Université De Lille
Faculté des Sciences et Technologies
Département de Mathématiques

Compte rendu du TP

Classification supervisée

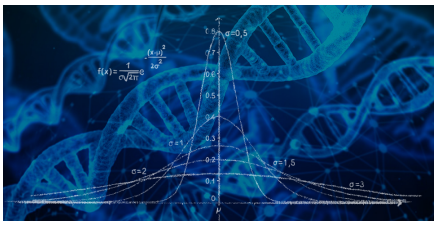
Régression logistique

Réalisé par :

- Abdelmonssif *Oufaska*
- Kamal *Samnouni*

Enseignant :

- *Prof.* Charlotte *Baey*



Année Universitaire :
2020/2021
23 octobre 2020

Introduction

Dans ce TP, on va s'intéresser à évaluer et caractériser les relations entre une variable de type binaire (par exemple : Malade / Non malade) ou qualitative à trois modalités, et une ou plusieurs variables explicatives. On étudiera pour cela deux bases données.

• La première provient d'une étude sur le diabète, dont l'objectif est de décrire et prédire la présence de diabète chez un patient en fonction de certaines caractéristiques cliniques.

• La deuxième provient d'une étude sur la contraception effectuée chez 1473 femmes Indonésiennes, dont le but de prédire le mode de contraception.
Nous verrons comment l'apport de la régression logistique tient dans l'interprétation des résultats du modèle.

Les packages utilisés dans ce tutoriel sont les suivants :
—**nnet**
—**broom**
—**gtsummary**
—**GGally**
—**forestmodel**
—**effects**
—**ggeffects**
—**ggplot2**
—**pROC**
—**MASS**

Prédiction du diabète

L'objectif ici est d'identifier les facteurs de risques associés à la présence du diabète et d'établir un modèle de prédiction capable de prédire l'existence du diabète pour une nouvelle personne donnée avec un minimum de risque de se tromper. La base de données que nous avons a huit variables ont été mesurées ou relevées chez 768 patients :
—**Pregnancies** : Nombre de fois enceinte
—**Glucose** : Concentration en glucose plasmatique
—**BloodPressure** : Pression artérielle diastolique (mm Hg)
—**SkinThickness** : Épaisseur du pli cutané des triceps (mm)
—**Insulin** : Insuline sérique de 2 heures (mu U / ml)
—**BMI** : Indice de masse corporelle (poids en kg / (taille en m^2))
—**DiabetesPedigreeFunction** : Fonction pedigree du diabète
—**Age**
Et une variable **Outcome** de type binaire indique la présence ou l'absence du diabète chez les patients (où 1 revient à dire que le patient a le diabète et 0 donne que le patient n'a pas le diabète)
Voici un aperçu de ces données :

	Pregnancies <int>	Glucose <int>	BloodPressure <int>	SkinThickness <int>	Insulin <int>	BMI <dbl>	DiabetesPedigreeFunction <dbl>	Age <int>	Outcome <int>
1	6	148	72	35	0	33.6	0.627	50	1
2	1	85	66	29	0	26.6	0.351	31	0
3	8	183	64	0	0	23.3	0.672	32	1
4	1	89	66	23	94	28.1	0.167	21	0
5	0	137	40	35	168	43.1	2.288	33	1
6	5	116	74	0	0	25.6	0.201	30	0

6 rows

La question qui se pose ici est :
Est ce que on devrait mesurer toutes ces caractéristiques cliniques pour prédire l'existence du diabète ?

