Université de Montpellier Année 2015-2016
HMMA304 - Modèles linéaires généralisés

Examen final - 17 décembre 2015

Durée 2h - Documents interdits

Exercice 1 (5 pts)

Nous nous intéressons ici à l'étude de la pièce ayant causé l'explosion de la navette spatiale Challenger en 1986. L'étanchéité du moteur de la navette spatiale est assurée par six pièces identiques appelées "O-ring". L'explosion de la navette Challenger est due à la défaillance d'au moins l'une de ces pièces.

Au cours des 24 vols précédents d'une navette spatiale, nous disposons des données suivantes : la variable `temp` qui correspond à la température au moment du lancement et la variable `defa` qui vaut 0 si aucun des "O-ring" n'a été endommagé au cours du lancement et 1 si au moins l'un d'entre eux a été endommagé.

Proposer une méthode permettant d'estimer la probabilité de défaillance d'au moins un "O-ring" pour une température de 31 degrés Fahrenheit (température au moment du lancement de la navette challenger). Nous supposons que les données sont stockées dans un data.frame R nommé `challenger`, donner le code R associé.

Exercice 2 (7 pts)

On considère n variables aléatoires indépendantes $Y_1, \dots, Y_n$ telles que $Y_i \sim \mathcal{N}\left(\exp(\alpha + \beta x_i), \sigma^2\right)$.

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.

2 (2 pts) On suppose dans la suite que $\sigma^2 = 1$, 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 $(Y_1, \dots, Y_n)$ sur les paramètres α et β notée $I_n(\alpha, \beta)$.

Exercice 3 (8 pts)

Commenter en détails le fichier R Markdown fourni en annexes. Il s'agit d'une étude sur les facteurs influençant la présence de ruissellement (runoff) lors de tempêtes.

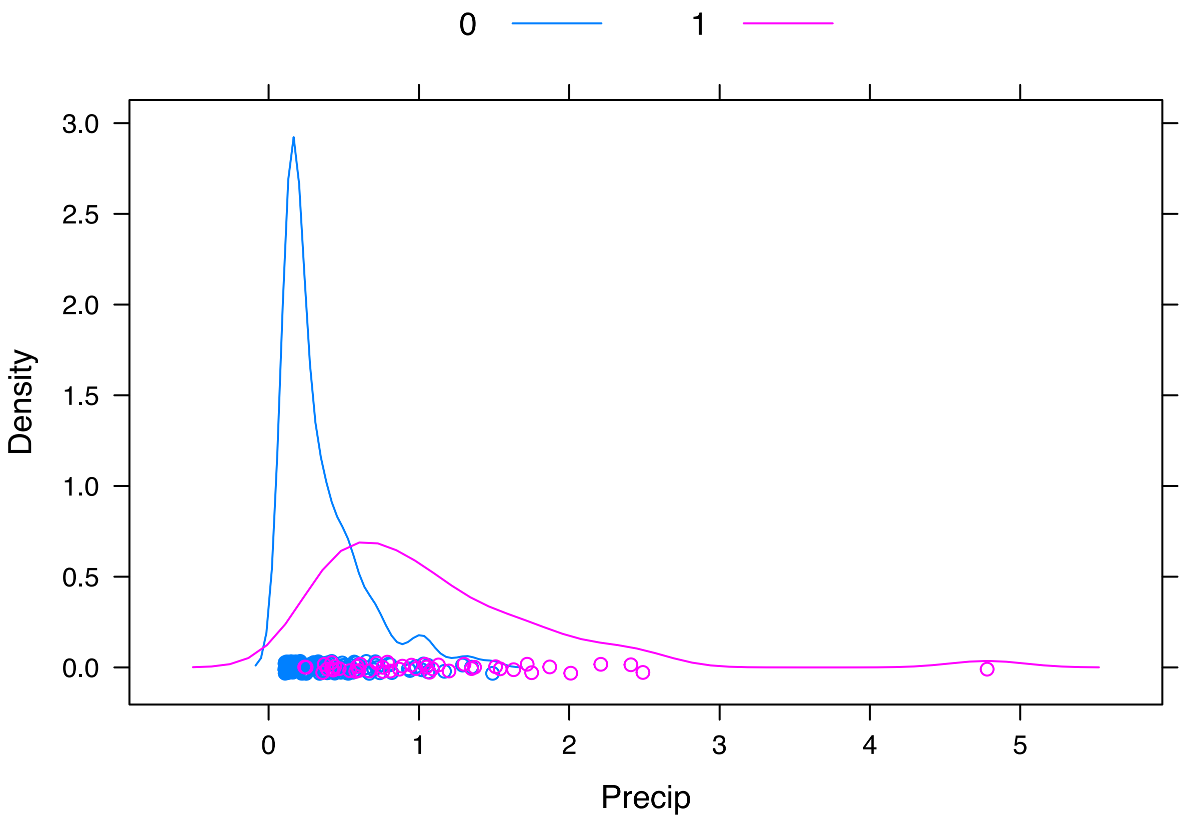
```
Data collected over a 4-year period from a Madison home.
Outcome: indicator if a rain storm produces runoff.
Multiple predictors.
```

