Quelques modèles linéaires généralisés

Cristian Preda

15/03/2021

Les modèles de régression

Soient

- Y la variable **réponse** et
- $X = (X_1, \ldots, X_p)$ les variables **explicatives**

Un modèle de régression exprime le lien entre l'espérance conditionnelles de Y sachant $X = x$, $x = (x_1, \ldots, x_p)$.

$$\mathbb{E}(Y|X = x) = f(x),$$

$f : \mathbb{R}^p \to \mathbb{R}$, fonction de régression

Quelques modèles classiques

- Le modèle logistique binaire

$$Y \in \{0, 1\}$$

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

$$\mathbb{P}(Y = 1|X = x) = \frac{e^{\beta_0+\beta_1 x_1+\ldots\beta_p X_p}}{1 + e^{\beta_0+\beta_1 x_1+\ldots\beta_p X_p}},$$

ou

$$\log\left(\frac{\mathbb{P}(Y = 1|X = x)}{\mathbb{P}(Y = 0|X = x)}\right) = \beta_0 + \beta_1 x_1 + \ldots \beta_p x_p.$$

Quelques modèles classiques (cont.)

- Le modèle logistique binaire

Exemple:

```
d = read.table("http://math.univ-lille1.fr/~preda/GIS5/data_roc.txt", sep="\t", header=TRUE)
head(d,3)
```

```
  AGE DIAB PREMATURE
1  26    0         1
2  25    0         1
3  28    0         1
```

```
knitr::kable(table(d$PREMATURE), col.names = c("PREM", "proba"), align ="c")
```

PREM	proba
0	124
1	266

```
knitr::kable(prop.table(table(d$PREMATURE)), col.names = c("PREM", "proba"), align ="c")
```

PREM	proba
0	0.3179487
1	0.6820513

Quelques modèles classiques (cont.)

- Le modèle logistique binaire

Exemple (cont.)

```
m = glm(PREMATURE~AGE+as.factor(DIAB), family = "binomial", data = d)
summary(m)
```

```
Call:
glm(formula = PREMATURE ~ AGE + as.factor(DIAB), family = "binomial",
    data = d)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7916  -1.4227   0.8144   0.8883   1.0990

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       1.84558    0.57924   3.186  0.00144 **
AGE              -0.04146    0.02135  -1.942  0.05218 .
as.factor(DIAB)1  0.57163    0.67296   0.849  0.39565
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 487.74  on 389  degrees of freedom
Residual deviance: 483.45  on 387  degrees of freedom
AIC: 489.45

Number of Fisher Scoring iterations: 4
```

Quelques modèles classiques (cont.)

- Le modèle logistique binaire : qualité d'ajustement

Exemple (cont.)

McFadden's R^2 ou quantité de "deveiance expliquée :

$$R^2 = 1 - \frac{LL(m)}{LL(m_{null})}$$

```
m_null = glm(PREMATURE~1, family = "binomial", data = d)
r2 = 1 - logLik(m)/logLik(m_null)
print(r2)
```

```
'log Lik.' 0.00881493 (df=3)
```

```
print((m$null.deviance - m$deviance)/m$null.deviance)
```

```
[1] 0.00881493
```

Le modèle semble assez mauvais! Un bon modèle à un $R^2 > 0.1$ (voir plus: http://eml.berkeley.edu/~mcfadden/travel.html)

Quelques modèles classiques (cont.)

- Le modèle logistique binaire : qualité d'ajustement

Exemple (cont.)

Test de Hosmer-Lemeshow (goodness of fit):

```
library(ResourceSelection)
```

```
## ResourceSelection 0.3-5   2019-07-22
```

```
hoslem.test(m$y, fitted(m))
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  m$y, fitted(m)
## X-squared = 5.2652, df = 8, p-value = 0.7289
```

Quelques modèles classiques (cont.)

- Le modèle logistique binaire : qualité d'ajustement

La courbe ROC et AUC. (vu en cours de classification supervisée).

Voir aussi l'exemple dans :

http://math.univ-lille1.fr/~preda/GIS5/exemple_roc.R

Quelques modèles classiques (cont.)

