

## Un exemple de régression logistique sous

A.B. Dufour & A. Viallefont

Etude de l'apparition ou non d'une maladie cardiaque des coronaires

### 1 Présentation des données

Les données d'étude sont accessibles à l'adresse suivante : <http://www.sph.emory.edu/~dkleinb/logreg2.htm#data>.

Elles contiennent les informations sur une cohorte de 609 hommes ayant été suivis sur une période de 7 ans. Il s'agit d'étudier la variable d'intérêt "apparition ou non d'une maladie cardiaque des coronaires"<sup>1</sup>.

```
evans <- read.table("http://www.sph.emory.edu/~dkleinb/datasets/evans.dat")
head(evans)
  V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12
1 21 0 0 56 270 0 0 80 138 0 0 0
2 31 0 0 43 159 1 0 74 128 0 0 0
3 51 1 1 56 201 1 1 112 164 1 1 201
4 71 0 1 64 179 1 0 100 200 1 1 179
5 74 0 0 49 243 1 0 82 145 0 0 0
6 91 0 0 46 252 1 0 88 142 0 0 0

names(evans) <- c("id", "chd", "cat", "age", "chl", "smk", "ecg",
                 "dbp", "sbp", "hpt", "ch", "cc")
head(evans)
  id chd cat age chl smk ecg dbp sbp hpt ch cc
1 21 0 0 56 270 0 0 80 138 0 0 0
2 31 0 0 43 159 1 0 74 128 0 0 0
3 51 1 1 56 201 1 1 112 164 1 1 201
4 71 0 1 64 179 1 0 100 200 1 1 179
5 74 0 0 49 243 1 0 82 145 0 0 0
6 91 0 0 46 252 1 0 88 142 0 0 0
```

Les variables du data frame sont définies ci-dessous.

**id** identifiant du sujet. Chaque observation a un identifiant unique soit une observation par individu.

**chd** une variable dichotomique prenant la valeur 1 si la maladie est présente, 0 sinon.

**cat** une variable dichotomique indiquant si le niveau de catecholamine est élevé (1) ou non (0).

**age** une variable continue exprimée en années.

**chl** une variable continue définissant le taux de cholestérol.

<sup>1</sup>Kleinbaum et Klein (2002). *Logistic Regression. A self-Learning Text*, Springer Edition

**smk** une variable dichotomique indiquant si le sujet est fumeur (1) ou s'il n'a jamais fumé (0).  
**ecg** une variable dichotomique indiquant la présence d'un électrocardiogramme anormal (1) ou non (0).  
**dbp** une variable continue indiquant la pression artérielle diastolique.  
**sbp** une variable continue indiquant la pression artérielle systolique.  
**hpt** une variable dichotomique indiquant la présence (1) ou non (0) d'une forte pression sanguine.  
**ch** une variable construite à partir du produit  $cat \times hpt$ .  
**cc** une variable construite à partir du produit  $cat \times chl$ .

L'objectif est de discuter le modèle logistique. Soit **chd** la présence (1) ou l'absence (0) de la maladie coronarienne avec comme variable d'exposition le niveau de catecholamine **cat**. **age**, **chl**, **smk**, **ecg** et **hpt** sont les variables de contrôle.

## 2 Quelques statistiques de base

```
summary(evans)
  id          chd          cat          age          chl
Min.   : 21   Min.   :0.0000   Min.   :0.0000   Min.   :40.00   Min.   : 94.0
1st Qu.: 4242 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:46.00   1st Qu.:184.0
Median : 9751 Median :0.0000   Median :0.0000   Median :52.00   Median :209.0
Mean   : 9213 Mean   :0.1166   Mean   :0.2003   Mean   :53.71   Mean   :211.7
3rd Qu.:13941 3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:60.00   3rd Qu.:234.0
Max.   :19161 Max.   :1.0000   Max.   :1.0000   Max.   :76.00   Max.   :357.0

  smk          ecg          dbp          sbp          hpt
Min.   :0.0000   Min.   :0.0000   Min.   : 60.00   Min.   : 92.0    Min.   :0.0000
1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.: 80.00   1st Qu.:125.0   1st Qu.:0.0000
Median :1.0000   Median :0.0000   Median : 90.00   Median :140.0   Median :0.0000
Mean   :0.6355   Mean   :0.2726   Mean   : 91.18   Mean   :145.5   Mean   :0.4187
3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:100.00  3rd Qu.:160.0   3rd Qu.:1.0000
Max.   :1.0000   Max.   :1.0000   Max.   :170.00  Max.   :300.0   Max.   :1.0000

  ch          cc
Min.   :0.0000   Min.   : 0.00
1st Qu.:0.0000   1st Qu.: 0.00
Median :0.0000   Median : 0.00
Mean   :0.1609   Mean   :39.96
3rd Qu.:0.0000   3rd Qu.: 0.00
Max.   :1.0000   Max.   :331.00
```

