Année 2017-2018 UNIVERSITÉ PAUL SABATIER M1 SID
Modélisation Les réponse au question en gras doivent être dans votre compte rendu en pdf

tout se trouve sous https ://synapse.math.univ-toulouse.fr/s/caTsW6vfCZIkoiH

Exercice 1 le jeux de donnée représente des tensions moyenne de groupe femme d'âge différent. Faire un régression linéaire simple de la tension sur l'âge en tapant le code suivant sous R

```
age = c(35,45,55,65,75)
tension=c(114,124,143,158,166)
Tens=data.frame(age,tension) # lecture des données

reg = lm( tension~age,data=Tens) # déclare le modèle
 summary(reg )
 anova(reg)
 plot(age,tension)
 abline(reg).
```

Répondre aux questions suivantes : **qu'elle l'estimation de σ ? ; age est il significatif ? ; quelle est la p-value ?, quelle est la tension estimée d'un femme de 40 ans ? le modèle vous parait il correct ?**

Adaptez le code pour faire un régression quadratique et interprétez.

Exercice 2. Analyse de la variance
Copiez les données et notez bien leur emplacement. Exécutez le code suivant

```
   dat = read.table("foret.txt", header=TRUE) # les noms des variables sont en première l
   dat$foret = as.factor(dat$foret) #déclare  le numéro de la foret comme un facteur

    lm.foret lm(hauteur~foret,dat)
   library(car) #appele une bibliothèque particulière peut ne pas fonctionner
   Anova(lm.foret,Type="III")  # technique


  coeff.foret = lm(hauteur~foret,dat)
summary(coeff.foret)
   par(mfrow= c(2,2))
   plot(lm.foret,las=1)
```

On va maintenant sortir des estimations dans un forme agréable pour cela il faut supprimer l'intercept

Interprétez : les forêts sont elle différentes ? (p-value) , quel est le nombre de d'arbres dans chaque forêt ?Le graphique des résidus correspond il bien aux hypothèses ?

Calculez la hauteur moyenne et retrouvez la dans les deux paramétrisations.

Exercice 3. Analyse de la variance à deux facteurs Pour étudier le pouvoir désinfectant de divers produits sur les racines de dents, on a construit l'experience suivante.

On prend des dents appartenant à diverses vaches (vache), et on les découpe en 4 morceaux. Chaque élément de dent est ensuite volontairement contaminé par un type de germe (germe) puis désinfecté par un type de désinfectant (trait). On note alors le nombre moyen de germes par spot d'observation microscopique après transformation en log (LNBAC). On note également l'age de la dent (age) et le nombre de spots d'observations (nobs) .

lires les données correspondant à des désinfections de dents par

```
dents=read.table("dents.txt",header=TRUE)
attach(dents)
germe=as.factor(germe)
trait=as.factor(trait)
summary(dents)
```

On va se limiter à l'explication de LNBAC par les deux facteurs germe et trait.

On fait une première analyse en faisant apparaitre le facteur croisé germe :trait

```
dents.lm=lm(LNBAC~germe:trait-1)#sans l'intercept pour avoir les bons estimateurs
summary(dents.lm)
plot(dents.lm$fit,dents.lm$res)
interaction.plot(trait,germe,LNBAC,fixed=TRUE,col = 2:3,leg.bty = "o")
interaction.plot(germe,trait,LNBAC,fixed=TRUE,col = 2:3,leg.bty = "o")
```

Faites une vraie analyse de la variance à deux facteurs en obtenant les tests des deux facteurs (germe*trait) Demandez les estimateurs dans la parametrisation classique et faites le lien avec la parametrisation précédente.

Exercice 4 : Un exemple jouet de regression logistique

```
> X <- factor(c("g1","g2","g3","g1","g2","g1"))
> Y <- factor(c(1,1,1,1,0,0))
> model <- glm(Y~X,family=binomial)
>summary model
```

Faites la table de X et Y. Décrivez complement la paramétrisation.
Rajoutez des données pour avoir un estimation non dégénérée.

