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

Examen de contrôle continu
sans document - 12 octobre 2016

Exercice 1 (5 pts)

Nous considérons le modèle de régression linéaire tel que pour tout $i \in \{1,\ldots,10\}$ et $j \in \{1,\ldots,3\}$, nous avons

$$y_{i,j} = \alpha x_i + \epsilon_{i,j}$$

où α est un paramètre inconnu et $\epsilon_{i,j}$ est un variable aléatoire réelle telle que :
$\mathbb{E}(\epsilon_{i,j}) = 0$, $\mathbb{V}(\epsilon_{i,j}) = 1$,
$\mathbb{C}(\epsilon_{i,j}, \epsilon_{i',j'}) = 0$ lorsque $i \neq i'$,
$\mathbb{C}(\epsilon_{i,j}, \epsilon_{i',j'}) = 1/2$ lorsque $i = i'$.

1 (2 pts) Donner l'expression de l'estimateur des Moindres Carrés Ordinaires $\widehat{\alpha}$ de α et calculer sa variance.

2 (3 pts) Donner le code R permettant de calculer l'estimateur des Moindres Carrés Généralisés de α.

Exercice 2 (9 pts)

On considère n variables aléatoires indépendantes $y_1,\ldots,y_n$ telles que y_i est distribué suivant une loi négative binomiale de paramètres (h,π_i). Nous supposons que h est fixé. La loi négative binomiale modélise, dans le contexte d'une suite d'épreuves de Bernoulli indépendantes, le nombre d'essais nécessaires pour obtenir h succès, π_i représente la probabilité de succès. Nous avons

$$f(y_i;\pi_i) = C_{y_i-1}^{h-1}\pi_i^h(1-\pi_i)^{y_i-h}\mathbf{1}_{\{h,h+1,\ldots\}}(y_i)\,.$$

Pour tout $i = 1,\ldots,n$, on suppose que $\log(\pi_i/(1-\pi_i)) = \alpha + \beta x_i$ où $x_i \in \mathbb{R}$ est supposé connu, α et β sont des paramètres inconnus.

1 (2 pts) Montrer qu'il s'agit d'un modèle linéaire généralisée. La fonction de lien canonique a-t-elle été utilisée ?

2 (2 pts) Donner la log-vraisemblance et les équations de vraisemblance. Pouvons-nous calculer les expressions analytique des estimateurs du maximum de vraisemblance de α et β ?

3 (2 pts) Calculer la matrice d'information de Fisher apportée par $(y_1, \ldots, y_n)$ sur les paramètres α et β notée $I_n(\alpha, \beta)$.

4 (3 pts) Modifier le modèle en utilisant le lien canonique et donner les équations de vraisemblance.

Exercice 3 (6 pts)

1 (3 pts) Expliquer les résultats produits par la code R suivant (le jeu de données est décrit en annexe).

```
library(klaR)
data(GermanCredit)
model1 <- glm(credit_risk~duration,family=binomial,data=GermanCredit)
model2 <- glm(credit_risk~.,data=GermanCredit,family=binomial)
model3 <- step(model2,direction="backward")
model4 <- step(model2,direction="backward",k=log(n))
model5 <- glm(credit_risk~1,data=GermanCredit,family=binomial)
model6 <- step(model5,direction="both",scope=formula(model2))
```

2 (3 pts) Donner le code R permettant d'estimer par validation croisée à 2 ensembles répétée 10 fois les taux d'erreurs associés aux modèles considérés dans la question précédente.

Annexes

GermanCredit (klaR) R Documentation
Statlog German Credit

Description

The dataset contains data of past credit applicants.
The applicants are rated as good or bad. Models of this
data can be used to determine if new applicants present
a good or bad credit risk.

A data frame containing 1,000 observations on 21 variables.

status
factor variable indicating the status of the existing checking account,
with levels ... < 100 DM, 0 <= ... < 200 DM, ... >= 200 DM/salary
for at least 1 year and no checking account

duration
duration in months

credit_history
factor variable indicating credit history, with levels no credits
taken/all credits paid back duly, all credits at this bank paid back duly,
existing credits paid back duly till now, delay in paying off
in the past and critical account/other credits existing

purpose
factor variable indicating the credit's purpose, with levels car (new),
car (used), furniture/equipment, radio/television, domestic appliances,
repairs, education, retraining, business and others

amount
credit amount.