Si l'on fait le résumé du data frame, on s'aperçoit que les données dichotomiques ne sont pas traitées comme des variables qualitatives.

```
transvar <- c(2, 3, 6, 7, 10)
for (i in 1:5) {
  evans[, transvar[i]] <- factor(evans[, transvar[i]])
}
```

On obtient alors le résumé statistique suivant :

```
summary(evans)
  id          chd          cat          age          chl          smk          ecg
Min.   : 21   0:538   0:487   Min.   :40.00   Min.   : 94.0   0:222   0:443
1st Qu.: 4242 1: 71   1:122   1st Qu.:46.00   1st Qu.:184.0   1:387   1:166
Median : 9751 Median : Median :52.00   Median :209.0
Mean   : 9213 Mean   : Mean   :53.71   Mean   :211.7
3rd Qu.:13941 3rd Qu.:3rd Qu.:60.00   3rd Qu.:234.0
Max.   :19161 Max.   :76.00   Max.   :357.0

  dbp          sbp          hpt          ch          cc
Min.   : 60.00   Min.   : 92.0   0:354   Min.   :0.0000   Min.   : 0.00
1st Qu.: 80.00   1st Qu.:125.0   1:255   1st Qu.:0.0000   1st Qu.: 0.00
Median : 90.00   Median :140.0   Median :0.0000   Median : 0.00
Mean   : 91.18   Mean   :145.5   Mean   :0.1609   Mean   :39.96
3rd Qu.:100.00  3rd Qu.:160.0   3rd Qu.:0.0000   3rd Qu.: 0.00
Max.   :170.00  Max.   :300.0   Max.   :1.0000   Max.   :331.00
```

Les données sont prêtes à être analysées.

### 3 Modèle sans interaction

```
options(show.signif.stars = FALSE)
attach(evans)
glm1 <- glm(chd ~ age + chl + smk + ecg + hpt + cat, family = "binomial")
summary(glm1)

Call:
glm(formula = chd ~ age + chl + smk + ecg + hpt + cat, family = "binomial")
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1388  -0.5284  -0.4043  -0.2944   2.5594

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.774707    1.140224  -5.942 2.82e-09
age          0.032228    0.015176   2.124 0.03370
chl          0.008746    0.003262   2.681 0.00735
smk1         0.834792    0.305222   2.735 0.00624
ecg1         0.369545    0.293635   1.259 0.20820
hpt1         0.439177    0.290814   1.510 0.13100
cat1         0.597780    0.351979   1.698 0.08944

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 438.56  on 608  degrees of freedom
Residual deviance: 400.39  on 602  degrees of freedom
AIC: 414.39

Number of Fisher Scoring iterations: 5

anova(glm1, test = "Chisq")

Analysis of Deviance Table
Model: binomial, link: logit

Response: chd

Terms added sequentially (first to last)

      Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                                608      438.56
age  1    11.33      607      427.22 0.000761
chl  1     5.20      606      422.02 0.02
smk  1     8.85      605      413.17 0.002932
ecg  1     5.16      604      408.01 0.02
hpt  1     4.76      603      403.26 0.03
cat  1     2.86      602      400.39 0.09
```

#### Exercice.

- 1) Commenter les résultats obtenus dans le `summary`.
- 2) Discuter de la pertinence des résultats obtenus dans `anova`. Pour cela, changer l'ordre d'entrée des variables de contrôle.

### 4 Modèle avec interaction

La variable `cat` est une variable d'exposition qu'il est intéressant de croiser avec d'une part le cholestérol et d'autre part le statut d'hypertension ou non du sujet.

```
glm2 <- glm(chd ~ age + smk + ecg + (chl * cat) + (hpt * cat), family = "binomial")
summary(glm2)

Call:
glm(formula = chd ~ age + smk + ecg + (chl * cat) + (hpt * cat),
    family = "binomial")
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0999  -0.4492  -0.3555  -0.2493   3.3095

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
```

```
(Intercept) -4.049738  1.255014 -3.227  0.00125
age          0.034963  0.016138  2.166  0.03028
smk1        0.773214  0.327267  2.363  0.01815
ecg1        0.367131  0.327803  1.120  0.26273
chl         -0.005455  0.004184 -1.304  0.19228
cat1       -12.689531  3.104650 -4.087  4.36e-05
hpt1        1.046649  0.331635  3.156  0.00160
chl:cat1    0.069170  0.014360  4.817  1.46e-06
cat1:hpt1   -2.331785  0.742668 -3.140  0.00169
```

