.00
##
##                                                    nivetud
##  Enseignement technique ou professionnel court       :463
##  Enseignement superieur y compris technique superieur:441
##  Derniere annee d'etudes primaires                   :341
##  1er cycle                                           :204
##  2eme cycle                                          :183
##  (Other)                                             :256
##  NA's                                                :112
##                    occup      freres.soeurs
##  Exerce une profession:1049   Min.   : 0.000
##  Chomeur              : 134   1st Qu.: 1.000
##  Etudiant, eleve      :  94   Median : 2.000
##  Retraite             : 392   Mean   : 3.283
##  Retire des affaires  :  77   3rd Qu.: 5.000
##  Au foyer             : 171   Max.   :22.000
##  Autre inactif        :  83
##                          relig     lecture.bd peche.chasse cuisine
##  Pratiquant regulier        :266   Non:1953   Non:1776     Non:1119
##  Pratiquant occasionnel     :442   Oui:  47   Oui: 224     Oui: 881
##  Appartenance sans pratique :760
##  Ni croyance ni appartenance:399
##  Rejet                      : 93
##  NSP ou NVPR                : 40
##
##  bricol     cinema     sport        heures.tv
##  Non:1147   Non:1174   Non:1277   Min.   : 0.000
##  Oui: 853   Oui: 826   Oui: 723   1st Qu.: 1.000
##                                   Median : 2.000
##                                   Mean   : 2.247
##                                   3rd Qu.: 3.000
##                                   Max.   :12.000
##                                   NA's   :5
```

```r
d$age <- cut(d$age, c(16, 25, 45, 65, 97), right = FALSE, include.lowest = TRUE)
d$etud <- d$nivetud
levels(d$etud)
```

```
## [1] "N'a jamais fait d'etudes"
## [2] "A arrete ses etudes, avant la derniere annee d'etudes primaires"
## [3] "Derniere annee d'etudes primaires"
## [4] "1er cycle"
## [5] "2eme cycle"
## [6] "Enseignement technique ou professionnel court"
## [7] "Enseignement technique ou professionnel long"
## [8] "Enseignement superieur y compris technique superieur"
```

```r
levels(d$etud) <- c("Primaire", "Primaire", "Primaire",
  "Secondaire", "Secondaire", "Technique/Professionnel",
  "Technique/Professionnel", "Supérieur")
freq(d$etud)
```

```
##                            n    % val%
## Primaire                 466 23.3 24.7
## Secondaire               387 19.4 20.5
## Technique/Professionnel 594 29.7 31.5
## Supérieur                441 22.1 23.4
## NA                       112  5.6   NA
```

```r
levels(d$etud)
```

```
## [1] "Primaire"                "Secondaire"
## [3] "Technique/Professionnel" "Supérieur"
```

```r
d$etud <- addNA(d$etud)
levels(d$etud)
```

```
## [1] "Primaire"                "Secondaire"
## [3] "Technique/Professionnel" "Supérieur"
## [5] NA
```

```r
d <- na.omit(d[,-3])
summary(d)
```

```
##       age         sexe                          occup      freres.soeurs
##  [16,25):168   Homme: 895   Exerce une profession:1047   Min.   : 0.000
##  [25,45):705   Femme:1100   Chomeur              : 131   1st Qu.: 1.000
##  [45,65):742                Etudiant, eleve      :  94   Median : 2.000
##  [65,97]:380                Retraite             : 392   Mean   : 3.283
##                             Retire des affaires  :  77   3rd Qu.: 5.000
##                             Au foyer             : 171   Max.   :22.000
##                             Autre inactif        :  83
##                          relig     lecture.bd peche.chasse cuisine
##  Pratiquant regulier        :266   Non:1948   Non:1772     Non:1116
##  Pratiquant occasionnel     :441   Oui:  47   Oui: 223     Oui: 879
##  Appartenance sans pratique :759
##  Ni croyance ni appartenance:397
##  Rejet                      : 92
##  NSP ou NVPR                : 40
##
##  bricol     cinema     sport        heures.tv
```

```
##  Non:1145   Non:1171   Non:1275   Min.   : 0.000
##  Oui: 850   Oui: 824   Oui: 720   1st Qu.: 1.000
##                                   Median : 2.000
##                                   Mean   : 2.247
##                                   3rd Qu.: 3.000
##                                   Max.   :12.000
##
##                       etud
##  Primaire               :465
##  Secondaire             :387
##  Technique/Professionnel:591
##  Supérieur              :440
##  NA                     :112
##
##
```