savings
factor. savings account/bonds, with levels ... < 100 DM, 100 <= ... < 500 DM,
500 <= ... < 1000 DM, ... >= 1000 DM and unknown/no savings account

employment_duration
ordered factor indicating the duration of the current employment,
with levels unemployed, ... < 1 year, 1 <= ... < 4 years, 4 <= ... < 7 years and
... >= 7 years

installment_rate
installment rate in percentage of disposable income.

personal_status_sex
factor variable indicating personal status and sex, with levels male:divorced/separated,
female:divorced/separated/married, male:single, male:married/widowed and female:single

other_debtors
factor. Other debtors, with levels none, co-applicant and guarantor

present_residence
present residence since

property
factor variable indicating the client's highest valued property, with levels real estate, building society savings agreement/life insurance, car or other and unknown/no property

age
client's age

other_installment_plans
factor variable indicating other installment plans, with levels bank, stores and none

housing
factor variable indicating housing, with levels rent, own and for free

number_credits
number of existing credits at this bank

job
factor indicating employment status, with levels unemployed/unskilled - non-resident, unskilled - resident, skilled employee/official and management/self-employed/highly qualified employee/officer

people_liable
number of people being liable to provide maintenance

telephone
binary variable indicating if the customer has a registered telephone number

foreign_worker
binary variable indicating if the customer is a foreign worker

credit_risk
binary variable indicating credit risk, with levels good and bad

Correction examen contrôle continu
12 octobre 2016
HMMA 304

Exercice 1

1) $\hat{\alpha} \in \arg\min_{\alpha} \sum_{i=1}^{10} \sum_{j=1}^{3} (y_{ij} - \alpha x_i)^2$

$\hat{\alpha} = (X^T X)^{-1} X^T y$

où $X = (x_1, x_1, x_1, \ldots, x_{10}, x_{10}, x_{10})$

Ainsi, $X^T X = 3 \sum_{i=1}^{10} x_i^2$

$(X^T X)^{-1} = \frac{1}{3 \sum_{i=1}^{10} x_i^2}$ et $X^T y = \sum_{i=1}^{10} \sum_{j=1}^{3} x_i y_{ij}$

$= \sum_{i=1}^{10} x_i \sum_{j=1}^{3} y_{ij}$

②

Nous obtenons

$$\hat{\alpha} = \frac{\sum_{i=1}^{10} x_i \left(\sum_{j=1}^{3} y_{ij} / 3 \right)}{\sum_{i=1}^{10} x_i^2}$$

$$\mathbb{V}(\hat{\alpha}) = \frac{1}{\left(\sum_{i=1}^{10} x_i^2\right)^2} \sum_{i=1}^{10} x_i^2 \left(\frac{1}{9}\right) \mathbb{V}\left(\sum_{j=1}^{3} y_{ij}\right)$$

$$\mathbb{V}\left(\sum_{j=1}^{3} y_{ij}\right) = \mathbb{C}\left(y_{i1}+y_{i2}+y_{i3}, y_{i1}+y_{i2}+y_{i3}\right)$$

$$= \mathbb{V}(y_{i1}) + \mathbb{V}(y_{i2}) + \mathbb{V}(y_{i3}) + 2\,\mathbb{C}(y_{i1}, y_{i2})$$

$$+ 2\,\mathbb{C}(y_{i1}, y_{i3}) + 2\,\mathbb{C}(y_{i2}, y_{i3})$$

$$= 3 + 3 \times 2 \times \frac{1}{2} = 6$$ puis on termine.

2] Soit $\tilde{\alpha}$ l'estimateur des moindres carrés généralisés de α

Nous avons

$$\tilde{\alpha} = (X^T \Omega^{-1} X)^{-1} X^T \Omega^{-1} y$$

où $\Omega = \mathbb{V}(y) = \mathbb{V}(\varepsilon) = \mathbb{E}(\varepsilon \varepsilon^T)$

Pour cet exemple nous avons

$$\mathbb{V}\left(\left(y_{i,1}, y_{i,2}, y_{i,3}\right)\right) = \begin{bmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{bmatrix}$$