Pour répondre à cette question on doit identifier les variables significatives i,e les variables qui portent le maximum possible de l'information sur le diabète, c'est une opération qui va nous aider à établir un modèle simple de prédiction par un minimum possible des variables explicatives.
Nous allons d'abord nous intéresser à l'analyse descriptive de notre base de données. D'après les sorties du logiciel R à l'aide de la fonction summary, on constate que l'âge des patients est entre 21ans et 81ans pour une moyenne de 33ans, et les autres variables se jouent entre 0.0 et 199.0 pour "Glucose" pour une moyenne de 120.9, entre 0.0 et 122.0 pour "BloodPressure" pour une moyenne de 69.11, entre 0.0 et 99.0 pour "SkinThickness" pour une moyenne de 20.54, entre 0.0 et 846.0 pour "Insulin" pour une moyenne de 79.8, entre 0.0 et 67.1 pour "BMI" pour une moyenne de 31.99, entre 0.078 et 2.42 pour "DiabetesPedigreeFunction" pour une moyenne de 0.4719 et entre 0 et 17 pour "Pregnancies" pour une moyenne de 3.845. On peut noter également que 268 des patients ont une présence du diabète d'après la variable "Outcome" et 500 des patients ont pas une présence du diabète.

Ce graphique nous représente la boîte à moustaches des âges des deux groupes (groupe 1 : Outcome=0 et groupe 2 : Outcome=1) indique que l'âge médiane du premier groupe est de 27ans et celle du deuxième groupe est de 36ans. En revanche l'âge de la plupart des patients qui sont atteints du diabète est située entre 28ans et 48ans.

Maintenant, on souhaite ajuster un modèle de régression logistique qui est l'objet de notre TP à l'aide de la fonction glm(), On veut expliquer la variable "Outcome" en fonction des autres variables.
On note log.diab1 le modèle trouvé par la fonction glm, on a comme résultat toutes les variables sont significatives sauf "Age", "Insulin" et "SkinThickness". Pour valider ces résultats on propose d'augmenter la valeur d'une variable non significative, et on laisse les autres pour voir l'impact de cette augmentation sur le risque d'être atteint du diabète.
On prend par exemple une augmentation de 10ans pour la variable "Âge", pour cela on multiplie le coefficient de la régression logistique β_{Age} lié à la variable âge par 10 afin de calculer l'odds-ratio.

$$\text{car } OR_{01}(x, x^{'}) = e^{\beta_{Age}(x^{'}_{Age} - x_{Age})}$$

Dans ce cas on trouve que le risque d'être atteint du diabète est 1.160313 ce qui est né-

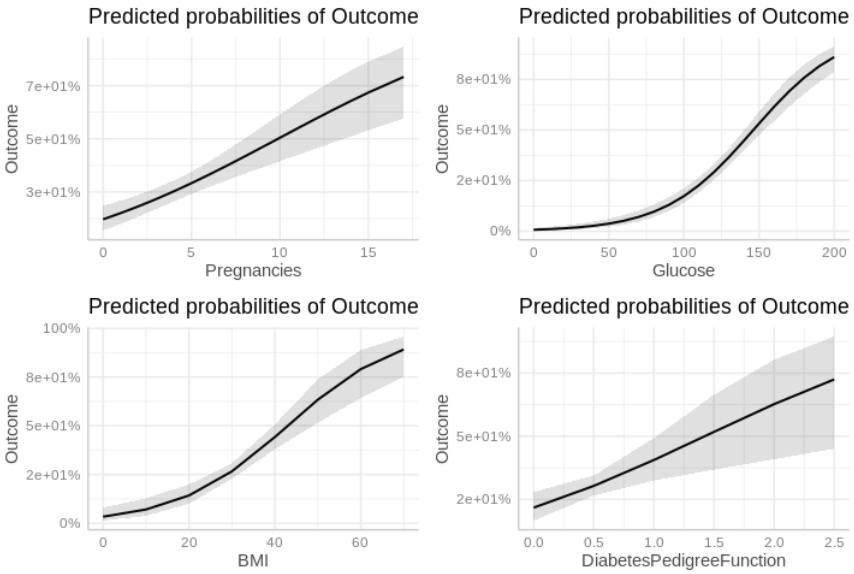
gligeable. De même respectivement pour "Insulin" et "SkinThickness", on propose une augmentation de 30mIU/L et 1mm(l'unité de SkinThickness dans la base est multiplié fois 10), on trouve le risque d'être atteint du diabète est 0.9648805 et 1.006209 également ce qui revient à dire que les trois variables ont aucun impact sur le diabète.
En revanche, si on fait une petite augmente de 0.2g/L (équivalent à une augmentation de 20 dans le tableau) de "Glucose" qui est significative, on trouve le risque d'être atteint du diabète est 2.020357, le risque est multiplié par 2. On fait la même chose pour les autres variables significatives respectivement pour "Pregnancies", "BloodPressure", "DiabetesPedigreeFunction" et "BMI", on propose une augmentation de 5, 20mmHg, 1.5 et 10 kg/m^2, on trouve le risque d'être atteint du diabète est 1.851343, 0.8755041, 4.127903 et 2.452259. On constate que les variables qui agissent comme des facteurs de risque sont "Glucose", "Pregnancies", "DiabetesPedigreeFunction" et "BMI".