Exercice 5. Régression logistique taux de travail de femmes aux USA

Les données (Jobson 1992) étudiées sont issues d'une enquête réalisée auprès de 200 femmes mariées du Michigan. Les variables considérées sont les suivantes :

THISYR, la variable à expliquer, (Woui) si la femme travaille l'année en cours, (Wnon) sinon ;

CHILD1 code la présence (Boui) ou l'absence (Bnon) d ?un enfant de moins de 2 ans ;

CHILD2 présence (Eoui) ou absence (Enon) d'un enfant entre 2 et 6 ans ;

ASCEND l ?ascendance noire (Anoi) ou blanche (Abla) ;

les autres variables, âge (AGE), nombre d ?années d ?études (EDUC), revenu du mari (HUBINC) sont quantitatives.

Les données sont disponibles dans le fichier jobpanel.dat.

Lire le fichier puis recoder les facteurs (exécutez ce script sans chercher à comprendre).

```
# Lecture des données:
type=c("character","character","numeric","numeric",
"numeric","character","character","character")
panel=read.table("jobpanel.dat",colClasses=type,
header=TRUE)
# Codage explicite des facteurs
panel[,"THISYR"]=factor(panel[,"THISYR"],
levels=c("0","1"),labels=c("Wnon","Woui"))
panel[,"CHILD1"]=factor(panel[,"CHILD1"],
levels=c("0","1"),labels=c("Bnon","Boui"))
panel[,"CHILD2"]=factor(panel[,"CHILD2"],
levels=c("0","1"),labels=c("Enon","Eoui"))
panel[,"ASCEND"]=factor(panel[,"BLACK"],
levels=c("0","1"),labels=c("Abla","Anoi"))
panel=panel[,-c(2,6)]
summary(panel)
```

Exploration Uni-dimensionnelle Vérifier les distributions des variables quantitatives, justifier la transformation.

```
hist(panel[,"HUBINC"])
panel[,"LHUBINC"]=log(1+panel[,"HUBINC"])
hist(panel[,"AGE"])
hist(panel[,"EDUC"])
```

Faire la regression logistique par :

```
panel.glm=glm(THISYR~LHUBINC+AGE+EDUC+CHILD1+
CHILD2+ASCEND,family=binomial,data=panel)
summary(panel.glm)
```

Interprétez les valeurs de logit(proba). Attention aux contraintes.

On pourra faire da la regression logistique sparse par l'utilisation de glmnet que l'on vous laisse découvrir

```
  library(glmnet)
res.glmnet <- glmnet(x=...., y=THISYR, family="binomial", alpha=1)
plot(res.glmnet, label=TRUE).
```

Exercice 6.

Courbes ROC

Essentiellement en medical la courbe ROC (Receiving Operating Characteristics) permet de gérer un compromis sensibilité-spécificité d'un test. On considère une maladie M et un test T qui veut la détecter

Si f est la prévision de la maladie issue du modèle logistique ou autre, on peut utiliser la règle suivante qui dépend d'un paramètre f_0 pour construire le test T
Si $f > f_0$ le sujet est déclaré positif on note cet événement T (le test est positif)
Si $f < f_0$ le sujet est déclaré négatif on note cet événement $\bar{T}$

On considère tous les points de l'échantillon qui sont atteint de la maladie M ou sains $\bar{M}$
En fonction de f_0 ils se répartissent en 4 catégories.

1. les vrais positifs MT
2. les faux positifs $\bar{M}T$
3. les vrais négatifs $\bar{M}\bar{T}$
4. les faux négatifs $M\bar{T}$.

on definit la **sensibilité** la proba que le test soit positif sur des sujet atteints et qui est estimée par

$$Se = \frac{\#\{vraispositif\}}{\#\{M\}}$$

et la spécificité la proba que le test soit négatif sur des sujet sains et qui est estimée par

$$Se = \frac{\#\{vraisnegatifs\}}{\#\{\bar{M}\}}$$

la courbe ROC représente en abcisse l 'estimation de la non spécificité et en ordonnée l'estimation de la sensibilité. C'est un courbe au dessus de la diagonale.

On l'obtient sous R dans le cas précédent par

```
   library(Deducer)
rocplot(panel.glm)
```