$= \frac{1}{2} I_3 + \frac{1}{2} J_3$ avec

I_3 matrice identité de dimension 3×3

J_3 matrice constante dont tous les termes sont égaux à 1 de dimension 3×3

Nous avons donc

$$\Omega = \begin{bmatrix} \frac{1}{2}I_3 + \frac{1}{2}J_3 & 0 & 0 & \cdots \\ 0 & \frac{1}{2}I_3 + \frac{1}{2}J_3 & 0 & \\ 0 & 0 & \frac{1}{2}I_3 + \frac{1}{2}J_3 & 0 \\ \vdots & 0 & 0 & \ddots \end{bmatrix}$$

Nous pouvons maintenant écrire le code R, nous supposons que l'objet R y contient le vecteur $(y_{11}, y_{12}, y_{13}, ---, y_{10,3})$ et l'objet R x contient le vecteur $(x_1, x_1, x_1, ---, x_{10}, x_{10}, x_{10})$

Code 1

```
H <- matrix(rep(1/2, 9), 3) + diag(rep(1/2), 3)
OMEGA <- bdiag(H, H, H, H, H, H, H, H, H, H)
#( fonction bdiag de la
#  bibliothèque Matrix
IOMEGA <- solve(OMEGA)
alphatilde <- solve(t(x)%*%IOMEGA%*%x)
              %*% t(x)%*%IOMEGA%*%y
```

Code 2 si l'on ne connait pas la fonction bdiag de la bibliothèque Matrix

```
H <- matrix(rep(1/2,9),3) + diag(rep(1/2),3)
IH <- solve(H)
z <- rep(0,30); w <- rep(0,30)
for (i in 1:10) {
  z[(1+3(i-1)):(3i)] <- IH %*% x[(1+3(i-1)):(3i)]
  w[(1+3(i-1)):(3i)] <- IH %*% y[(1+3(i-1)):(3i)]
}
alphatilde <- solve(t(x)%*%z)%*%t(x)%*%w
```

L'exactitude du code ci-dessus vient du fait que l'inverse d'une matrice bloc-diagonale et aussi bloc diagonale avec les matrices inverses des blocs en

Exercice 2 ⑥

1)

$$f(y;\pi) = C_{y-1}^{h-1} \pi^{h} (1-\pi)^{y-h} \mathbb{1}_{\{h, h+1, ---\}}(y)$$

$$f(y;\pi) = \exp\left\{ y \log(1-\pi) + h \log\left(\frac{\pi}{1-\pi}\right) \right\} C_{y-1}^{h-1} \mathbb{1}_{\{h, h+1, ----\}}(y)$$

Donc la loi négative binomiale appartient bien à la famille exponentielle avec paramètre de nuisance :

$$\theta = \log(1-\pi) \Rightarrow \pi = 1 - e^{\theta}$$

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

$$\phi = 1$$

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

avec la mesure de référence

$$\nu(y) = C_{y-1}^{h-1} \delta_{\{h, h+1, ----\}}(y)$$

Par ailleurs, nous savons que

$$\mathbb{E}(y) = b'(\theta) = n + \frac{n e^{\theta}}{1-e^{\theta}}$$

et donc $\mathbb{E}(y) = \dfrac{n}{1-e^{\theta}} = \dfrac{n}{\pi}$

D'après l'énoncé, $\log\left(\dfrac{\pi}{1-\pi}\right) = \alpha + \beta x$

et donc $\pi = \left[\dfrac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}\right]$

Ainsi, $\mathbb{E}(y) = \dfrac{\left[1+e^{\alpha+\beta x}\right] n}{e^{\alpha+\beta x}}$

et $\mathbb{E}(y) = r(\alpha + \beta x)$

avec $r(q) = n\left(\dfrac{1+e^{q}}{e^{q}}\right)$

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

Par ailleurs,

$$\Pi(q) = \ln\left[\frac{1+e^q}{e^q}\right] \neq b'(q) = \frac{\ln}{1-e^q}$$

Ce n'est pas le lien canonique qui a été utilisé.

2) $$\mathcal{L}(\alpha,\beta) = \sum_{i=1}^{n} y_i \log\left(\frac{1}{1+e^{\alpha+\beta x_i}}\right) + \ln \sum_{i=1}^{n}\left[\alpha+\beta x_i\right] + cst$$

$$\frac{dL}{d\alpha}(\alpha,\beta) = -\sum_{i=1}^{n} y_i \left[\frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}}\right] + n\ln$$

$$\frac{dL}{d\beta}(\alpha,\beta) = -\sum_{i=1}^{n} y_i x_i \left[\frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}}\right] + \ln\sum_{i=1}^{n} x_i$$