On souhaite maintenant proposer une méthode pour construire un modèle réduit contenant moins de variables, pour cela on va procéder comme suit.
Notre premier modèle indique l'existence de trois variables non significatives !. Que se passe t-il si on retire une variable significative ? On retire par exemple la variable "Glucose" et on construit un nouveau modèle log.diab2 par la fonction glm on remarque que toutes les variables sont significatives sauf la variable "SkinThickness" ! donc quel est le modèle le plus pertinent ? la fonction stepAIC de la librairie MASS revient à rependre à cette question, c'est une fonction qui choisit le meilleur modèle le plus adapté au variable prédite avec un minimum de variables on trouve qu'on peut exprimer la variable binaire "Outcome" par quatre variables seulement : "Glucose", "Pregnancies", "BMI" et "DiabetesPedigreeFunction".
Ce tableau ci-dessous résume les résultats précédents :

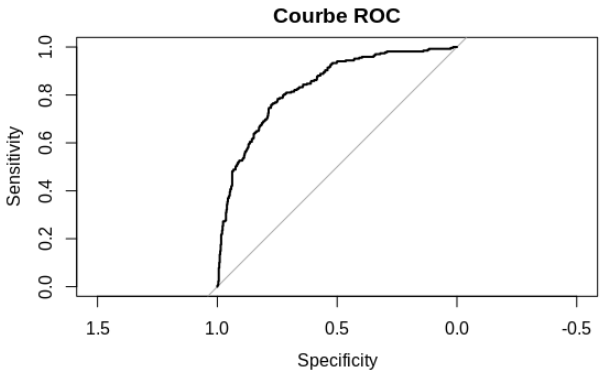
Variable	Valeur d'augmentation	Valeur du risque	Facteur du risque
Insulin	30mIU/L	0.9648805	Non
SkinThickness	1mm	1.006209	Non
Age	10ans	1.160313	Non
Pregnancies	5	1.851343	Oui
BloodPressure	20mmHg	0.8755041	Non
DiabetesPedigreeFunction	1.5	4.127903	Oui
BMI	10 kg/m^2	2.452259	Oui

TABLE 1 – Effet d'une augmentation de la valeur des variables

Maintenant, on veut visualiser l'effet de chaque variable sur la variable à prédire. D'après notre étude on va s'intéresser à l'effet des variables qui agissent comme des facteurs de risque. On a le graphique suivant :
On constate que les quatre variables ont un effet important sur la variable à prédire, tel que la probabilité d'être atteint du diabète augmente avec une petite augmentation de ces variables



