FICHE TP AVEC LE LOGICIEL

REGRESSION LOGISTIQUE SOUS
Mise en œuvre de la fonction « `glm` »

LE CANCER DU SEIN

En 2012, on estime à 48 763 le nombre de nouveaux cas en France, soit plus du tiers de l'ensemble des nouveaux cas de cancer chez la femme. Dans plus de 8 cas sur 10, il touche des femmes âgées de 50 ans et plus. Ce cancer peut aussi apparaître chez l'homme, mais c'est extrêmement rare (moins de 1 % des cancers du sein). Aujourd'hui, le taux global de survie relative à 5 ans (proportion de personnes atteintes d'une maladie et vivantes 5 années après le diagnostic en l'absence des autres causes de décès) après le diagnostic d'un cancer du sein est estimé à près de 89 %.

LES TYPES DE CANCER DU SEIN

Les cancers du sein les plus fréquents (95 %) sont des adénocarcinomes : ils se développent à partir des cellules épithéliales de la glande mammaire :

- **carcinome in situ** (limitées aux canaux ou aux lobules du sein)
- **cancer ou carcinome infiltrant** (franchissement de la membrane basale et infiltration du tissu qui entoure les canaux et les lobules)
- **cancer métastatique** (des cellules cancéreuses peuvent se détacher de la tumeur et emprunter les vaisseaux sanguins ou les vaisseaux lymphatiques pour atteindre d'autres parties du corps)

CANCER DU SEIN ET AUTOANTICORPS

Le dépistage par mammographie a permis de significativement augmenter la détection des cancers du sein à des stades précoces. Dans ce contexte, le dosage d'un panel d'autoanticorps (dirigés contre des antigènes associés aux tumeurs) est également une technique de détection précoce des cancers du sein, et plus particulièrement, des carcinomes in situ (CIS, non invasif) et pourrait s'avérer une alternative intéressante à la mammographie.

OBJECTIF

Notre étude visera à expliquer :

- la survenue d'un cancer du sein (non malade vs malade)
- le type de la maladie (contrôle vs non invasif vs invasif)

et sélectionner (dans ces deux situations) les auto-anticorps (+/- variables démographiques et cliniques) les plus associés à l'évènement d'intérêt.

PRESENTATION / LECTURE DES DONNEES

Les données à analyser (fichier Data_cancer_sein.xlsx) contiennent les informations sur une cohorte de 195 patientes. Il s'agit d'étudier, dans un premier temps, la variable d'intérêt « survenue ou non d'un cancer du sein ».

```
data = read.csv(file = "Data_cancer_sein.csv", header = T)
head(data)
tail(data)
names(data)
```

Les variables sont définies ci-dessous :

`npat` identifiant de la patiente dans l'étude (données toujours anonymes). Chaque observation a un identifiant unique soit une observation par patiente.
`surv_cancer` survenue d'un cancer, variable dichotomique prenant la valeur (1) si la patiente est malade et (0) sinon
`type` type de la maladie, variable polytomique prenant la valeur (12) si la patiente présente un cancer non invasif (ou CIS), (11) si la patiente présente un cancer invasif (ou infiltrant) et (0) sinon (la patiente n'est pas atteinte de cancer).
`age` variable continue exprimée en année
`re*` récepteur œstrogènes, variable dichotomique indiquant si les récepteurs aux œstrogènes sont positif (1) ou négatif (0)
`rp**` récepteur progestérone, variable dichotomique indiquant si les récepteurs à la progestérone sont positif (1) ou négatif (0)
`her2` protéines HER2, variable dichotomique indiquant si le cancer est HER2 positif (1) c'est-à-dire si les cellules présentent à leur surface une quantité importante de protéines HER2 qui ont pour propriété de favoriser la croissance des cellules tumorales et (0) sinon
`hsp60` variable continue quantifiant l'auto-anticorps « heat shock protein 60 » (HSP60)
`muc1` variable continue quantifiant l'auto-anticorps mucine 1 (MUC1)
`prdx2` variable continue quantifiant l'auto-anticorps peroxirédoxine 2 (PRDX2)
`ppia` variable continue quantifiant l'auto-anticorps cyclophilline A (PPIA)
`fkbp52` variable continue quantifiant l'auto-anticorps immunophiline (FKBP52)

Remarque : Le choix du traitement dans le cancer du sein dépend de plusieurs facteurs, par exemple, s'il est ou n'est pas **hormonosensible**, c'est-à-dire si sa croissance est stimulée par les hormones féminines (oestrogènes, progestérone) naturellement produites par l'organisme.