On ne peut pas résoudre

$$\frac{dL}{d\alpha}(\alpha^*,\beta^*) = 0 \text{ et } \frac{dL}{d\beta}(\alpha^*,\beta^*) = 0$$

3]

$$\frac{d^2L}{d\alpha^2}(\alpha,\beta) = -\sum_{i=1}^{m} y_i \left[\frac{e^{\alpha+\beta x_i}\left(1+e^{\alpha+\beta x_i}\right) - e^{\alpha+\beta x_i}\, e^{\alpha+\beta x_i}}{\left(1+e^{\alpha+\beta x_i}\right)^2} \right]$$

$$\frac{d^2L}{(d\alpha)^2}(\alpha,\beta) = -\sum_{i=1}^{m} y_i \frac{e^{\alpha+\beta x_i}}{\left(1+e^{\alpha+\beta x_i}\right)^2}$$

$$\frac{d^2L}{(d\beta)^2}(\alpha,\beta) = -\sum_{i=1}^{m} y_i x_i^2 \frac{e^{\alpha+\beta x_i}}{\left(1+e^{\alpha+\beta x_i}\right)^2}$$

$$\frac{d^2L}{d\alpha\, d\beta}(\alpha,\beta) = -\sum_{i=1}^{m} y_i x_i \frac{e^{\alpha+\beta x_i}}{\left(1+e^{\alpha+\beta x_i}\right)^2}$$

$$I_m(\alpha,\beta) = n \left[\begin{array}{c|c} \displaystyle\sum_{i=1}^{m} \frac{1}{\left(1+e^{\alpha+\beta x_i}\right)} & \displaystyle\sum_{i=1}^{m} \frac{x_i}{\left(1+e^{\alpha+\beta x_i}\right)} \\ \hline \displaystyle\sum_{i=1}^{m} \frac{x_i}{\left(1+e^{\alpha+\beta x_i}\right)} & \displaystyle\sum_{i=1}^{m} \frac{x_i^2}{\left(1+e^{\alpha+\beta x_i}\right)} \end{array} \right]$$

4] Pour le lien canonique

$$\mathbb{E}(y) = \frac{h}{\pi} = \left[\frac{h}{1 - e^{\alpha+\beta x}}\right]$$

Dans ce cas, $\pi = 1 - e^{\alpha+\beta x}$

$$L(\alpha,\beta) = \sum_{i=1}^{n} y_i \left[\alpha + \beta x_i\right]$$

$$+ h \sum_{i=1}^{n} \log\left[\frac{1 - e^{\alpha+\beta x_i}}{e^{\alpha+\beta x_i}}\right]$$

$$+ \text{cst}$$

$$= \alpha \sum_{i=1}^{n} y_i + \beta \sum_{i=1}^{n} y_i x_i - hn\alpha$$

$$- h\beta \sum_{i=1}^{n} x_i + h \sum_{i=1}^{n} \log\left(1 - e^{\alpha+\beta x_i}\right)$$

$$+ \text{cst}$$

$$\frac{dL}{d\alpha}(\alpha,\beta) = \sum y_i - nh - h\sum_{i=1}^{n} \frac{e^{\alpha+\beta x_i}}{1-e^{\alpha+\beta x_i}}$$

$$\frac{dL}{d\beta}(\alpha,\beta) = \sum_{i=1}^{n} y_i x_i - h\sum_{i=1}^{n} x_i - h\sum_{i=1}^{n} x_i \left(\frac{e^{\alpha+\beta x_i}}{1-e^{\alpha+\beta x_i}}\right)$$

Exercice 3

1] Il s'agit de six modèles de régression logistique

model 1 : le risque de crédit expliqué par les données.

model 2 : le risque de crédit expliqué par toutes les variables

model 3 : le risque de crédit expliqué par les variables les plus pertinentes au sens du critère AIC et méthode descendante model 4 : idem que model 3 pour le critère BIC

model 5 : le risque de crédit expliqué uniquement par la constante (model null)

model 6 : le risque de crédit expliqué par les variables les plus pertinentes au sens du critère AIC et méthode progressive.

2) Nous donnons le code pour le model 1, il s'applique à tous les autres en changeant model 1 par model i.

```
library(caret)
train(formula(model 1), data = GermanCredit,
      method = "glm", family = "binomial",
      metric = "Accuracy", trControl = trainControl(repeats=10,
                                    method="repeatedcv", number=2))
```