```
levels(d$sport)
```

```
## [1] "Non" "Oui"
```

```
model0 <- glm(sport ~ sexe,data=d,family=binomial)
summary(model0)
```

```
##
## Call:
## glm(formula = sport ~ sexe, family = binomial, data = d)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0255  -1.0255  -0.8813   1.3373   1.5058
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.36836    0.06799  -5.418 6.03e-08 ***
## sexeFemme   -0.37707    0.09374  -4.022 5.76e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2609.2  on 1994  degrees of freedom
## Residual deviance: 2593.0  on 1993  degrees of freedom
## AIC: 2597
##
## Number of Fisher Scoring iterations: 4
```

```
d$sexe <- relevel(d$sexe, ref="Femme")
model1 <- glm(sport ~ sexe,data=d,family=binomial)
summary(model1)
```

```
##
## Call:
## glm(formula = sport ~ sexe, family = binomial, data = d)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
```

```
## -1.0255  -1.0255  -0.8813   1.3373   1.5058
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.74543    0.06454 -11.550  < 2e-16 ***
## sexeHomme    0.37707    0.09374   4.022 5.76e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2609.2  on 1994  degrees of freedom
## Residual deviance: 2593.0  on 1993  degrees of freedom
## AIC: 2597
##
## Number of Fisher Scoring iterations: 4
```

```
levels(d$sport)
```

```
## [1] "Non" "Oui"
```

```
d$sport <- relevel(d$sport, ref="Oui")
table(d$sport,d$sexe)
```

```
##
##       Femme Homme
##   Oui   354   366
##   Non   746   529
```

```
model2 <- glm(sport ~ sexe,data=d,family=binomial)
summary(model2)
```

```
##
## Call:
## glm(formula = sport ~ sexe, family = binomial, data = d)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5058  -1.3373   0.8813   1.0255   1.0255
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.74543    0.06454  11.550  < 2e-16 ***
## sexeHomme   -0.37707    0.09374  -4.022 5.76e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2609.2  on 1994  degrees of freedom
## Residual deviance: 2593.0  on 1993  degrees of freedom
## AIC: 2597
##
## Number of Fisher Scoring iterations: 4
```

```r
model3 <- glm(sport ~ sexe + etud + relig + heures.tv,data=d,family=binomial)
anova(model3,test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: sport
##
## Terms added sequentially (first to last)
##
##
##           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                       1994     2609.2
## sexe       1    16.21      1993     2593.0 5.668e-05 ***
## etud       4   319.19      1989     2273.8 < 2.2e-16 ***
## relig      5     1.65      1984     2272.2 0.8953896
## heures.tv  1    13.18      1983     2259.0 0.0002831 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
model4 <-  glm(sport~.,data=d,family=binomial)
anova(model4,test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: sport
##
## Terms added sequentially (first to last)
##
##
##               Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                           1994     2609.2
## age             3  213.226      1991     2396.0 < 2.2e-16 ***
## sexe            1   19.253      1990     2376.7 1.145e-05 ***
## occup           6   30.610      1984     2346.1 3.009e-05 ***
## freres.soeurs   1   24.625      1983     2321.5 6.963e-07 ***
## relig           5    4.401      1978     2317.1 0.4932429
## lecture.bd      1    3.730      1977     2313.4 0.0534341 .
## peche.chasse    1    0.123      1976     2313.2 0.7260114
## cuisine         1   17.894      1975     2295.3 2.335e-05 ***
## bricol          1   13.545      1974     2281.8 0.0002329 ***
## cinema          1   71.717      1973     2210.1 < 2.2e-16 ***
## heures.tv       1   21.002      1972     2189.1 4.588e-06 ***
## etud            4   67.891      1968     2121.2 6.325e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
library(caret)
```

```
## Loading required package: lattice
```

```
## Loading required package: ggplot2
```

```r
res3 <- train(formula(model3),data=d,method="glm",family=binomial,
    metric="Accuracy",trControl=trainControl(method="repeatedcv",number=10,
        repeats=50))
confusionMatrix(res3)
```

```
## Cross-Validated (10 fold, repeated 50 times) Confusion Matrix
##
## (entries are percentual average cell counts across resamples)
##
##           Reference
## Prediction  Oui  Non
##        Oui 16.0  9.7
##        Non 20.1 54.2
##
##  Accuracy (average) : 0.7023
```