OBJECTIF DE MODELISATION

On souhaite modéliser la probabilité de survenue de la maladie (`surv_cancer` la survenue (1) ou l'absence (0) de la maladie) compte tenu des covariables démographiques (`age`), cliniques (`re`, `rp` et `her2`) et biologiques (`hsp60`, `muc1`, `prdx2`, `ppia` et `fkbp52`) sur les 100 patientes de la cohorte étudiée.

ANALYSE EXPLORATOIRE

```
str(data)
summary(data)
```

Les données dichotomiques ne sont pas traitées comme des variables qualitatives. On les transforme donc en type facteur :

```
names(data[,c(2,3,5,6,7)])
data_col = c(2,3,5,6,7)
for (i in 1:5){
data[,data_col[i]] = factor(data[,data_col[i]])}
```

Labélisation des données qualitatives :

```
data[,2] = factor(data[,2],levels=c(levels(data[,2])),c("Controle","Cancer"))
data[,3] =
factor(data[,3],levels=c(levels(data[,3])),c("Controle","Infiltrant","CIS"))
data[,5] = factor(data[,5],levels=c(levels(data[,5])),c("Negatif","Positif"))
data[,6] = factor(data[,6],levels=c(levels(data[,6])),c("Negatif","Positif"))
data[,7] = factor(data[,7],levels=c(levels(data[,7])),c("Negatif","Positif"))

str(data)
summary(data)
```

Description des données qualitatives (exemple) :

```
table(data[,5])
table(data[,2],data[,3])
```

Description des données quantitatives (exemple) :

```
median(data[,8])
min(data[,8])
max(data[,8])
```

Représentation graphique des données quantitatives (exemple):

```
boxplot(data[,8],main = "",xlab = "HSP60",ylab = "Densité optique")
hist(data[,4],col = "lightblue",main = "Age",xlab = "",ylab = "")

attach(data)
```

Discrétisation de la variable âge :

```
age_cl = rep(0,length(data$age))
age_cl[data$age<=55] = 1
age_cl[data$age>55 & data$age<=65] = 2
age_cl[data$age>65] = 3
age_cl = as.factor(age_cl)
age_cl = factor(age_cl,levels=c(levels(age_cl)),c("<=55","]55;65]",">65"))
table(age_cl)
```

ou

```
borne = c(min(data$age),55,65,max(data$age))
age_cl2 = cut(data$age, breaks = borne, include.lowest = TRUE)
table(age_cl, age_cl2)
```

MODELE LOGISTIQUE

```
data = read.csv(file = "Data_cancer_sein.csv",header = T)
table(data$surv_cancer)

modele = glm(surv_cancer ~ age_cl+hsp60+muc1+prdx2+ppia+fkbp52,family = binomial(link="logit"), data = data)
summary(modele)
attributes(modele)
```

Par défaut, R ne donne pas les odds-ratios (OR). Il est toutefois facile de les recalculer car un OR n'est rien de plus que l'estimation des coefficients de la régression. On peut donc obtenir les ORs et leurs intervalles de confiance avec les commandes suivantes :

```
exp(cbind(OR = coef(modele), confint(modele)))
```

L'OR de la constante n'est en général pas interprété. On interprète un OR d'une variable qualitative seulement à partir de la modalité de référence. Par exemple, le fait d'avoir un « age > 65 ans » augmente la probabilité de survenue de cancer de 2,5 fois. De manière générale, un intervalle de confiance de l'OR qui contient 1 implique que la variable n'est pas significative. Mais, sous R, cela peut déjà se vérifier en regardant les résultats de la fonction glm.

