DÉMO6 STAT 5100 : REGRESSION DE POISSON

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)
```

```
##   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
##   SATISFACTION Y
## 1            4 0
## 2            4 0
## 3            4 0
## 4            5 0
## 5            3 0
## 6            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 de POISSON sur les données, nous devons nous

rassurer que la variable réponse Y est un élément de $\mathbb{N}$.

```
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 a ses valeurs dans $\mathbb{N}$. Nul besoin de transformations pour la régression de poisson.

RÉGRESSION DE POISSON.

```
library(MASS)
out <- glm(Y~0+.,data=base,family=poisson(link="log"))
#sélection du meilleur modèle
l <- stepAIC(out, direction = 'both')

## Start:  AIC=1468.79
## Y ~ 0 + (SEX + AGE + YEARMARRIAGE + CHILDREN + RELIGIOUS + EDUCATION +
##     OCCUPATION + SATISFACTION + SEXE + ENFANT)
##
##
## Step:  AIC=1468.79
## Y ~ SEX + AGE + YEARMARRIAGE + CHILDREN + RELIGIOUS + EDUCATION +
##     OCCUPATION + SATISFACTION + SEXE - 1
##
##
## Step:  AIC=1468.79
## Y ~ AGE + YEARMARRIAGE + CHILDREN + RELIGIOUS + EDUCATION + OCCUPATION +
##     SATISFACTION + SEXE - 1
##
##                Df Deviance    AIC
## - SEXE          2   1132.9 1464.9
## - OCCUPATION    1   1133.0 1467.1
## <none>              1132.7 1468.8
## - CHILDREN      1   1141.5 1475.5
## - EDUCATION     1   1150.2 1484.3
## - AGE           1   1151.2 1485.2
## - YEARMARRIAGE  1   1151.8 1485.9
## - RELIGIOUS     1   1160.2 1494.3
## - SATISFACTION  1   1190.8 1524.9
##
```

```
## Step:  AIC=1464.93
## Y ~ AGE + YEARMARRIAGE + CHILDREN + RELIGIOUS + EDUCATION + OCCUPATION +
##     SATISFACTION - 1
##
##                Df Deviance    AIC
## - OCCUPATION    1   1133.5 1463.5
## <none>              1132.9 1464.9
## + SEX           1   1132.8 1466.8
## + SEXE          2   1132.7 1468.8
## - CHILDREN      1   1142.0 1472.0
## - YEARMARRIAGE  1   1153.6 1483.6
## - AGE           1   1154.7 1484.7
## - RELIGIOUS     1   1162.5 1492.6
## - EDUCATION     1   1164.8 1494.9
## - SATISFACTION  1   1200.4 1530.4
##
## Step:  AIC=1463.51
## Y ~ AGE + YEARMARRIAGE + CHILDREN + RELIGIOUS + EDUCATION + SATISFACTION -
##     1
##
##                Df Deviance    AIC
## <none>              1133.5 1463.5
## + OCCUPATION    1   1132.9 1464.9
## + SEX           1   1133.1 1465.1
## + SEXE          2   1133.0 1467.1
## - CHILDREN      1   1142.1 1470.2
## - YEARMARRIAGE  1   1154.2 1482.3
## - AGE           1   1154.9 1483.0
## - RELIGIOUS     1   1164.0 1492.0
## - EDUCATION     1   1187.6 1515.7
## - SATISFACTION  1   1204.2 1532.3
```

```
summary(l)
```

```
##
## Call:
## glm(formula = Y ~ AGE + YEARMARRIAGE + CHILDREN + RELIGIOUS +
##     EDUCATION + SATISFACTION - 1, family = poisson(link = "log"),
##     data = base)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6882  -1.1268  -0.8569  -0.5870   5.6450
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## AGE          -0.041156   0.009267  -4.441 8.95e-06 ***
## YEARMARRIAGE  0.074748   0.016527   4.523 6.10e-06 ***
```

```
## CHILDREN      0.477739   0.164742   2.900  0.00373 **
## RELIGIOUS    -0.259263   0.047150  -5.499 3.83e-08 ***
## EDUCATION     0.123338   0.016864   7.314 2.60e-13 ***
## SATISFACTION -0.369905   0.043296  -8.544  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1383.8  on 563  degrees of freedom
## Residual deviance: 1133.5  on 557  degrees of freedom
## AIC: 1463.5
##
## Number of Fisher Scoring iterations: 6
```

```
# -1 dans le modèle est une autre façon de supprimer l'intercept
out0 <- glm(Y ~ AGE + YEARMARRIAGE + CHILDREN + RELIGIOUS +
    EDUCATION + SATISFACTION -1 , family = poisson(link = "log"),
    data = base)
summary(out0)
```

```
##
## Call:
## glm(formula = Y ~ AGE + YEARMARRIAGE + CHILDREN + RELIGIOUS +
##     EDUCATION + SATISFACTION - 1, family = poisson(link = "log"),
##     data = base)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6882  -1.1268  -0.8569  -0.5870   5.6450
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## AGE          -0.041156   0.009267  -4.441 8.95e-06 ***
## YEARMARRIAGE  0.074748   0.016527   4.523 6.10e-06 ***
## CHILDREN      0.477739   0.164742   2.900  0.00373 **
## RELIGIOUS    -0.259263   0.047150  -5.499 3.83e-08 ***
## EDUCATION     0.123338   0.016864   7.314 2.60e-13 ***
## SATISFACTION -0.369905   0.043296  -8.544  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1383.8  on 563  degrees of freedom
## Residual deviance: 1133.5  on 557  degrees of freedom
## AIC: 1463.5
##
```

```
## Number of Fisher Scoring iterations: 6
```

SUR/SOUS-DISPERSION

Le problème qu'on rencontre généralement avec la régression de poisson est celui de la sous-dispersion ou de la sur-dispersion. En effet, une des hypothèses sous-jacentes à ce modèle est l'égalité entre la moyenne et la variance de la variable à expliquer. Il existe donc plusieurs tests pour détecter un éventuel problème de sur/sous-dispersion, notamment le test d'égalité de la moyenne et de la variance. On pourrait aussi effectuer la régression quasi-poisson puis vérifier si le paramètre de dispersion ϕ vaut 1 [1]. Dans le cas d'un tel problème, une solution serait de faire un ajustement sur le Khi-deux ou un ajustement avec un modèle de régression binomial négative [2].