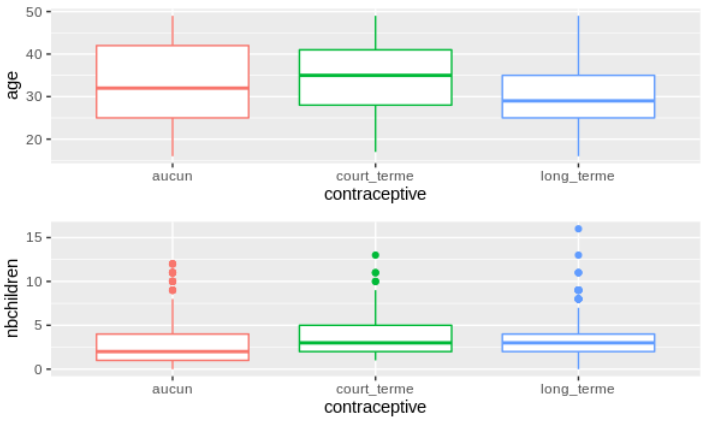
On propose de tracer la courbe ROC, on a le graphique à coté, à l'aide du logiciel R on trouve que le seuil optimal qui mesure la qualité du classement des prédictions est $s = 0.83$, donc pour prédire avec une grande performance la classe d'une nouvelle personne avec le modèle précèdent on doit fixé le seuil à 0.83.



Prédiction du mode de contraception

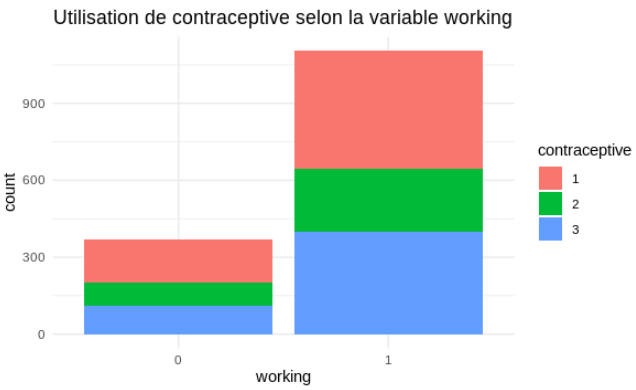
Dans cette partie on va s'intéresser à la prédiction d'une variable qualitative à trois modalités. L'objectif est d'identifier les facteurs influençant la prise d'une contraception, et étudier les différences éventuelles entre les types d'une contraception.
La base de données qui nous intéresse provient d'une étude chez 1473 femmes Indonésiennes. Dix variables ont été mesurées ou relevées sont les suivantes :
— **age** : l'âge en années
— **education** : le niveau d'éducation (codé de 1 : faible à 4 : élevé)
— **husband-education** : le niveau d'éducation du mari (codé de 1 : faible à 4 : élevé)
— **nbchildren** : le nombre d'enfants
— **religion** : la religion de la femme (1 : mulsumane, 0 : autre)
— **working** : est-ce que la femme travaille (1 : oui, 0 : non)
— **husband-occupation** : le niveau d'occupation du mari (codé de 1 : faible à 4 : élevée)
— **standard-of-living** : le niveau de vie du ménage (codé de 1 : faible à 4 : élevé)
— **media** : l'exposition aux médias (1 : oui, 0 : non)
— **contraceptive** : le type de contraception (1 : aucune, 2 : court-terme, 3 : long-terme)
Nous allons d'abord faire une analyse descriptive de la base.

On trouve que l'âge des femmes est entre 16ans et 49ans pour une moyenne de 32ans, et le nombre d'enfant se joue entre 0 et 16. D'après le graphique à coté représente un Boxplot pour les deux variables "age" et "nbchildren", on remarque que l'âge et le nombre d'enfant ont pas un grand effet sur le type de contraception. Est ce qu'on peut avoir confiance dans ces résultats ?

