DÉMO5 STAT 5100 : REGRESSION LOGISTIQUE

TRANSFORMATION DE LA BASE DE DONNÉES

Nous utilisons dans ce tutoriel la base de données construite à partir des données de l'article *A Theory of Extramarital Affairs, de Ray Fair, paru en 1978 dans Journal of Political Economy* avec 563 observations.

```
base=read.table("http://freakonometrics.free.fr/baseaffairs.txt",header=TRUE)

# la commande head Affiche les 10 premières lignes de notre base de données
head(base,10)
```

```
##    SEX AGE YEARMARRIAGE CHILDREN RELIGIOUS EDUCATION OCCUPATION
## 1    1  37        10.00        0         3        18          7
## 2    0  27         4.00        0         4        14          6
## 3    0  32        15.00        1         1        12          1
## 4    1  57        15.00        1         5        18          6
## 5    1  22         0.75        0         2        17          6
## 6    0  32         1.50        0         2        17          5
## 7    0  22         0.75        0         2        12          1
## 8    1  57        15.00        1         2        14          4
## 9    0  32        15.00        1         4        16          1
## 10   1  22         1.50        0         4        14          4
##    SATISFACTION Y
## 1              4 0
## 2              4 0
## 3              4 0
## 4              5 0
## 5              3 0
## 6              5 0
## 7              3 0
## 8              4 0
## 9              2 0
## 10             5 0
```

```
#la commande tail affiche quelques dernières lignes de la base de données
tail(base)
```

```
##     SEX AGE YEARMARRIAGE CHILDREN RELIGIOUS EDUCATION OCCUPATION
## 596   1  47         15.0        1         3        16          4
## 597   1  22          1.5        1         1        12          2
## 598   0  32         10.0        1         2        18          5
## 599   1  32         10.0        1         2        17          6
## 600   1  22          7.0        1         3        18          6
## 601   0  32         15.0        1         3        14          1
##     SATISFACTION Y
```

```
## 596           2 7
## 597           5 1
## 598           4 6
## 599           5 2
## 600           2 2
## 601           5 1
```

On pourrait de ce fait déterminer toutes les variables de notre base de données. Ainsi, cette base contient les variables :

- SEX : 0 pour une femme, et 1 pour un homme.
- AGE : âge de la personne interrogée.
- YEARMARRIAGE : nombre d'années de mariage.
- CHILDREN : 0 si la personne n'a pas d'enfants (avec son épouse) et 1 si elle en a.
- RELIGIOUS : degré de "religiosité", entre 1 (anti-religieuse) à 5 (très religieuse).
- EDUCATION : nombre d'années d'éducation, 9=grade school, 12=high school, à 20=PhD.
- OCCUPATION : construit suivant l'échelle d'Hollingshead.
- SATISFACTION : perception de son mariage, de très mécontente (1) à très contente (5).
- Y : nombre d'aventures extra-conjugales hétérosexuelles pendant l'année passée

Nous allons créer deux autres variables pour faciliter l'analyse.

- ENFANTS: OUI si la personne en a, NON sinon.
- SEXE: F pour une femme, et H pour un homme.

```
 base$SEXE="H"
base$SEXE[base$SEX=="0"]="F"
base$SEXE=as.factor(base$SEXE)
table(base$SEXE)

##
##   F   H
## 295 268

base$ENFANT="OUI"
base$ENFANT[base$CHILDREN==0]="NON"
base$ENFANT=as.factor(base$ENFANT)
table(base$ENFANT)

##
## NON OUI
## 164 399

table(base$CHILDREN)
```

```
##
##   0   1
## 164 399

#table utilise les facteurs de classification croisée
#pour créer un tableau de contingence des comptes à chaque
#combinaison de niveaux de facteurs.
```

