« Etude observationnelle en épidémiologie Clinique »

Contexte :

Le jeu de données à analyser provient d'une étude observationnelle multicentrique (5 centres hospitaliers) réalisée aux Etats-Unis.
Cette étude porte sur 5 735 patients adultes (≥18 ans) admis en unité de soins intensifs (USI) entre Juin 1989 et Janvier 1994, et suivis pendant 6 mois.

La question sous-jacente initiale de cette étude était de tester l'efficacité d'une technique de réanimation (pose d'une sonde de Swan Ganz, ou Right Heart Catheterization (RHC)). En d'autres termes, le gain théorique que cette technique est sensée apporté était-il supérieur aux conséquences potentiellement délétères de ce geste ?

A l'admission en USI, les données sociodémographiques des patients ont été recueillies par questionnaire. Le jour suivant l'admission en USI, différentes caractéristiques cliniques (quantitatives et qualitatives) ont été mesurées.

Le jeu de données comporte 67 variables dont le descriptif est présenté en page suivante.

Variables du jeu de données :

Variable	Type	Libellé
PTID	Texte	Identifiant Patient
Diagnostic à l'admission		
CAT1	Texte	Catégorie principale de maladie
CAT2	Texte	Catégorie secondaire de maladie
CA	Texte	Cancer
RESP	Texte	Respiratoire
CARD	Texte	Cardiovasculaire
NEURO	Texte	Neurologique
GASTR	Texte	Gastro-intestinal
RENAL	Texte	Rénal
META	Texte	Métabolique
HEMA	Texte	Hématologique
SEPS	Texte	Infectieux
TRAUMA	Texte	Traumatique
ORTHO	Texte	Orthopédique
Maladies associées		
CARDIOHX	Num.	Atteinte vasculaire, cardiovasc.
CHFHX	Num.	Crise cardiaque
DEMENTHX	Num.	Démence, infarctus céréb., Parkinson
PSYCHHX	Num.	Psychose, dépression
CHRPULHX	Num.	Atteinte pulmonaire
RENALHX	Num.	Atteinte rénale
LIVERHX	Num.	Cirrhose, atteinte hépatique
GIBLEDHX	Num.	Hémorragie gastro-intestinale Haute
MALIGHX	Num.	Tumeur solide, hémopathie maligne
IMMUNHX	Num.	Immunosuppression, VIH, diabète…
TRANSHX	Num.	Transfert d'un autre hôpital (>24h)
AMIHX	Num.	Infarctus du myocarde
Scores Cliniques		
SURV2MD1	Num.	Probabilité Estimée de survie à 2 mois
DAS2D3PC	Num.	Score DASI (range
APS1	Num.	Score Apache 3
SCOMA1	Num.	Score Glasgow
ADLD3P	Num.	Score ADL

Variable	Type	Libellé
Caractéristiques physiologiques *(dans les 24h)*		
WTKILO1	Num.	Poids (kg)
TEMP1	Num.	Température corporelle (°C)
MEANBP1	Num.	Tension Artérielle (mm Hg)
RESP1	Num.	Fréquence respiratoire (resp/mn)
HRT1	Num.	Fréquence cardiaque (Bpm)
PAFI1	Num.	Ratio PaO2/FIO2 (mm Hg)
PACO21	Num.	PaCo2 (mm Hg)
PH1	Num.	PH
WBLC1	Num.	Leucocytes (10^9 cell./L)
HEMA1	Num.	Hématocrite (%)
SOD1	Num.	Sodium (mmol/L)
POT1	Num.	Potassium (mmol/L)
CREA1	Num.	Créatinine (mg/L)
BILI1	Num.	Bilirubine (mg/L)
ALB1	Num.	Albumine (g/dL)
URIN1	Num.	Diurèse (mL)
Evénements		
SWANG1	Texte	Right Heart Catheterization (RHC)
DEATH	Texte	Décès durant le suivi
Dates		
d_SADMDTE	Num.	Date inclusion étude (jour)
m_SADMDTE	Num.	Date inclusion étude (mois)
y_SADMDTE	Num.	Date inclusion étude (année)
d_DSCHDTE	Num.	Date de sortie de l'hôp. (jour)
m_DSCHDTE	Num.	Date de sortie de l'hôp. (mois)
y_DSCHDTE	Num.	Date de sortie de l'hôp. (année)
d_DTHDTE	Num.	Date de décès (jour)
m_DTHDTE	Num.	Date de décès (mois)
y_DTHDTE	Num.	Date de décès (année)
d_LSTCTDTE	Num.	Date de dernière nouvelle (jour)
m_LSTCTDTE	Num.	Date de dernière nouvelle (mois)
y_LSTCTDTE	Num.	Date de dernière nouvelle (année)
Caractéristiques sociodémographiques		
AGE	Num.	Age (année)
SEX	Texte	Sexe
RACE	Texte	Ethnie
EDU	Num.	Année d'étude (année)
INCOME	Texte	Revenu
NINSCLAS	Texte	Type d'assurance médicale