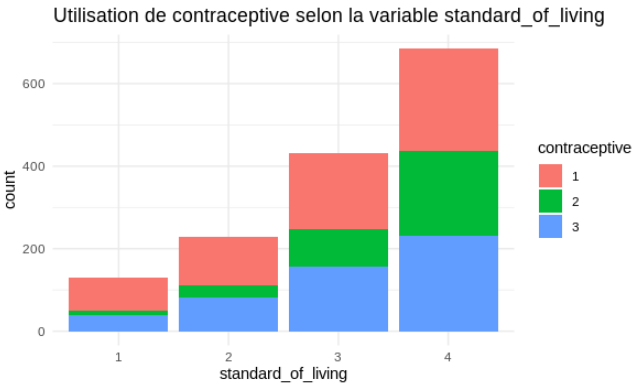


On peut noter également que 152 des femmes ont une niveau d'éducation faible, 334 niveau passable, 410 niveau bien et 577 niveau élevé. 44 des femmes le niveau d'éducation du mari est faible, 178 niveau passable, 352 niveau bien et 899 niveau élevé. 1253 des femmes sont des musulmanes et 220 ont autre religion. 436 des femmes le niveau d'occupation du mari est faible, 425 niveau passable, 585 niveau bien et 27 niveau élevé. 1364 des femmes ont pas l'exposition aux médias et 109 ont l'exposition aux médias.

On a aussi que 1104 des femmes elles ont un travail et 369 ont pas de travail. 129 des femmes leur niveau de vie du ménage est faible, 229 est du niveau passable, 431 est du niveau bien et 684 est du niveau élevé. Le graphique ci-dessous représente l'utilisation de contraception selon les deux variables "standard-of-living" et "working".



On constate que le fait que la femme travaille ou non n'a pas d'effet sur le type de contraception car le nombre des femmes qui ne travaillent pas et n'utilisent pas de contraception est proportionnel au nombre de femmes qui travaillent et n'utilisent pas de contraception, au contraire si on compare les femmes qui ont un niveau faible et les femmes qui ont un niveau élevé on remarque que les valeurs ne sont pas proportionnels, donc on peut dire qu'il y a un effet sur la contraception.

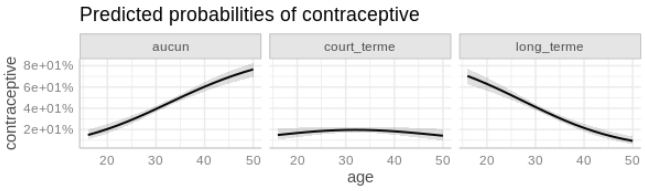


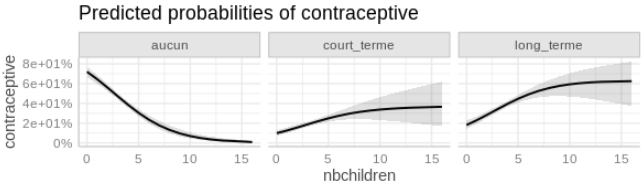
On peut noter également que 629 des femmes ont pas de type de contraception, 333 ont un type court-terme et 511 ont un type long-term.

Maintenant, on souhaite ajuster un modèle de régression logistique multinomial, à l'aide de la fonction "multinom" de la librairie "nnet". On veut expliquer la variable "contraceptive" en fonction des autres variables, Quelles sont les variables qui ont un grand effet sur l'utilisation de contraceptif ?

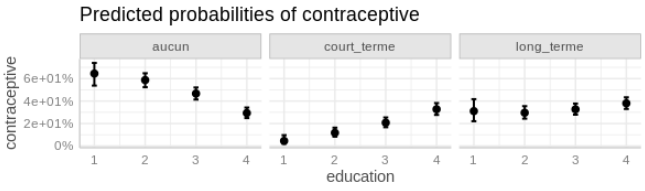
Les graphes ci-dessous représentent les odds-ratio des variables explicatives

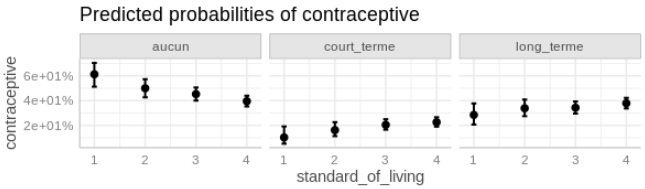
On constate que l'âge et le nombre d'enfants ont un effet sur l'utilisation de contraception tel que, plus que l'âge des femmes augmente, elles n'utilisent pas des contraceptifs au contraire plus que les femmes ont beaucoup d'enfants elles utilisent des contraception de long-terme. ils sont deux résultats qu'on n'a pas pu extraire dans le Boxplot mais grâce à odds-ratio on remarque les effets de ces deux variables sur la contraception.