```
#test de sous/sur-dispersion
#regression quasipoisson pour voir si le paramètre de dispersion vaut 1
out1 <-  glm(Y ~ AGE + YEARMARRIAGE + CHILDREN + RELIGIOUS +
    EDUCATION + SATISFACTION -1 , family = quasipoisson,
    data = base)
summary(out1)

##
## Call:
## glm(formula = Y ~ AGE + YEARMARRIAGE + CHILDREN + RELIGIOUS +
##     EDUCATION + SATISFACTION - 1, family = quasipoisson, data = base)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6882  -1.1268  -0.8569  -0.5870   5.6450
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## AGE          -0.04116    0.01737  -2.369 0.018183 *
## YEARMARRIAGE  0.07475    0.03098   2.413 0.016165 *
## CHILDREN      0.47774    0.30885   1.547 0.122477
## RELIGIOUS    -0.25926    0.08840  -2.933 0.003496 **
## EDUCATION     0.12334    0.03162   3.901 0.000107 ***
## SATISFACTION -0.36990    0.08117  -4.557 6.38e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

1. Si $\phi > 1$, surdispersion. Si $\phi < 1$, sous-dispersion
2. Poisson, quasi-poisson, géométrique, binomiale négative sont des régressions pour des données de comptage

```
## (Dispersion parameter for quasipoisson family taken to be 3.514804)
##
##     Null deviance: 1383.8  on 563  degrees of freedom
## Residual deviance: 1133.5  on 557  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 6
```

```
summary(out1)$dispersion
```

```
## [1] 3.514804
```

```
# une autre façon simple de procéder. Cependant ce test nécessite la library AER
library(AER)
```

```
## Loading required package: car
## Loading required package: carData
## Loading required package: lmtest
## Loading required package: zoo
##
## Attaching package: 'zoo'
## The following objects are masked from 'package:base':
##
##    as.Date, as.Date.numeric
## Loading required package: sandwich
## Loading required package: survival
```

```
dispersiontest(out0)
```

```
##
##  Overdispersion test
##
## data:  out0
## z = 5.6885, p-value = 6.407e-09
## alternative hypothesis: true dispersion is greater than 1
## sample estimates:
## dispersion
##   3.508133
```

On peut remarquer que le paramètre de dispersion qui vaut 3.514804 est supérieur à 1. Il y a donc surdispersion.

On pourrait également pour ce test regarder l'écart entre la déviance et le nombre de dégrés de liberté. Dans notre cas, la déviance est deux fois plus élevée que le nombre dégré de liberté, ce qui indique un problème d'ajustement. On pourrait soupçonner une surdispersion.

CONSÉQUENCES DE LA SUR-DISPERSION ET SOLUTION

— Les estimations de l'erreur standard, de la statistique du Khi-deux et de la valeur p faussées [3].
— On risque de déclarer trop souvent les paramètres du modèle (qui sont sans biais) significatifs.

Une des solutions serait d'ajuster la surdispersion. Procéder comme suit :

— Calculer $c =$

Une autre solution serait d'effectuer une régression binomiale négative. SI LE PARAMÈTRE DE DISPERSION EST ÉGAL À 0, ON GARDE NOTRE MODÈLE DE RÉGRESSION DE POISSON.

```
out2 <-  glm.nb(Y ~ AGE + YEARMARRIAGE + CHILDREN + RELIGIOUS +
    EDUCATION + SATISFACTION -1 ,  data = base)
summary(out2)
##
## Call:
## glm.nb(formula = Y ~ AGE + YEARMARRIAGE + CHILDREN + RELIGIOUS +
##     EDUCATION + SATISFACTION - 1, data = base, init.theta = 0.1629456013,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0435  -0.7187  -0.6032  -0.4610   2.2006
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## AGE          -0.04563    0.02057  -2.218 0.026541 *
## YEARMARRIAGE  0.08875    0.03864   2.297 0.021614 *
## CHILDREN      0.40627    0.34049   1.193 0.232794
## RELIGIOUS    -0.26707    0.10610  -2.517 0.011833 *
## EDUCATION     0.14097    0.03961   3.559 0.000372 ***
## SATISFACTION -0.42168    0.10757  -3.920 8.85e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.1629) family taken to be 1)
##
##     Null deviance: 330.05  on 563  degrees of freedom
## Residual deviance: 280.11  on 557  degrees of freedom
## AIC: 1002.2
##
## Number of Fisher Scoring iterations: 1
##
##
```

3. Erreur standard sous-estimée et statistique du Khi-deux sur-estimée

```
##                Theta:  0.1629
##            Std. Err.:  0.0236
##
##  2 x log-likelihood:  -988.2180
summary(out2)$dispersion
## [1] 1
```

Le paramètre de dispersion tel que lu sur la sortie vaut 0.1629. Il est recommandé par `R` de le prendre égal à 1. À chacun de conclure. Sinon, on serait tenté d'opter pour le modèle de régression de Poisson ajusté.