```r
res4 <- train(formula(model4),data=d,method="glm",family=binomial,
    metric="Accuracy",trControl=trainControl(method="repeatedcv",number=10,
        repeats=50))
confusionMatrix(res4)
```

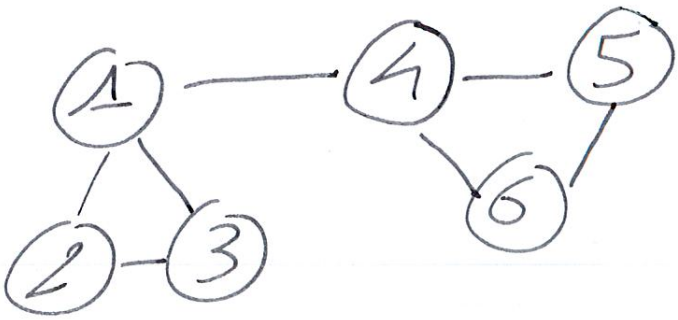
```
## Cross-Validated (10 fold, repeated 50 times) Confusion Matrix
##
## (entries are percentual average cell counts across resamples)
##
##           Reference
## Prediction  Oui  Non
##        Oui 18.8 10.7
##        Non 17.3 53.2
##
##  Accuracy (average) : 0.7208
```

Correction Examen final
HMMA304 - 07/11/2016

Exercice 1

1| Les modèles log-linéaires graphiques sont utilisés pour étudier les relations qui existent entre différentes variables qualitatives.

2|



$(X_5, X_6) \perp\!\!\!\perp (X_1, X_2, X_3) \mid X_4$

$X_2 \perp\!\!\!\perp (X_4, X_5, X_6) \mid X_1, X_3$

3] 2 variables qualitatives

Les données sont synthétisées dans une table de contingence 2×2.

Le modèle d'indépendance s'écrit

$$\log(\mu_{ij}) = \mu + \alpha_i + \beta_j \quad i \in \{1,2\} \quad j \in \{1,2\}$$

avec les contraintes

$$\alpha_1 = -\alpha_2$$

$$\beta_1 = -\beta_2$$

$$y_{ij} \sim \mathcal{P}(\mu_{ij})$$

$$y = (y_{11}, y_{12}, y_{21}, y_{22})$$

$$V(\mu, \alpha_1, \beta_1; y) = \prod_{i=1}^{2} \prod_{j=1}^{2} \left[e^{-e^{\mu+\alpha_i+\beta_j}} \right] \frac{e^{y_{ij}[\mu+\alpha_i+\beta_j]}}{y_{ij}!}$$

Il s'agit bien d'un modèle linéaire généralisé car

i] famille exponentielle (trivial)

ii] $\log\left[\mathbb{E}(Y_{ij})\right] = \mu + \alpha_i + \beta_j$

$= \mathbb{1}_{ij}^T \theta$ avec

$\theta = (\mu, \alpha_1, \alpha_2)$

$\mathbb{1}_{11} = (1, 1, 1)$

$\mathbb{1}_{12} = (1, 1, -1)$

$\mathbb{1}_{21} = (1, -1, 1)$

$\mathbb{1}_{22} = (1, -1, -1)$

4] library(glim)