Annexes Examen HMMA304

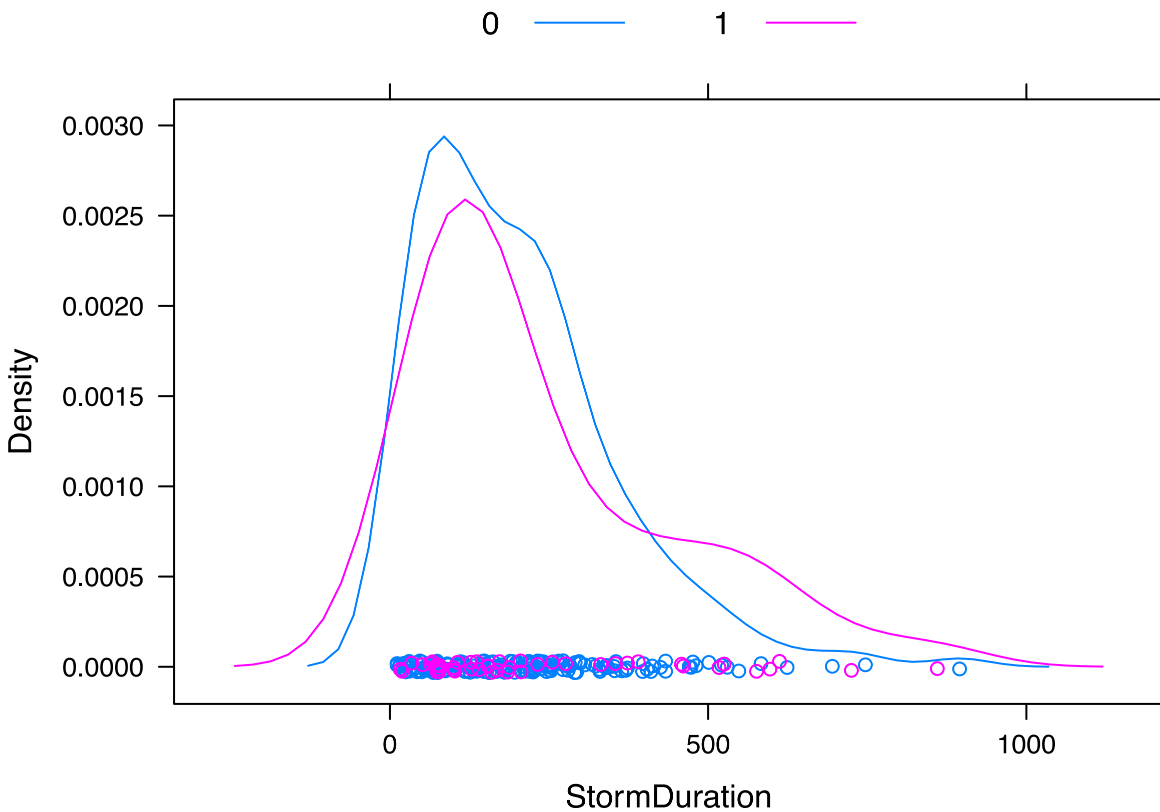
Jean-Michel Marin

17 décembre 2015

```
runoff <- read.table("runoff.txt",header=TRUE)
library(lattice)
densityplot(~Precip,groups=factor(RunoffEvent),
data=runoff,auto.key=list(columns=2))
```




```
densityplot(~StormDuration,groups=factor(RunoffEvent),
data=runoff,auto.key=list(columns=2))
```




```
model1 <- glm(RunoffEvent~Precip,data=runoff,family=binomial)
summary(model1)
```

```
##
## Call:
## glm(formula = RunoffEvent ~ Precip, family = binomial, data = runoff)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0749  -0.4512  -0.3184  -0.2821   2.3629
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -3.6418     0.4152  -8.771  < 2e-16 ***
## Precip        3.8059     0.5801   6.560 5.37e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 227.82  on 230  degrees of freedom
## Residual deviance: 148.13  on 229  degrees of freedom
## AIC: 152.13
##
## Number of Fisher Scoring iterations: 5
```

```r
model2 <-  glm(RunoffEvent~StormDuration,data=runoff,family=binomial)
summary(model2)
```

```
##
## Call:
## glm(formula = RunoffEvent ~ StormDuration, family = binomial,
##     data = runoff)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9721  -0.6704  -0.6199  -0.5845   1.9296
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -1.7143475  0.2718360  -6.307 2.85e-10 ***
## StormDuration  0.0013520  0.0009357   1.445    0.148
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 227.82  on 230  degrees of freedom
## Residual deviance: 225.81  on 229  degrees of freedom
## AIC: 229.81
##
## Number of Fisher Scoring iterations: 4
```

```r
model3 <-  glm(RunoffEvent~StormDuration+LastStorm+Precip+
        MaxIntensity60+EI,data=runoff, family=binomial)
summary(model3)
```

```