(Dispersion parameter for binomial family taken to be 1)

```
Null deviance: 438.56  on 608  degrees of freedom
Residual deviance: 347.23  on 600  degrees of freedom
AIC: 365.23
```

Number of Fisher Scoring iterations: 6

```
anova(glm2, test = "Chisq")
```

Analysis of Deviance Table  
Model: binomial, link: logit

Response: chd

Terms added sequentially (first to last)

	Df	Deviance	Resid. Df	Resid. Dev	P(> Chi )
NULL			608	438.56	
age	1	11.33	607	427.22	7.609e-04
smk	1	8.59	606	418.63	3.373e-03
ecg	1	4.45	605	414.18	0.03
chl	1	6.17	604	408.01	0.01
cat	1	5.34	603	402.67	0.02
hpt	1	2.28	602	400.39	0.13
chl:cat	1	43.35	601	357.05	4.582e-11
cat:hpt	1	9.82	600	347.23	1.730e-03

## Exercice.

Commenter les résultats obtenus.

## 5 Comparaison de modèles

### 5.1 Comparaison de deux modèles pré-choisis

Lorsque l'on cherche à comparer des modèles emboîtés, il est possible de procéder comme suit :

```
anova(glm1, glm2, test = "Chisq")
Analysis of Deviance Table
Model 1: chd ~ age + chl + smk + ecg + hpt + cat
Model 2: chd ~ age + smk + ecg + (chl * cat) + (hpt * cat)
  Resid. Df Resid. Dev  Df Deviance P(>|Chi|)
1      602      400.39
2      600      347.23  2    53.16 2.854e-12
```

### 5.2 Comparaison par ajout ou retrait d'une variable

Deux fonctions sont à connaître. Elles sont valables pour les `glm` comme pour les `lm`.

`add1` ajouter un paramètre

`drop1` enlever un paramètre

On part d'un modèle simple, par exemple une variable comme l'âge et on introduit les autres variables sur un mode simple : `age + chl, age + smk, ...`

```
glm0 <- glm(chd ~ age, "binomial")
summary(glm0)
```

```

Call:
glm(formula = chd ~ age, family = "binomial")
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7579 -0.5170 -0.4464 -0.3929  2.3518

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.47833    0.75610  -5.923 3.16e-09
age          0.04445    0.01315   3.381 0.000723

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 438.56  on 608  degrees of freedom
Residual deviance: 427.22  on 607  degrees of freedom
AIC: 431.22

Number of Fisher Scoring iterations: 5

add1(glm0, ~age + chl + smk + ecg + hpt + cat)

Single term additions
Model:
chd ~ age
      Df Deviance   AIC
<none>  1  427.22 431.22
chl     1  422.02 428.02
smk     1  418.63 424.63
ecg     1  423.10 429.10
hpt     1  419.61 425.61
cat     1  420.38 426.38

glm(chd ~ age + chl, "binomial")
Call: glm(formula = chd ~ age + chl, family = "binomial")
Coefficients:
(Intercept)          age          chl
-6.027068      0.044676      0.007112

Degrees of Freedom: 608 Total (i.e. Null);  606 Residual
Null Deviance:          438.6
Residual Deviance: 422      AIC: 428

glm(chd ~ age + smk, "binomial")
Call: glm(formula = chd ~ age + smk, family = "binomial")
Coefficients:
(Intercept)          age          smk1
-5.40622      0.05079      0.83445

Degrees of Freedom: 608 Total (i.e. Null);  606 Residual
Null Deviance:          438.6
Residual Deviance: 418.6      AIC: 424.6

On part d'un modèle sans interaction contenant toutes les variables de contrôle.
On en enlève une et on étudie la modification apportée sur le critère AIC.

glmcc <- glm(chd ~ age + chl + smk + ecg + hpt + cat, "binomial")
drop1(glmcc, ~age + chl + smk + ecg + hpt + cat)

Single term deletions
Model:
chd ~ age + chl + smk + ecg + hpt + cat
      Df Deviance   AIC
<none>  1  400.39 414.39
age     1  404.83 416.83
chl     1  407.50 419.50
smk     1  408.62 420.62
ecg     1  401.95 413.95
hpt     1  402.67 414.67
cat     1  403.26 415.26

glm(chd ~ chl + smk + ecg + hpt + cat, "binomial")
Call: glm(formula = chd ~ chl + smk + ecg + hpt + cat, family = "binomial")
Coefficients:
(Intercept)          chl          smk1          ecg1          hpt1          cat1
-5.105800      0.009124      0.745719      0.383710      0.445640      0.891532

Degrees of Freedom: 608 Total (i.e. Null);  603 Residual
Null Deviance:          438.6
Residual Deviance: 404.8      AIC: 416.8

glm(chd ~ age + smk + ecg + hpt + cat, "binomial")