```
model1 <- glmod(~.^., data=vary)
model2 <- glmod(~.^1, data=vary)
model3 <- backward(model1,
                  crit="aic")
```

Exercice 2

1) Soit $\pi_x = \mathbb{P}(Y=1|x)$

$$f(y;\pi_x) = \pi_x^{y}(1-\pi_x)^{1-y}\,\mathbb{1}_{\{y\in\{0,1\}\}}$$

$$f(y;\pi_x) = \exp\{y\log(\pi_x) + (1-y)\log(1-\pi_x)\}$$

$$f(y;\pi_x) = \exp\{y\log\left(\frac{\pi_x}{1-\pi_x}\right) + \log(1-\pi_x)\}$$

$\mathbb{1}_{\{y\in\{0,1\}\}}$ $\mathbb{1}_{\{y\in\{0,1\}\}}$

La loi de Bernoulli appartient à la famille exponentielle

$$\theta = \log\left(\frac{\pi_x}{1-\pi_x}\right) \Rightarrow \pi_x = \frac{e^\theta}{1+e^\theta}$$

$$b(\theta) = -\log\left(1-\frac{e^\theta}{1+e^\theta}\right) = \log(1+e^\theta)$$

$$\phi = 1 \quad , \quad c(y,\phi) = 0$$

$$V(y) = \pi(y)\,\mathbb{1}_{\{y\in\{0,1\}\}}$$

Par ailleurs,

$$\mathbb{E}(Y) = \pi_x = \mathbb{P}(Y=1|x)$$

$$= \mathbb{P}(Y^* \leq 0 | x)$$

$$= \mathbb{P}(Y^* - \omega - bx \leq -\omega - bx | x)$$

$$= \mathbb{P}(N(0,1) \leq -\omega - bx | x)$$

$$= \Phi(-\omega - bx)$$

$$\Rightarrow -\Phi^{-1}(\mathbb{E}(Y)) = \omega + bx$$

Il s'agit donc bien d'un modèle linéaire généralisé -

$$\eta(q) = \Phi(-q)$$

2] Nous avons $b'(q) = \frac{e^q}{1+e^q}$

$\neq \eta(q)$ - ce n'est pas le lien canonique qui a été utilisé.

3] $y = (y_1, \dots, y_m)$ et $x = (x_1, \dots, x_m)$

$$\mathcal{L}V(a, b; \underline{y}, \underline{x})$$

$$= \sum_{i=1}^{m} \left[y_i \log\left(\Phi(-a - b x_i)\right) + (1 - y_i) \log\left(1 - \Phi(-a - b x_i)\right) \right]$$

$$\mathcal{L}V(a, b; \underline{y}, \underline{x}) =$$

$$\sum_{i=1}^{m} \left[y_i \log\left(\Phi(-a - b x_i)\right) + (1 - y_i) \log\left(\Phi(a + b x_i)\right) \right]$$

$$\mathcal{L}V(a, b; \underline{y}, \underline{x}) = 16 \log\left(\Phi(-a - b)\right) + 96 \log\left(\Phi(-a)\right) + 32 \log\left(\Phi(a + b)\right) + 26 \log\left(\Phi(a)\right)$$

$$LV(\alpha, \beta; \underline{y}, \underline{k})$$

$$= 16 \log(1-\beta) + 26 \log(1-\alpha) + 32 \log(\beta) + 26 \log(\alpha)$$

4] $$\frac{\partial LV}{\partial \alpha}(\alpha^*, \beta^*; \underline{y}, \underline{k}) = -\frac{26}{1-\alpha^*} + \frac{26}{\alpha^*} = 0$$

$$\frac{\partial LV}{\partial \beta}(\alpha^*, \beta^*; \underline{y}, \underline{k}) = -\frac{16}{1-\beta^*} + \frac{32}{\beta^*} = 0$$

$$\Leftrightarrow \begin{cases} 26\alpha^* = 26(1-\alpha^*) \\ 32(1-\beta^*) = 16\beta^* \end{cases}$$

$$\Leftrightarrow \begin{cases} \alpha^* = \frac{26}{52} = \frac{1}{2} \\ \beta^* = \frac{32}{48} = \frac{2}{3} \end{cases}$$

$$\frac{\partial^2 LV}{\partial\alpha\partial\beta}(\alpha,\beta;\underline{y},\underline{x}) = 0$$

$$\frac{\partial^2 LV}{(\partial\alpha)^2}(\alpha,\beta;\underline{y},\underline{x}) = -\frac{26}{\alpha^2} - \frac{16}{(1-\alpha)^2} < 0$$

$$\frac{\partial^2 LV}{(\partial\beta)^2}(\alpha,\beta;\underline{y},\underline{x}) = -\frac{32}{\beta^2} - \frac{26}{(1-\beta)^2} < 0$$

La fonction LV est strictement -concave-

Ainsi,

$$\widehat{\alpha} = \frac{13}{21} \text{ et } \widehat{\beta} = \frac{16}{29}$$

Exercice 3

model 0, model 1, model 2

trois modèles de regression logistique où la variable sport est expliquée par la taille

Les trois modèles sont identiques, seules les modalités de référence des facteurs (à expliquer et explicatif) changent.

Pour le modèle 0,

$$TP(Sport = \text{"Oui"} \mid Sexe = \text{"Femme"})$$

$$\# \frac{e^{-0.37-0.38}}{1+e^{-0.37-0.38}} \# 0.32$$

$$TP(Sport = \text{"Oui"} \mid Sexe = \text{"Homme"})$$

$$\# \frac{e^{-0.37}}{1+e^{-0.37}} \# 0.40$$

Pour le modèle 1,

$$\mathbb{P}(\text{Sport} = \text{"Oui"} \mid \text{Sexe} = \text{"Homme"}) \approx \frac{e^{-0.74+0.37}}{1+e^{-0.74+0.37}} \approx 0.40$$

$$\mathbb{P}(\text{Sport} = \text{"Oui"} \mid \text{Sexe} = \text{"Femme"}) \approx \frac{e^{-0.75}}{1+e^{-0.75}} \approx 0.32$$

Pour le modèle 2,

$$\mathbb{P}(\text{Sport} = \text{"Non"} \mid \text{Sexe} = \text{"Homme"}) \approx \frac{e^{0.74-0.37}}{1+e^{0.74-0.37}} \approx 0.60$$

$$\mathbb{P}(\text{Sport} = \text{"Non"} \mid \text{Sexe} = \text{"Femme"}) \approx \frac{e^{0.75}}{1+e^{0.75}} \approx 0.68$$

Les modèles de regression logistique modèle 3 et modèle 4 sont comparés à l'aide d'une technique de validation croisé à 10 ensembles répétés 50 fois.

Le modèle 4 a de meilleures performances en terme de taux d'erreur de classification.