##
## Call:
## glm(formula = RunoffEvent ~ StormDuration + LastStorm + Precip +
##     MaxIntensity60 + EI, family = binomial, data = runoff)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.20621  -0.38734  -0.23867  -0.05348   2.83316
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -3.821e+00  9.623e-01  -3.970 7.18e-05 ***
## StormDuration  -9.522e-04  3.347e-03  -0.285   0.7760
## LastStorm      -1.299e-04  5.397e-05  -2.408   0.0161 *
## Precip          3.210e+00  2.198e+00   1.461   0.1441
## MaxIntensity60  4.292e+00  3.714e+00   1.156   0.2477
## EI             -3.107e-02  1.896e-01  -0.164   0.8698
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 200.58  on 194  degrees of freedom
## Residual deviance: 100.83  on 189  degrees of freedom
##   (36 observations deleted due to missingness)
## AIC: 112.83
##
## Number of Fisher Scoring iterations: 7
```

```r
model <- step(model3,direction="backward")
```

```
## Start:  AIC=112.83
## RunoffEvent ~ StormDuration + LastStorm + Precip + MaxIntensity60 +
##     EI
##
##                  Df Deviance    AIC
## - EI              1   100.86 110.86
## - StormDuration   1   100.92 110.92
## - MaxIntensity60  1   102.09 112.09
## <none>                100.83 112.83
## - Precip          1   102.95 112.95
## - LastStorm       1   110.82 120.82
##
## Step:  AIC=110.86
## RunoffEvent ~ StormDuration + LastStorm + Precip + MaxIntensity60
##
##                  Df Deviance    AIC
## - StormDuration   1   100.93 108.93
## - MaxIntensity60  1   102.61 110.61
## <none>                100.86 110.86
## - Precip          1   103.13 111.13
## - LastStorm       1   110.82 118.82
##
## Step:  AIC=108.93
## RunoffEvent ~ LastStorm + Precip + MaxIntensity60
##
##                  Df Deviance    AIC
## <none>                100.93 108.93
## - Precip          1   108.13 114.13
## - MaxIntensity60  1   110.73 116.73
## - LastStorm       1   110.90 116.90
```

```r
summary(model)
```