```

```
Call: glm(formula = chd ~ age + smk + ecg + hpt + cat, family = "binomial")
Coefficients:
(Intercept)      age      smk1      ecg1      hpt1      cat1
-4.98857      0.03451      0.83163      0.34280      0.53661      0.38186

Degrees of Freedom: 608 Total (i.e. Null); 603 Residual
Null Deviance:      438.6
Residual Deviance: 407.5      AIC: 419.5
```

**Problème :** chaque variable est traitée séparément dans un sens (ajout) comme dans un autre (retrait). Cela peut aider dans une phase finale mais non dans une sélection globale.

### 5.3 Généralisation

Afin de procéder d'une manière optimum, il est préférable d'utiliser une autre fonction `step` valable en `glm` comme en `lm`.

On part d'un modèle simple pour le complexifier.

```
glm0 <- glm(chd ~ age + chl, "binomial")
step(glm0, scope = ~age + chl + smk + ecg + hpt + cat, dir = "forward")

Start: AIC=428.02
chd ~ age + chl
      Df Deviance  AIC
+ cat  1   412.16 420.16
+ smk  1   413.17 421.17
+ hpt  1   414.93 422.93
+ ecg  1   417.30 425.30
<none>      422.02 428.02

Step: AIC=420.16
chd ~ age + chl + cat
      Df Deviance  AIC
+ smk  1   404.66 414.66
+ hpt  1   409.65 419.65
<none>      412.16 420.16
+ ecg  1   410.82 420.82

Step: AIC=414.66
chd ~ age + chl + cat + smk
      Df Deviance  AIC
+ hpt  1   401.95 413.95
<none>      404.66 414.66
+ ecg  1   402.67 414.67

Step: AIC=413.95
chd ~ age + chl + cat + smk + hpt
      Df Deviance  AIC
<none>      401.95 413.95
+ ecg  1   400.39 414.39

Call: glm(formula = chd ~ age + chl + cat + smk + hpt, family = "binomial")
Coefficients:
(Intercept)      age      chl      cat1      smk1      hpt1
-6.680112      0.032770      0.008608      0.715810      0.802906      0.476272

Degrees of Freedom: 608 Total (i.e. Null); 603 Residual
Null Deviance:      438.6
Residual Deviance: 401.9      AIC: 413.9
```

On part d'un modèle complet pour le simplifier.

```
glmc <- glm(chd ~ age + chl + smk + ecg + hpt + cat, "binomial")
step(glmc, dir = "backward")

Start: AIC=414.39
chd ~ age + chl + smk + ecg + hpt + cat
      Df Deviance  AIC
- ecg  1   401.95 413.95
<none>      400.39 414.39
- hpt  1   402.67 414.67
```

```
- cat 1 403.26 415.26
- age 1 404.83 416.83
- chl 1 407.50 419.50
- smk 1 408.62 420.62

Step: AIC=413.95
chd ~ age + chl + smk + hpt + cat

      Df Deviance   AIC
<none> 401.95 413.95
- hpt  1 404.66 414.66
- cat  1 406.33 416.33
- age  1 406.52 416.52
- chl  1 408.86 418.86
- smk  1 409.65 419.65

Call: glm(formula = chd ~ age + chl + smk + hpt + cat, family = "binomial")

Coefficients:
(Intercept)      age      chl      smk1      hpt1      cat1
-6.680112    0.032770    0.008608    0.802906    0.476272    0.715810

Degrees of Freedom: 608 Total (i.e. Null); 603 Residual
Null Deviance: 438.6
Residual Deviance: 401.9      AIC: 413.9
```

## 5.4 Exercice

Donner le modèle permettant, selon la minimisation du critère AIC, d'expliquer au mieux la présence ou l'absence de maladie cardiaque des coronaires.