- Le modèle logistique binaire

Un autre exemple :

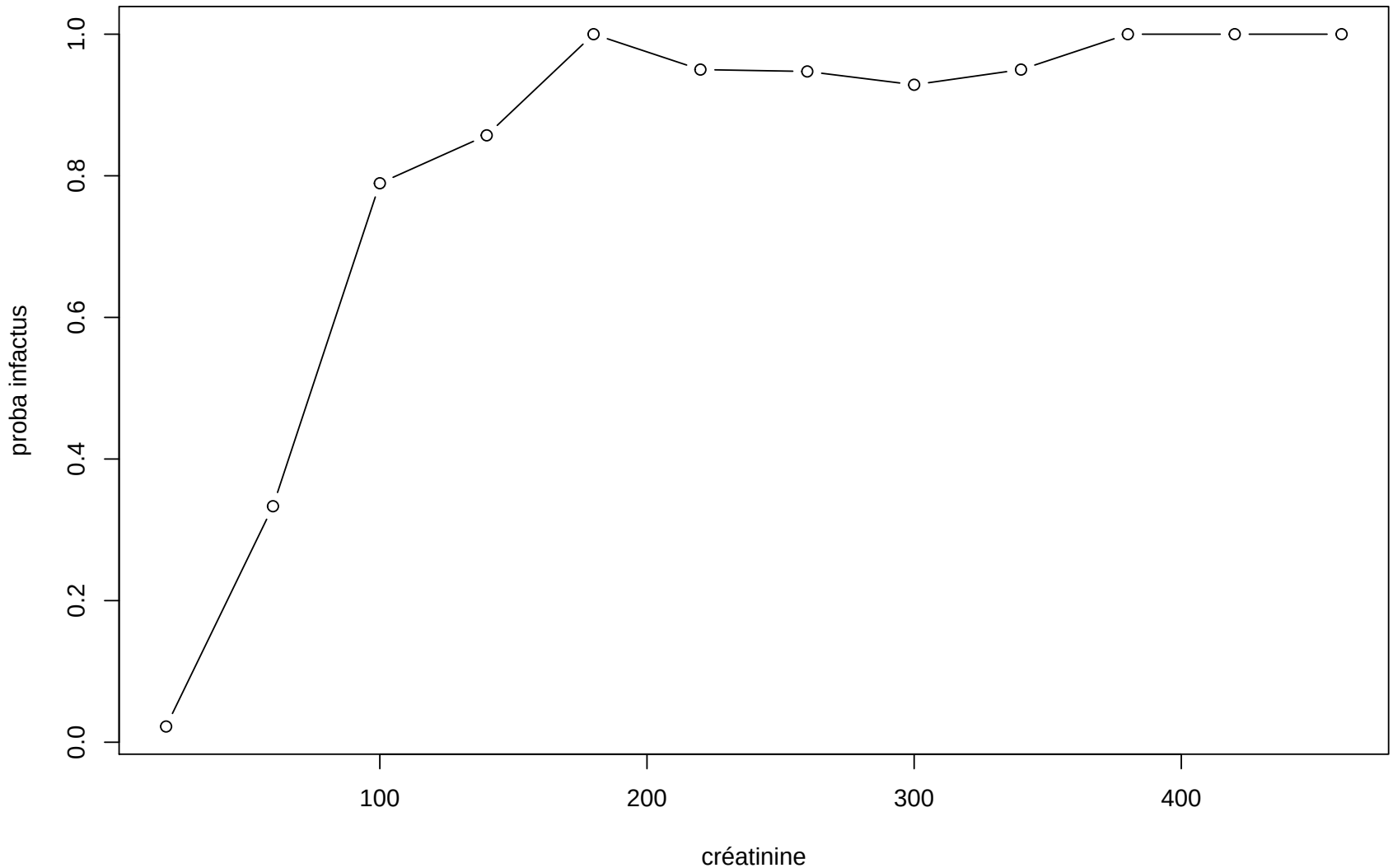
Les données (infarctus): un bon prédicteur pourrait etre l'enzyme créatinine-Kinase (CK)

CK	nb_pat_infarctus	nb_pat_sans_infarctus
20	2	88
60	13	26
100	30	8
140	30	5
180	21	0
220	19	1
260	18	1
300	13	1
340	19	1
380	15	0
420	7	0
460	8	0

Quelques modèles classiques (cont.)