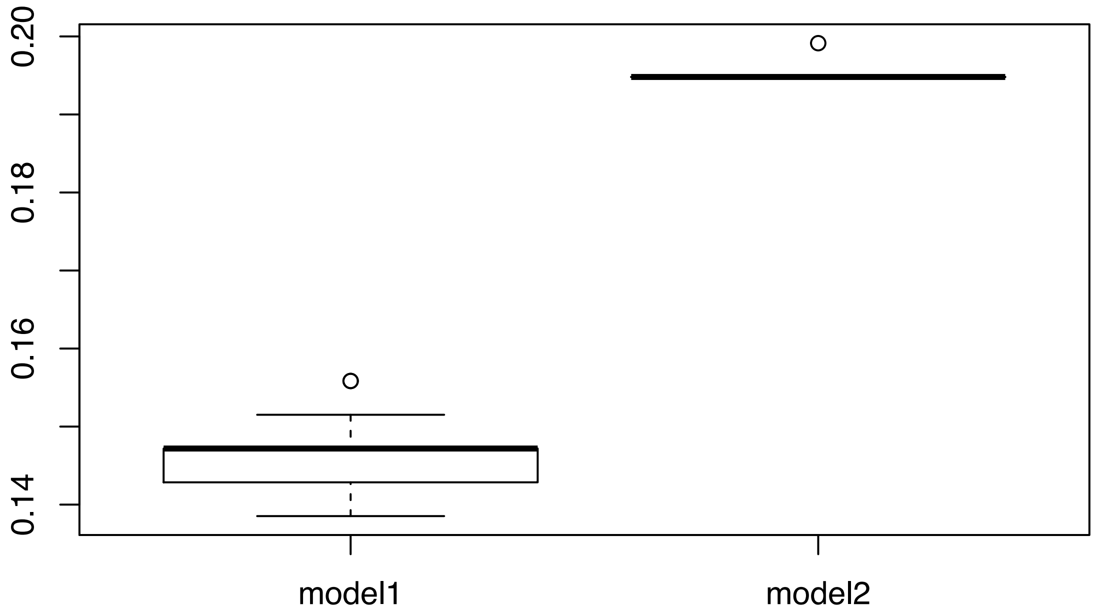
```
##
## Call:
## glm(formula = RunoffEvent ~ LastStorm + Precip + MaxIntensity60,
##     family = binomial, data = runoff)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.21622  -0.39232  -0.24822  -0.05326   2.78648
##
## Coefficients:
```

```
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -3.902e+00  6.259e-01  -6.234 4.53e-10 ***
## LastStorm      -1.291e-04  5.389e-05  -2.395  0.01660 *
## Precip          2.609e+00  9.773e-01   2.670  0.00759 **
## MaxIntensity60  4.623e+00  1.593e+00   2.902  0.00371 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 200.58  on 194  degrees of freedom
## Residual deviance: 100.93  on 191  degrees of freedom
##   (36 observations deleted due to missingness)
## AIC: 108.93
##
## Number of Fisher Scoring iterations: 7
```

```r
library(boot)
```

```
##
## Attaching package: 'boot'
##
## The following object is masked from 'package:lattice':
##
##     melanoma
```

```r
cost <- function(true,pred,a=1,b=1)
{
    res <- true
    res[true==1 & pred > a/(a+b)] <- 0
    res[true==1 & pred <= a/(a+b)] <- b
    res[true==0 & pred <= a/(a+b)] <- 0
    res[true==0 & pred > a/(a+b)] <- a
    mean(res)
}
resu <- matrix(0,100,2)
for (i in 1:100)
{
    resu[i,1] <- cv.glm(runoff,model1,cost,K=10)$delta[1]
    resu[i,2] <- cv.glm(runoff,model2,cost,K=10)$delta[1]
}
boxplot(resu,names=c("model1","model2"))
```


0.20
0.18
0.16
0.14
model1
model2

Correction HMMA304
Examen final - 17/12/2015

1

Exercice 1

La variable à expliquer est un facteur à deux modalités. Nous allons utiliser le modèle linéaire généralisé de la régression logistique : $Y \in \{0,1\}$

$$\mathbb{P}(Y=1 \mid X=x) = \frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)}$$

où x est la température au moment du lancement.