Le but ici étant d'effectuer la régression logistique sur les données, nous devons nous rassurer que la variable réponse Y est binaire (i.e 0 ou 1). Et si ce n'est pas le cas, nous la transformerons comme tel.

```
table(base$Y)

##
##   0   1   2   3   4   5   6   7   8  10
## 451  34  17  19  12  11  11   5   2   1
```

On constate que la variable Y n'est pas binaire. Pour la transformer, une règle serait de lui donner la valeur 0 si aucune aventure extra-conjugale hétérosexuelle pendant l'année passée et 1 si au moins une aventure extra-conjugale hétérosexuelle pendant l'année passée. Pour ce faire, voir le code ci-dessous :

```
base$Y0=as.numeric(base$Y>0)
table(base$Y0)

##
##   0   1
## 451 112

head(base,3)

##   SEX AGE YEARMARRIAGE CHILDREN RELIGIOUS EDUCATION OCCUPATION
## 1   1  37           10        0         3        18          7
## 2   0  27            4        0         4        14          6
## 3   0  32           15        1         1        12          1
##   SATISFACTION Y SEXE ENFANT Y0
## 1            4 0    H    NON  0
## 2            4 0    F    NON  0
## 3            4 0    F    OUI  0
```

La commande `baseY>0` est booléenne. Elle retourne 'FALSE' si la condition n'est pas respectéee et 'TRUE' dans le cas contraire. `as.numeric` vient transformer 'FALSE' en 0 et 'TRUE' en 1. `table(base$Y0)` donne le nombre de fois qu'on a 0 et 1.

Ayant effectué toutes les transformations nécessaires à notre modélisation à notre base de données, faisons ensuite place à la modélisation (régression logistique) proprement dite.

SÉLECTION DU MEILLEUR MODÈLE.

Il existe plusieurs méthodes de sélection de modèles. Nous allons ici illustrer deux cas. Le premier consistera à sélectionner un modèle parmi deux, et le second à sélectionner le meilleur modèle parmi tous les modèles existants.

1. **Sélection d'un modèle parmi deux modèles imbriqués [1].**
   On pourrait dans ce cas, pour faire simple utiliser le critère du AIC, BIC, R^2 ou encore R^2_{adj}. N'oublions pas le fameux test de rapport de vraisemblance. Nous n'allons bien évidemment pas tous les explorer.
   Le meilleur modèle sera celui qui aura le plus grand R^2/R^2_{adj} ou bien le plus petit AIC/BIC.

   Considérons deux modèles qu'on veut comparer : l'un avec toutes les covariables et l'autre sans les covariables `OCCUPATION` et `SATISFACTION`. La question qu'on se pose est de savoir quel modèle parmi les deux est le 'meilleur' ?

```
#modèle complet avec toutes les covariables sans l'intercept, ni Y.
model1 <- glm(Y0~0+.-Y, family=binomial(link="logit"), data = base)

#modèle complet sans les covariables AGE et OCCUPATION.
model2 <- glm(Y0~0+.-Y-AGE-OCCUPATION, family=binomial(link="logit"),
              data = base)

#AIC des deux modèles.
c(model1$aic, model2$aic)
## [1] 541.0011 541.7181
```

   Critère AIC/BIC : On retient le modèle ayant le plus petit AIC/BIC.
   De la même façon (`model$bic`), on obtient le BIC.

   Il n'apparait dans les sorties `R` ni le R^2, ni le R^2_{adj}. Nous allons déterminer analytiquement le R^2 comme suit :

```
R2.model1=1-model1$deviance/model1$null.deviance
R2.model2=1-model2$deviance/model2$null.deviance
c(R2.model1,R2.model2)
## [1] 0.3299013 0.3238576
```

   Critère R^2 : On retient le modèle ayant le plus grand R^2.
   Dans tous les cas, le meilleur modèle parmi les deux est le `model1`.