VALIDATION DU MODELE

La qualité d'ajustement du modèle ainsi obtenu peut alors être évalué au moyen des critères d'adéquation classiques : la statistique de vraisemblance ou déviance (= –2 Log (L(α, βi))) et la statistique de Pearson ou selon les Critères d'Information de Akaike (AIC = –2Log(L(α,βi)) + 2p) et de Schwarz (Bayesian Information Criterion) (BIC = –2Log(L(α,βi)) + pLog(n))) qui pénalise la vraisemblance quand le nombre de paramètres augmentent. Le modèle est d'autant plus intéressant que la valeur de ces deux critères est faible.

```
modele$deviance
AIC(modele)
BIC(modele)
```

Ensuite, la calibration du modèle, c'est-à-dire la comparaison des probabilités prédites par le modèle à celles observées dans l'échantillon, doit être vérifiée à l'aide du test de Hosmer- Lemeshow. Le modèle sera dit calibré si on ne rejette pas l'hypothèse nulle (i.e. les probabilités théoriques sont proches de celles observées).

```
library(generalhoslem)
logitgof(modele$y, fitted(modele))
```

Remarque : Graphique de calibration !

Le principe de la sélection de variables, ou en d'autres termes le test de l'apport d'une variable ou d'un groupe de variables explicatives dans l'ajustement du modèle, est fondé sur la comparaison du modèle complet avec le modèle dont on a exclu la variable (ou les variables) à évaluer. Dans cette optique, les deux critères habituellement utilisé sont le test du rapport de vraisemblance et le test de Wald.

Test du rapport de vraisemblance entre le modèle réduit à la constante et le modèle retenu :

```
library(lmtest)
modele1 = glm(surv_cancer ~ 1,family = binomial(link="logit"), data = data)
modele2 = glm(surv_cancer ~ age_cl+hsp60+muc1+prdx2+ppia+fkbp52,family = binomial(link="logit"), data = data)
lrtest(modele1,modele2)
```

EXERCICE 1

1/ A partir du modèle complet, essayer d'enlever au fur et à mesure les variables (de façon manuelle) les moins significatives pour déterminer le modèle le plus adéquat

2/ Justifier votre démarche en vous appuyant sur les valeurs des AIC, BIC et test du rapport de vraisemblance à chaque étape de la modélisation

CONSTRUCTION DE MODELES

Plusieurs façons de sélectionner (de façon automatique) le meilleur modèle :

- Backward : On part de toutes les variables disponibles et on enlève au fur et à mesure les variables non significatives.
- Forward : Le contraire de Backward.
- Both : Dans les deux directions.

En général on s'appuie sur le critère d'Akaïke (AIC) ou de Scharwz (BIC) que l'on souhaite minimiser. Ces critères, à peu de choses près équivalents, traduisent la complexité du modèle au fur et à mesure que l'on rajoute des variables (donc des paramètres à estimer).

Nous allons mettre en œuvre les différentes sélections sur notre exemple.

Modèle réduit à la constante :

```
str_constant = "~ 1"
```

Modèle complet incluant toutes les variables explicatives potentielles :

```
str_all = "~age_cl+hsp60+muc1+prdx2+ppia+fkbp52"
```

Installation package MASS :

```
library(MASS)
```

Modélisation « Forward » :

```
modele_constant = glm(surv_cancer~1,family = binomial(link="logit"), data = data)
```

```
modele_forward = stepAIC(modele_constant,scope = list(lower = str_constant,
upper = str_all),trace = TRUE, direction = "forward", data = data)
summary(modele_forward)
```

Modélisation « Both » :

```
modele_constant = glm(surv_cancer ~ 1, family = binomial(link="logit"), dat
= data)
modele_both = stepAIC(modele_constant , scope = list(lower = str_constant,
upper = str_all),trace = TRUE,direction = "both")
summary(modele_both)
```

EXERCICE 2

1/ Proposer la commande à utiliser pour implémenter une modélisation « Backward »

2/ Comparer les résultats obtenus avec les 3 modélisations

3/ Interpréter les résultats obtenus avec le modèle « Both »

MODELE LOGISTIQUE POLYTOMIQUE

On souhaite à présent modéliser la probabilité de survenue du type de la maladie (`type : contrôle vs non invasif vs invasif`) compte tenu des covariables démographiques (`age`), cliniques (`re`, `rp` et `her2`) et biologiques (`hsp60`, `muc1`, `prdx2`, `ppia` et `fkbp52`).

Installation package nnet :

```
library(nnet)
```

Aide de la fonction multinom

```
? multinom
```

EXERCICE 3

1/ Proposer la commande à utiliser pour implémenter une modélisation « Backward » avec la fonction multinom

2/ Interpréter les résultats