

# analyse groupe de variables

```
load("../RData")
require(knitr)
```

```
## Loading required package: knitr
```

```
colnames(dataPropre)
```

```
## [1] "Experimentateur" "Pédagogie"      "Classe"
## [4] "TypeClasse"      "Sexe"           "Langues"
## [7] "Lateralite"      "DateNaissance"  "DateEval"
## [10] "AgeChar"         "AgeNum"         "AgeInt"
## [13] "T1"              "T21"            "T22"
## [16] "T23"             "T31"            "T32"
## [19] "T41a"            "T41b"           "T41c"
## [22] "T41d"            "T42a"           "T42b"
## [25] "T42c"            "T42d"           "T51"
## [28] "T52"             "T61"            "T62"
## [31] "T71"             "T72"            "T81"
## [34] "T82"             "T83"            "T84"
## [37] "T85"             "T86"            "T87"
## [40] "T88"             "T89"            "AnneeScolaire"
```

```
noms<-c("T23","T31","T52","T62","T86","T87","T88","T89","T42a","T42b","T42c","T42d")
audela<-rowSums(apply(dataPropre[,noms],2,as.numeric))#somme de chaque question concerné par la var aud
audela<-audela+ifelse(dataPropre$T1>7,1,0)#ajout de la t1 si ils savent compter au dela de 7
```

```
noms<-c("T41a","T41b","T41c","T41d","T42a","T42b","T42c","T42d","T51","T52","T62","T61")
outils<-rowSums(apply(dataPropre[,noms],2,as.numeric))
```

```
noms<-c("T21","T22","T23","T31","T32","T81","T82","T83",
        "T84","T85","T86","T87","T88","T89")
objet<-rowSums(apply(dataPropre[,noms],2,as.numeric))
```

Etude du chi2 de chaque variable du groupe au dela chacune entre elles

```
noms<-c("T23","T31","T42a","T42b","T42c","T42d","T52","T62","T86","T87","T88","T89")
res.chi2<-matrix(nrow = length(noms),ncol = length(noms))
colnames(res.chi2)<-noms
row.names(res.chi2)<-noms
for (i in 1:length(noms)){
  for (j in 1:length(noms)){
    res.chi2[i,j]<-chisq.test(table(dataPropre[,noms[i]],dataPropre[,noms[j]]))$p.value
  }
}
```

```
kable(iffelse(res.chi2>10**-5,"0","1"))
```

	T23	T31	T42a	T42b	T42c	T42d	T52	T62	T86	T87	T88	T89
T23	1	1	0	0	0	0	0	0	0	0	1	0
T31	1	1	0	0	1	0	0	0	1	1	1	1
T42a	0	0	1	1	0	1	0	0	0	0	0	0
T42b	0	0	1	1	1	1	0	0	0	0	0	1
T42c	0	1	0	1	1	1	0	0	0	0	0	0
T42d	0	0	1	1	1	1	0	0	0	0	0	0
T52	0	0	0	0	0	0	1	0	0	0	0	0
T62	0	0	0	0	0	0	0	1	0	0	0	0
T86	0	1	0	0	0	0	0	0	1	1	1	1
T87	0	1	0	0	0	0	0	0	1	1	1	1
T88	1	1	0	0	0	0	0	0	1	1	1	1
T89	0	1	0	1	0	0	0	0	1	1	1	1

Un “1” est mis dans les cases qui sont significatives afin que le tableau soit plus facilement lisible. On peut voir que les couples de variables qui ont un chi2 significatif sont : T21 T22, T22 T32, T84 T22, T84 T21, T83 T22, T85 T32, et toutes les T8 entre elles.

## Création des groupes sur la Q1

```
Classe_T1<-cut(dataPropre$T1,breaks = c(-1,3,7,11,16,100))
levels(Classe_T1)<-c(1/3,2/3,1,4/3,5/3)
Classe_T1<-as.numeric(Classe_T1)
don.groupe<-data.frame(Pedagogie=dataPropre$Pedagogie,Classe_T1,audela,outils,objet)
```

## Regression sur les nouveaux groupes

```
# reg<-glm(Pedagogie~.,data = don.groupe,family = binomial)
# step(reg) ##on ne garde que la regression avec le regroupement "au dela" les autres ne sont pas signifi.
reg<-glm(Pedagogie~audela,data = don.groupe,family = binomial)
summary(reg)##audela significatif avec la pédagogie
```

```
##
## Call:
## glm(formula = Pedagogie ~ audela, family = binomial, data = don.groupe)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.2743  -0.9914  -0.8320   1.2978   1.6133
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.98361    0.30651  -3.209  0.00133 **
## audela       0.10071    0.04929   2.043  0.04101 *
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 207.88  on 155  degrees of freedom
## Residual deviance: 203.62  on 154  degrees of freedom
## AIC: 207.62
##
## Number of Fisher Scoring iterations: 4
```

Seule la variable audela est significative, afin de prédire la pédagogie. On peut donc dire que les question de type “audela” peuvent définir un élève classique d’un élève montessorien.

La partie suivante consiste à créer un échantillon sur lequel refaire le même modèle, pour ensuite prédire la pédagogie sur le reste du jeu de données. Cependant un problème est rencontré : le modèle n’est plus adéquate, la variable n’est pas significative..

```
# install.packages("ResourceSelection")
require(ResourceSelection)
```

```
## Loading required package: ResourceSelection
```

```
## Warning: package 'ResourceSelection' was built under R version 3.5.3
```

```
## ResourceSelection 0.3-4    2019-01-08
```

```
hoslem.test(don.groupe$Pédagogie,fitted(reg))
```

```
## Warning in Ops.factor(1, y): '-' not meaningful for factors
```

```
##
## Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  don.groupe$Pédagogie, fitted(reg)
## X-squared = 156, df = 8, p-value < 2.2e-16
```

Le test de hosmer-Lemeshow nous donne une pvalue significative, il y a donc une différence entre le modèle et les données observées, le modèle ne serait donc pas adéquate...

## Réalisation du modèle sur un échantillon :

```
train.id<-sample(seq_len(nrow(don.groupe)),size = 109)
don.train<-don.groupe[train.id,]
don.test<-don.groupe[-train.id,]
```

## Réalisation du modèle :

```
reg<-glm(Pedagogie~Classe_T1,data = don.train,family = binomial)
summary(reg)
```

```
##
## Call:
## glm(formula = Pedagogie ~ Classe_T1, family = binomial, data = don.train)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.2994  -1.0451  -0.8186   1.1853   1.5851
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.2221     0.4826  -2.533  0.0113 *
## Classe_T1      0.3009     0.1439   2.091  0.0365 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 148.44  on 108  degrees of freedom
## Residual deviance: 143.91  on 107  degrees of freedom
## AIC: 147.91
##
## Number of Fisher Scoring iterations: 4
```