- Le modèle logistique binaire

```
p = d$nb_pat_infarctus/(d$nb_pat_infarctus+d$nb_pat_sans_infarctus)
plot(d$CK, p, xlab = "créatinine", ylab = "proba infactus", type="b")
```



Quelques modèles classiques (cont.)

- Le modèle logistique binaire

```
m = glm(cbind(nb_pat_infarctus, nb_pat_sans_infarctus)~CK, family =binomial, data=d)
summary(m)
```

```
Call:
glm(formula = cbind(nb_pat_infarctus, nb_pat_sans_infarctus) ~
    CK, family = binomial, data = d)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-3.08184  -1.93008   0.01652   0.41772   2.60362

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.758358   0.336696  -8.192 2.56e-16 ***
CK           0.031244   0.003619   8.633  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 271.712  on 11  degrees of freedom
Residual deviance:  36.929  on 10  degrees of freedom
AIC: 62.334

Number of Fisher Scoring iterations: 6
```

Quelques modèles classiques

- Régression logistique multinomiale

$$Y \in \{1, 2, \ldots, K\}, \quad K \geq 2.$$

On choisit une modalité de référence : 1

$$\log\left(\frac{\mathbb{P}(Y = i|X = x)}{\mathbb{P}(Y = 1|X = x)}\right) = \beta_{i0} + \beta_{i1}x_1 + \ldots + \beta_{ip}x_p, \quad i = 2, ..., K$$

Quelques modèles classiques

- Régression logistique multinomiale

Exemple : évalution des vins

```
library(ordinal)
library(nnet)
data("wine", package="ordinal")
head(wine)
```

```
  response rating temp contact bottle judge
1       36      2 cold      no      1     1
2       48      3 cold      no      2     1
3       47      3 cold     yes      3     1
4       67      4 cold     yes      4     1
5       77      4 warm      no      5     1
6       60      4 warm      no      6     1
```

Quelques modèles classiques

- Régression logistique multinomiale

Exemple : évalution des vins

```
m =multinom(rating~temp, data = wine)
```

```
# weights:  15 (8 variable)
initial  value 115.879530
iter  10 value 90.682525
final  value 90.134520
converged
```

```
summary(m)
```

```
Call:
multinom(formula = rating ~ temp, data = wine)

Coefficients:
  (Intercept)  tempwarm
2   1.1630698  7.147386
3   0.9552492  8.128180
4  -0.9165646  9.738184
5  -8.4736298 16.937776

Std. Errors:
  (Intercept) tempwarm
2   0.5123225 26.04038
3   0.5262232 26.03893
4   0.8366930 26.04750
5  30.9428493 40.43853

Residual Deviance: 180.269
```

Quelques modèles classiques

- Régression logistique multinomiale **ordinale**

$$Y \in \{1, 2, \ldots, K\}, \quad K \geq 2.$$

$$1 < 2 < \ldots < K$$

$$\log\left(\frac{\mathbb{P}(Y \leq j | X = x)}{\mathbb{P}(Y > j | X = x)}\right) = \theta_j - \beta_1 x_1 - \ldots - \beta_p x_p, \quad j = 1, \ldots K-1.$$

Interprétation de β_i :

$$exp(-\beta_i) = \frac{\frac{\mathbb{P}(Y \leq j | X_i = x_i + 1)}{\mathbb{P}(Y > j | X_i = x_i + 1)}}{\frac{\mathbb{P}(Y \leq j | X_i = x_i)}{\mathbb{P}(Y > j | X_i = x_i)}}$$

}

https://cran.r-project.org/web/packages/ordinal/vignettes/clm_article.pdf

Quelques modèles classiques

- Régression de Poisson

Y = variable quantitative à valeurs dans $\{0, 1, \ldots\}$

$$Y \sim \mathcal{P}(\lambda)$$

$$P(Y = k) = e^{-\lambda} \frac{\lambda^k}{k!}.$$

$$E(Y) = Var(Y) = \lambda.$$

Modèle de Poisson :

$$\log(\lambda | X = x) = \beta_0 + \beta_1 X1 + \ldots + \beta_p X_p.$$

Exemple :

https://stats.idre.ucla.edu/r/dae/poisson-regression/