0 = "pas de défaillance"
1 = "défaillance"

②

Nous disposons d'un échantillon de taille $n = 24$ $(y_1, x_1), \dots, (y_n, x_n)$.
Nous l'utilisons pour estimer α et β puis nous prédisons la probabilité de défaillance pour $x = 31$ par

$$\widehat{\mathbb{P}(Y=1 \mid x=31)} = \left[\frac{\exp(\hat{\alpha} + 31\hat{\beta})}{1+\exp(\hat{\alpha}+31\hat{\beta})}\right].$$

où $\hat{\alpha}$ et $\hat{\beta}$ sont des estimations de α et β.

```
model <- glm(y ~ x, data = challenger,
             family = binomial)
topred = data.frame(x = 31)
predict(model, data=topred, type=
        "response")
```

Exercice 2

1] $f(y;\alpha,\beta,\sigma^2) = (2\pi\sigma^2)^{-1/2} e^{-\frac{1}{2\sigma^2}(y-e^{\alpha+\beta x})^2}$

$$\Leftrightarrow f(y;\alpha,\beta,\sigma^2) = \exp\left\{ \frac{y e^{\alpha+\beta x} - e^{2(\alpha+\beta x)}/2}{\sigma^2} + \left[-\frac{1}{2}\log(2\pi\sigma^2) - \frac{y^2}{2\sigma^2}\right]\right\}$$

Nous obtenons ainsi que

$\theta = e^{\alpha+\beta x}$, $\varphi = \sigma^2$

$b(\theta) = \frac{\theta^2}{2}$

Par ailleurs, $E(y) = e^{\alpha+\beta x}$

$= \exp(\alpha+\beta x)$

et donc $r(z) = \exp(z)$.

Il s'agit bien d'un modèle linéaire généralisé associé à la famille gaussienne avec lien log (lien inverse de $r(.)$).

2] Le lien canonique correspond au cas où $\pi(z) = b'(z)$ (1)

Nous avons $b'(z) = z \neq \exp(z)$.

Donc ce n'est pas le lien canonique qui a été utilisé ici.

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

$$\frac{d\mathcal{L}}{d\alpha}(\alpha,\beta) = \sum_{i=1}^{n}\left(y_i - e^{\alpha+\beta x_i}\right)e^{\alpha+\beta x_i} = 0$$

$$\frac{d\mathcal{L}}{d\beta}(\alpha,\beta) = \sum_{i=1}^{n} x_i\left(y_i - e^{\alpha+\beta x_i}\right)e^{\alpha+\beta x_i} = 0$$

On ne peut pas résoudre analytiquement ce système.

4]

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

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

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

Comme $\mathbb{E}(y_i) = e^{\alpha+\beta x_i}$, nous obtenons

$$I_n(\alpha,\beta) = \begin{bmatrix} \sum_{i=1}^{n}\left(e^{\alpha+\beta x_i}\right)^2 & \sum_{i=1}^{n}x_i\left(e^{\alpha+\beta x_i}\right)^2 \\ \sum_{i=1}^{n}x_i\left(e^{\alpha+\beta x_i}\right)^2 & \sum_{i=1}^{n}x_i^2\left(e^{\alpha+\beta x_i}\right)^2 \end{bmatrix}$$

Exercice 3

Les deux premiers graphiques d'estimation non paramétrique de densités par classe de la variable à expliquer (variable binaire) montrent qu'un modèle basé sur les précipitations sera plus pertinent qu'un modèle basé sur la durée de la tempête.

Cela se confirme par la mise en œuvre de modèles de régression logistique. Le critère AIC à minimiser avec R est plus bas pour le modèle ayant pour variable explicative "Precip".

Par une méthode descendante avec le critère AIC on retient les variables : LotStorm, Precip et MaxIntensity60.

Par validation croisée à 10 ensembles, on évalue l'erreur de classification associée aux modèles logistique contenant la variable Precip seulement et la variable StormDuration seulement. Le taux d'erreur médian, pour 100 réplications Monte-Carlo est plus faible avec la variable Precip.