1. Deux modèles sont imbriqués lorsque l'un est égal à l'autre sans au moins une covariable

2. **Sélection du meilleur modèle.**
Nous allons considérer le modèle avec toutes les covariables et à partir d'une fonction `R`, déterminer le meilleur modèle.

```
#stepAIC nécessite le package MASS.
library(MASS)

out <- stepAIC(model1, direction = "both")
## Start:  AIC=541
## Y0 ~ 0 + (SEX + AGE + YEARMARRIAGE + CHILDREN + RELIGIOUS + EDUCATION +
##     OCCUPATION + SATISFACTION + Y + SEXE + ENFANT) - Y
##
##
## Step:  AIC=541
## Y0 ~ SEX + AGE + YEARMARRIAGE + CHILDREN + RELIGIOUS + EDUCATION +
##     OCCUPATION + SATISFACTION + SEXE - 1
##
##
## Step:  AIC=541
## Y0 ~ AGE + YEARMARRIAGE + CHILDREN + RELIGIOUS + EDUCATION +
##     OCCUPATION + SATISFACTION + SEXE - 1
##
##                Df Deviance    AIC
## - SEXE          2   524.31 538.31
## - OCCUPATION    1   523.04 539.04
## - EDUCATION     1   524.79 540.79
## <none>              523.00 541.00
## - YEARMARRIAGE  1   526.27 542.27
## - CHILDREN      1   526.33 542.33
## - AGE           1   527.61 543.61
## - RELIGIOUS     1   529.43 545.43
## - SATISFACTION  1   538.51 554.51
##
## Step:  AIC=538.31
## Y0 ~ AGE + YEARMARRIAGE + CHILDREN + RELIGIOUS + EDUCATION +
##     OCCUPATION + SATISFACTION - 1
##
##                Df Deviance    AIC
## - OCCUPATION    1   524.40 536.40
## <none>              524.31 538.31
## + SEX           1   523.00 539.00
## - YEARMARRIAGE  1   527.32 539.32
## - EDUCATION     1   527.67 539.67
## - CHILDREN      1   528.18 540.18
## + SEXE          2   523.00 541.00
## - AGE           1   529.34 541.34
## - RELIGIOUS     1   531.43 543.43
```

```
## - SATISFACTION  1   543.07 555.07
##
## Step:  AIC=536.4
## Y0 ~ AGE + YEARMARRIAGE + CHILDREN + RELIGIOUS + EDUCATION +
##     SATISFACTION - 1
##
##                Df Deviance    AIC
## <none>              524.40 536.40
## + SEX           1   523.04 537.04
## - YEARMARRIAGE  1   527.41 537.41
## - CHILDREN      1   528.18 538.18
## + OCCUPATION    1   524.31 538.31
## + SEXE          2   523.04 539.04
## - AGE           1   529.35 539.35
## - EDUCATION     1   529.97 539.97
## - RELIGIOUS     1   531.67 541.67
## - SATISFACTION  1   543.88 553.88
#Summary(out) nous produit le meilleur modèle.
summary(out)
##
## Call:
## glm(formula = Y0 ~ AGE + YEARMARRIAGE + CHILDREN + RELIGIOUS +
##     EDUCATION + SATISFACTION - 1, family = binomial(link = "logit"),
##     data = base)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3121  -0.6976  -0.5448  -0.3812   2.4088
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## AGE          -0.04052    0.01869  -2.168  0.03016 *
## YEARMARRIAGE  0.05900    0.03409   1.731  0.08344 .
## CHILDREN      0.60904    0.31697   1.921  0.05468 .
## RELIGIOUS    -0.25476    0.09503  -2.681  0.00734 **
## EDUCATION     0.08229    0.03554   2.315  0.02059 *
## SATISFACTION -0.41164    0.09398  -4.380 1.19e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 780.48  on 563  degrees of freedom
## Residual deviance: 524.40  on 557  degrees of freedom
## AIC: 536.4
##
## Number of Fisher Scoring iterations: 4
```

Remarque : On aurait utilié `stepAIC(model1, direction = "forward")` ou `stepAIC(model1, direction = "backward")` qu'on obtiendrait le même résultat.