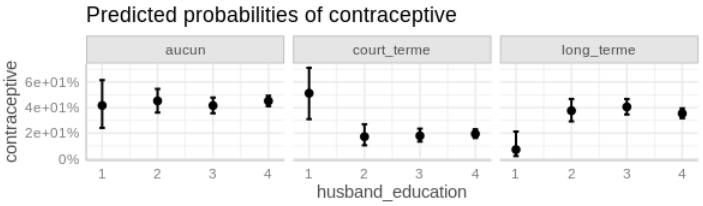


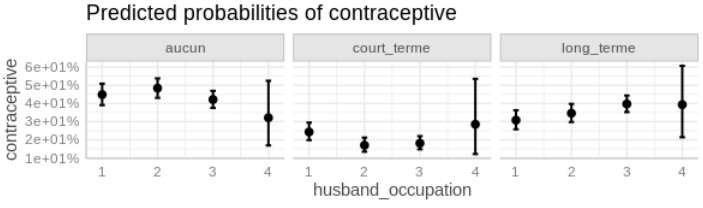
Ici le niveau d'éducation de la femme affecte sur l'utilisation de la contraception plus que les femmes ont un niveau élevé d'éducation, elles utilisent les contraceptifs, la même chose pour le niveau de vie du ménage



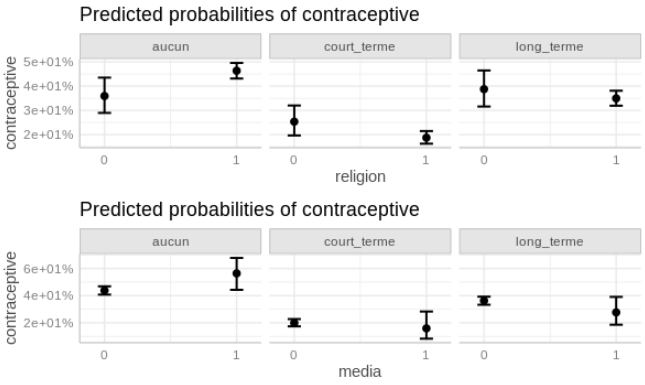


Le niveau d'éducation du mari n'affecte pas sur l'utilisation des contraceptifs, mais on constate que la probabilité d'utiliser les contraceptions de court-terme est plus grande que celle de long-terme quand l'homme a un niveau d'éducation faible, au contraire des autre niveaux. Pour le niveau d'occupation du mari lorsque c'est élevé les femmes utilisent des contraceptifs et la probabilité d'utiliser les contraceptions de long-terme est plus grande que celle de court-terme pour les quatre niveaux.





On remarque que les femmes qui sont exposées aux médias et qui sont musulmanes ont une tendance de ne pas utiliser les contraceptifs, et la probabilité d'utiliser les contraceptions de long-terme est plus grande que celle de court-terme.



