TP5 : Test multiple dans le modèle linéaire gaussien

M2 Statistique, 2019-2020

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

Considérons le modèle linéaire gaussien

$$Y = X\beta^\star + \varepsilon,$$

pour ε_i, $1 \leq i \leq n$, i.i.d. $\mathcal{N}(0, \sigma^2)$ (σ inconnue) et pour une matrice réelle X de taille $n \times m$ de telle sorte que X^tX soit de plein rang ($n \geq m$). Le paramètre d'intérêt est $\beta^\star \in \mathbb{R}^m$.

Commencer par télécharger le fichier `dataforTP5` sur la page web `http://etienne.roquain.free.fr/teaching.html`. Charger des jeux de données `y` et `X` à l'aide de la commande.

```
load("dataforTP5")
n=dim(X)[1]
m=dim(X)[2]
plot(y,pch=16,col=gray(0.6),xlab="",ylab="y")
image(t(X),col=c(0,1))
```

Pour plus de détail sur ce jeu de données, consulter la page web

`https://web.stanford.edu/group/candes/knockoffs/software/knockoff/tutorial-4-r.html`

Exercice 1 : Contrôle du FWER

1) Faire la régression linéaire de Y sur X et calculer

$$T_j = \frac{\hat\beta_j}{\hat\sigma[(X^tX)^{-1}]_{j,j}^{1/2}},\ 1 \leq j \leq m$$

et les p-values associées à l'aide des commandes

```
fit=lm(y~X+0)
stat=summary(fit)$coefficients[,3]
pvalues=summary(fit)$coefficients[,4]
```

Représenter les statistiques de tests et les p-values. Combien de p-values sont plus petites que $\alpha = 0.2$. S'agit-il des mutations associées à la résistance au traitement ?

2) Appliquer la correction de Bonferroni au niveau $\alpha = 0.2$. Quelle garantie a t-on sur les variables sélectionnées ?

3) On se propose d'appliquer la procédure `maxT` en utilisant une approximation de Monte-Carlo avec $B = 1000$.

a) Commencer par calculer la matrice

$$M_{j,j'} = \frac{[(X^tX)^{-1}]_{j,j'}}{[(X^tX)^{-1}]_{j,j}^{1/2}[(X^tX)^{-1}]_{j',j'}^{1/2}}$$

à l'aide de la commande

```
Gamma = solve(t(X)%*%X)
M = diag(1/sqrt(diag(Gamma)))%*%Gamma%*%diag(1/sqrt(diag(Gamma)))
```

b) On rappelle que, sous $\beta = 0$, la loi de $(T_j)_{1\leq j\leq m}$ correspond à une gaussienne multivariée centrée de matrice de covariance M, divisée par $\sqrt{\chi^2(n-m)/(n-m)}$. Utiliser la commande `mvrnorm` du package `MASS` qui permet de générer $B = 1000$ Gaussiennes multivariées avec la matrice de covariance Γ. En déduire un B-échantillon pour la loi de $\max_{1\leq j\leq m} |T_j|$ lorsque $\beta = 0$.

c) En déduire le seuil s_α du test multiple `maxT` au niveau $\alpha = 0.2$.

d) Comparer le seuil de Bonferroni et celui `maxT`, ainsi que le nombre de variables découvertes par les deux procédures. Commenter.

4) Refaire l'étude pour un Y (prendre une lette différente de y!) simulé selon le modèle $\sigma = 1$, pour un $\beta_j = 4$ pour $1 \leq j \leq 30$ et $\beta_j = 0$ sinon.

Exercice 2 : Contrôle du FDR

On reprend les données du précédent exercice (à recharger si besoin).

1) Représenter les p-values ordonnées à l'aide du code suivant.

```
fit=lm(y~X+0)
stat=summary(fit)$coefficients[,3]
pvalues=summary(fit)$coefficients[,4]
sortpvalues=sort(pvalues)
plot(sortpvalues,xlab="",ylab="",pch=16,col=gray(0.5),cex=0.4)
```

Pour $\alpha = 0.2$, ajouter le seuil non corrigé ($t = \alpha$), le seuil de Bonferroni, la droite de Benjamini-Hochberg (BH) et de Benjamini-Yekutieli (BY).

2) Calculer les seuils de BH et de BY puis leur ensembles de rejets. Comparer à la procédure de Bonferroni.

3) Quelle procédure utiliser ?