TP3 : Estimation dans le modèle logistique

M2 Statistique, 2019-2020

Statistique mathématique en grande dimension et applications, cours de M. Roquain

Dans tout ce TP, nous considérons le modèle logistique

$$Y \sim \bigotimes_{i=1}^{n} \mathcal{B}(g(X_i^t \beta^\star)),$$

pour $g(u) = e^u/(1+e^u)$ et pour une matrice réelle X de taille $n \times p$. Le paramètre d'intérêt est $\beta^\star \in \mathbb{R}^p$.

Commencer par charger (`dataforTP3` peut être téléchargé sur le site web `http://etienne.roquain.free.fr/teaching.html`)

```
load("dataforTP3")
```

Exercice 1 : Modélisation/prévision de la maladie coronarienne

On cherche à expliquer la présence ou non de la maladie coronarienne (insuffisance cardiaque dû à la diminution du débit sanguin de l'artère coronarienne). Plusieurs facteurs peuvent expliquer l'apparition de cette pathologie, et les données `donneescor` regroupent le cas de plusieurs individus/variables. Plus de détails sont donnés sur la page `https://web.stanford.edu/~hastie/ElemStatLearn//datasets/SAheart.info.txt`.

1) Ajuster un modèle logistique sur les données avec le maximum de vraisemblance, à l'aide de la commande suivante.

```
res <- glm(chd ~ ., family = binomial , data=donneescor)
summary(res)
```

Calculer l'EMV $\hat{\beta}$ pour $\beta^\star$ et en déduire les estimateurs $g(X_i^t \hat{\beta})$ des probabilités $g(X_i^t \beta^\star)$, $1 \leq i \leq n$. Comment les interpréter ?

2) L'`odds` (la côte) d'avoir la pathologie pour un individu ayant les caractéristiques $x \in \mathbb{R}^p$ est définie par

$$\text{odds}(x) = \frac{g(x^t \beta^\star)}{1 - g(x^t \beta^\star)} = e^{x^t \beta^\star}$$

Calculer une estimation des `odds` des 10 premiers individus et interpréter l'`odds` de l'individu $i = 1$. Comparer les `odds` estimés de l'individu $i = 1$ et $i = 5$. Interpréter le résultat.

3) Pour une variable $j \in \{1, \ldots, p\}$, on définit l'`odds ratio (partiel)` la quantité e_j^β. Estimer les `odds ratios` de toutes les variables. Proposer une interprétation de ces quantités pour les variables `famhistPresent` et `ldl`.

4) On cherche à tester l'hypothèse qu'aucune des variables n'est significativement associées à la pathologie. Formaliser l'hypothèse nulle, puis exécuter la commande suivante qui effectue un test de rapport de vraisemblance.

```
res0 =glm(chd ~ 1, family = "binomial", data=donneescor)
anova(res0,res,test="Chisq")
```

Conclure.

5) On cherche à trouver les variables qui sont individuellements associées à la pathologie. Formaliser les hypothèses nulles à tester, puis éxécuter la commande suivante qui effectue un test de Wald (analogue du test de Student).

```
summary(res)$coefficients[,4]
```

Conclure.

Exercice 2 : Limite de la théorie asymptotique

Dans cet exercice on explore le cas de la dimension "modérée" et d'une matrice de design d'entrée i.i.d. $\mathcal{N}(0,1)$:

```
p=100
n=1000
X=matrix(rnorm(n*p),n,p)
```

1) On cherche à mettre en évidence le biais de l'estimateur du maximum de vraisemblance. Générer Y dans le modèle logistique avec les $\beta_j^\star$ tirés de manière i.i.d. selon une gaussien de moyenne 3 et de variance 16. Calculer l'EMV avec la commande suivante:

```
betachap=glm(Y~X+0,family=binomial)$coeff
```

Comparer graphiquement l'EMV $\hat{\beta}$ et la cible $\beta^\star$. Que constate t-on ?

2) Charger le package `glmnet`

```
library(glmnet)
```

et prendre cette fois $\beta_j^\star$ s-sparse (disons $s = 10$) avec des coefficients non-nuls tirés de manière i.i.d. selon une gaussien de moyenne 3 et de variance 16. Générer Y dans le modèle logistique, calculer l'EMV comme précédemment et calculer l'estimateur LASSO cross validé comme suit:

```
cv.fit=cv.glmnet(X,Y,family="binomial",intercept=FALSE)
betaLASSO=coef(cv.fit,s="lambda.min")[-1]
```

L'effet constaté à la question précédente est-il corrigé ?

Exercice 3 : Modélisation/prévision du cancer de la prostate à partir du génome

Considérons les données de grande dimension du cancer de la prostate, que l'on peut obtenir de la façon suivante :

```
library(sda)
data(singh2002)
X=singh2002$x
Y=(singh2002$y=="cancer")
n=dim(X)[1]
p=dim(X)[2]
```

1) Ajuster un modèle logistique avec un estimateur LASSO cross-validé comme suit :

```
cv.fit=cv.glmnet(X,Y,family="binomial")
betachap=coef(cv.fit,s="lambda.1se")[-1]
```

2) Un nouvel individu a apporté ses données génomiques personnelles dans le vecteur `newx` (dans `dataforTP3`). Estimer la probabilité qu'il contracte le cancer de la prostate et son `odds`.