Annexe : Code R

```r

library(nnet)
library(broom)
library(ggplot2)
library(GGally)
library(forestmodel)
library(ggeffects)
library(effects)
library(MASS)
library(pROC)
library(cowplot) # graphe combiner
library(forcats) # facteur


############################################## Exercice 1 : Pr diction du diab te

#Importer la base de donn es
diab <- read.csv("/home/samnouni/Bureau/M2 ISN/TP R/diabetes.csv",sep=",", header=
    TRUE)


#Neufs variables ont   t   mesur es ou relev es chez 768 patients

# faire une analyse descriptive
str(diab)
head(diab)
summary(diab)

# Graphe des patients ayont le diab te par age
diab$Outcome=as.factor(diab$Outcome)
ggplot(diab)+aes(x = Outcome, y = Age)+geom_boxplot()+xlab("Outcome")+ylab(" ge ")+
    ggtitle("R partition par  ge  selon l'absence ou la pr sence du diab te")


# R gression logistique
log.diab=glm(Outcome~.,data= diab, family=binomial)  # Premier modele
summary(log.diab)  #Afficher les r sultats du mod le
coef.reg = coef(log.diab)  # pour savoir les beta chapeaux estim s
ci.reg=confint(log.diab) # intrvalle de confiance pour les coefficients

exp(coef.reg) # pour r cuperer les odds ratio
exp(ci.reg) # l'intervalle de confiance des odds ratio
cbind(exp(coef.reg),exp(ci.reg))


#Effet d'une augmentation de l'age de 10
```

```
coef.age=coef.reg["Age"]
coef.age.plu10=10*coef.age
or.age.plus10=exp(coef.age.plu10)
ci.age.plus10=10*ci.reg["Age",]
ci.OR.age.plus10=exp(ci.age.plus10)

print(paste0("OR pour l'Age qui augmente de 10 ans :",or.age.plus10,
             "IC     95% : [",ci.OR.age.plus10[1],";",ci.OR.age.plus10[2]))

# pour choisir le modle le plus pertinent
stepAIC(log.diab, trace = TRUE, data = diab)

log.diab.2=glm(Outcome~Pregnancies+Glucose+BMI+DiabetesPedigreeFunction,data=diab,
    family = binomial) # modle pertinent

#Visualiser  l effet  de chaque variable sur la variable    prdire
g=plot(ggeffect(log.diab.2))
plot_grid(g$Pregnancies,g$Glucose,g$BMI,g$DiabetesPedigreeFunction, ncol = 2, nrow =
    2)

# Tracer la courbe ROC associe au modle
pred.diab=predict(log.diab,type = "response") # probabilite d'etre dans la classe 1.
plot(roc(diab$Outcome,pred.diab)) # Courbe ROC
pred <- prediction(pred.diab, diab$Outcome)
perf <- performance(pred, "auc")
perf@y.values[[1]] # le seuil optimal pour predire Outcome = 1


#####################################  Exercice 2 : Prdiction du mode de
    contraception


#Importer la base de donnes
data<- read.table("/home/samnouni/Bureau/M2 ISN/TP R/cmc.data",sep=",", header=TRUE)

# faire une analyse descriptive
str(data)
head(data)
summary(data)


## Variable en factor
data$education=as.factor(data$education)
data$husband_education=as.factor(data$husband_education)
data$religion=as.factor(data$religion)
data$working=as.factor(data$working)
data$husband_occupation=as.factor(data$husband_occupation)
data$standard_of_living=as.factor(data$standard_of_living)
data$media=as.factor(data$media)
data$contraceptive=as.factor(data$contraceptive)

# Recodage de la variable contraceptive
levels(data$contraceptive)=c( "aucun","court_terme","long_terme")

# refaire une analyse descriptive
str(data)
summary(data)

########## Boxplot Age et nb des enfants
box1= ggplot(data, aes(x=contraceptive, y=age, color=contraceptive)) + geom_boxplot()
```

```
    + theme(legend.position = "none")
box2= ggplot(data, aes(x=contraceptive, y=nbchildren, color=contraceptive)) + geom_
    boxplot() + theme(legend.position = "none")
plot_grid(box1,box2, ncol = 1, nrow = 2)

### Graphe working ##
ggplot(data, aes(x = working , fill =contraceptive)) +geom_bar()

### Graphe standart_of living ##
ggplot(data, aes(x = standard_of_living , fill =contraceptive)) +geom_bar()

# Ajuster un mod le de r gression logistique multinomial,
log.data= multinom(contraceptive~.,data=data)
summary(log.data) # Pour afficher les resultats du modele
coef(log.data)  # pour savoir les beta chapeaux estim s

# l'effet des variables sur la variables    predire
g=plot(ggeffect(log.data))
plot_grid(g$age,g$nbchildren, ncol = 1, nrow = 2)
plot_grid(g$education,g$standard_of_living, ncol = 1, nrow = 2)
plot_grid(g$husband_education,g$husband_occupation, ncol = 1, nrow = 2)
plot_grid(g$religion,g$media ,ncol = 1, nrow = 2)
```