DÉMO4 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
```

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. `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.

MODÉLISATION

FORMULE MATHÉMATIQUE.
La transformation la plus courante est

$$g(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \mathbf{X_i}^\top \beta,$$

$$p_i := P[Y=1] = \frac{e^{\mathbf{X_i}^\top \beta}}{1+e^{\mathbf{X_i}^\top \beta}}.$$

CODE R.

```r
out = glm(Y0~YEARMARRIAGE+SEXE,data=base, family=binomial(link="logit"))
summary(out)
```

```
##
## Call:
## glm(formula = Y0 ~ YEARMARRIAGE + SEXE, family = binomial(link = "logit"),
##     data = base)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7921  -0.6983  -0.6339  -0.5517   1.9849
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.84763    0.22411  -8.244   <2e-16 ***
## YEARMARRIAGE  0.03738    0.01911   1.957   0.0504 .
## SEXEH         0.28865    0.21247   1.358   0.1743
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 561.79  on 562  degrees of freedom
## Residual deviance: 555.94  on 560  degrees of freedom
## AIC: 561.94
##
## Number of Fisher Scoring iterations: 4
```

On constate que les variables `YEARMARRIAGE` et `SEXE`, prises individuellement, ne sont pas significatives pour le modèle. La variable $\mathbb{1}$ est significative. Rappelons ici que l'intercept contient une partie [1] des paramètres correspondant aux variables `YEARMARRIAGE` et `SEXE`

On retire l'intercept pour voir comment se comporte le modèle.

```r
out0 = glm(Y0~YEARMARRIAGE+SEXE,data=base, family=binomial(link="logit"))
summary(out0)
```

```
##
## Call:
## glm(formula = Y0 ~ YEARMARRIAGE + SEXE, family = binomial(link = "logit"),
##     data = base)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7921  -0.6983  -0.6339  -0.5517   1.9849
##
## Coefficients:
```

1. à expliquer

```
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.84763    0.22411  -8.244   <2e-16 ***
## YEARMARRIAGE  0.03738    0.01911   1.957   0.0504 .
## SEXEH         0.28865    0.21247   1.358   0.1743
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 561.79  on 562  degrees of freedom
## Residual deviance: 555.94  on 560  degrees of freedom
## AIC: 561.94
##
## Number of Fisher Scoring iterations: 4
```

Ici, toutes les variables sont significatives pour le modèle.

```
out1 = glm(Y0~SEXE,data=base, family=binomial(link="logit"))
summary(out1)

##
## Call:
## glm(formula = Y0 ~ SEXE, family = binomial(link = "logit"), data = base)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7120  -0.7120  -0.6228  -0.6228   1.8632
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.5418     0.1528 -10.091   <2e-16 ***
## SEXEH         0.2986     0.2117   1.411    0.158
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 561.79  on 562  degrees of freedom
## Residual deviance: 559.79  on 561  degrees of freedom
## AIC: 563.79
##
## Number of Fisher Scoring iterations: 4
```

```
out2 = glm(Y0~YEARMARRIAGE+SEXE+SATISFACTION+CHILDREN,data=base, family=binomial(link="logit"))
summary(out2)

##
```

```
## Call:
## glm(formula = Y0 ~ YEARMARRIAGE + SEXE + SATISFACTION + CHILDREN,
##     family = binomial(link = "logit"), data = base)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1631  -0.7185  -0.5657  -0.4143   2.2424
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.496450   0.485240  -1.023 0.306259
## YEARMARRIAGE -0.003209   0.023529  -0.136 0.891519
## SEXEH         0.278693   0.216968   1.284 0.198970
## SATISFACTION -0.382175   0.098316  -3.887 0.000101 ***
## CHILDREN      0.614351   0.318600   1.928 0.053820 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 561.79  on 562  degrees of freedom
## Residual deviance: 536.02  on 558  degrees of freedom
## AIC: 546.02
##
## Number of Fisher Scoring iterations: 4
```

Exercice :
Considérez le jeu de données ci-dessus. On souhaite utiliser la régression logistique comme technique de scoring pour prédire par exemple si un emprunteur sera un bon ou un mauvais payeur (et donc, si on lui accordera un crédit ou non).

N.B. : Dans ce qui suit, on ne tiendra pas compte de la QUALITÉ du modèle. L'exercice est donné ici dans le but de s'exercer. Les exercices sur la qualité du modèle et les tests d'hypothèses vous seront éventuellement donnés ultérieurement.

a. Pourquoi faut-il utiliser le modèle de régression logistique plutôt que le modèle de régression normale ?
b. Le sexe d'un individu a-t-il un impact sur sa probabilité d'avoir des relations extraconjugales ? Si oui, quelle est la différence entre la probabilité d'avoir des relations extraconjugales pour un homme et pour une femme ?
c. Pour un homme de 5 ans de mariage, qui est très mécontent et qui n'a pas d'enfants, quel serait l'impact pour lui d'avoir au moins un enfant sur sa probabilité d'avoir des relations extraconjugales ?

Proposition de solution :

a. Étant donné que la variable dépendante est binaire (0 ou 1), l'hypothèse de normalité ne sera pas respectée. Il faut alors utiliser le modèle de régression logistique.

b. — Calculer avec le modèle `out1` pour un homme $S = \hat{\beta}_0 + \hat{\beta}_1$

$$P[Y = 1|SEXE = 1] = \frac{e^S}{1 + e^S},$$

et pour une femme $S' = \hat{\beta}_0$

$$P[Y = 1|SEXE = 0] = \frac{S'}{1 + e^{S'}},$$

— Conclure.

c. — Calculer en considérant le modèle `out2` la quantité

$$S_1 = \hat{\beta}_0 + \hat{\beta}_1(5) + \hat{\beta}_2(1) + \hat{\beta}_3(0),$$

puis

$$P_0 = \frac{e^{S_1}}{1 + e^{S_1}}$$

— Calculer la quantité

$$S_2 = \hat{\beta}_0 + \hat{\beta}_1(5) + \hat{\beta}_2(1) + \hat{\beta}_3(1),$$

puis

$$P_1 = \frac{e^{S_2}}{1 + e^{S_2}}$$

— Comparer enfin les deux résultats obtenus. Si $P_0 \geq P_1$, alors baisse de $P_0 - P_1$ ...

Les exercices sur les tests et la qualité du modèle vous seront *éventuellement* proposés ultérieurement.