Descriptif des scores cliniques :

- Le score ADL (Activities of Daily Living), version modifiée de l'index de Katz, évalue la capacité des patients dans la conduite des activités quotidiennes (valeurs entre 0 et 7, 7 indiquant un état grabataire).
- Le score de Duke modifié (DASI), évalue la capacité fonctionnelle (valeurs entre 0 et 33, 33 indiquant une excellente capacité).
- FACTOCHINY cluster sur RSTUDIO

- Le score Apache 3, est un score de gravité synthétisant plusieurs paramètres biologiques, (valeurs entre 0 et 299, les valeurs élevées indiquant un pronostic péjoratif).
- Le score de Glasgow modifié, est un indice évaluant l'état neurologique (valeurs : 0-100, corrélation positive avec l'altération de l'état neurologique).

```
setwd("/Users/mbena/OneDrive/Bureau/R studio/SAE épidémiologie")

epi<-read.table("Epi_Clin.txt", header = T,na.strings = """,sep="\t")

library(dplyr)

library(ggplot2)

library(tidyr)

nrow(epi)

nrow(distinct(epi, PTID, .keep_all = T))

# pas de doublon

View(epi)

#transfomer les va qui prennent 0;1 en facteur

epi$CARDIOHX <- as.factor(epi$CARDIOHX)

epi$CHFHX <- as.factor(epi$CHFHX)

epi$DEMENTHX <- as.factor(epi$DEMENTHX)

epi$PSYCHHX <- as.factor(epi$PSYCHHX)

epi$CHRPULHX <- as.factor(epi$CHRPULHX)

epi$RENALHX <- as.factor(epi$RENALHX)

epi$LIVERHX <- as.factor(epi$LIVERHX)

epi$GIBLEDHX <- as.factor(epi$GIBLEDHX)

epi$MALIGHX <- as.factor(epi$MALIGHX)

epi$IMMUNHX <- as.factor(epi$IMMUNHX)

epi$TRANSHX <- as.factor(epi$TRANSHX)

epi$AMIHX <- as.factor(epi$AMIHX)

epi$AMIHX <- as.factor(epi$AMIHX)

#description des données

summary(epi)

table(epi$CAT1)
```

```
table(epi$CAT2)

table(epi$CA)

table(epi$DEATH)

table(epi$SEX)

table(epi$RESP)

table(epi$CARD)

table(epi$NEURO)

table(epi$GASTR)

table(epi$RENAL)

table(epi$META)

table(epi$HEMA)

table(epi$SEPS)

table(epi$TRAUMA)

table(epi$ORTHO)

table(epi$RACE)

table(epi$INCOME)

# filtrons les données abérrante

table(epi$SEX)

summary(epi$SURV2MD1)

summary(epi$ADLD3P)

summary(epi$DAS2D3PC)

range(epi$APS1)

# %>% PIPE

# création

epi_propre = epi %>%

filter(
```

```
AGE >= 18 & AGE < 120,

SURV2MD1 > 0 & SURV2MD1 < 1,

DAS2D3PC >= 0 & DAS2D3PC <= 33,

APS1 >= 0 & APS1 <= 299,

SCOMA1 >=0 & SCOMA1 <= 100

)

View(epi_propre)

# 233 ligne supprimé

install.packages("lubridate")

library(lubridate)

# création de variable sous la forme date

date_in <- ymd(paste(epi_propre$y_SADMDTE,  epi_propre$m_SADMDTE,

epi_propre$d_SADMDTE, sep = "-"))

View(epi_propre)

date_out <- ymd(paste(epi_propre$y_DSCHDTE,  epi_propre$m_DSCHDTE,

epi_propre$d_DSCHDTE, sep = "-"))

date_mort <- ymd(paste(epi_propre$y_DTHDTE, epi_propre$m_DTHDTE,

epi_propre$d_DTHDTE,

sep = "-"))

epi1 <- cbind(epi_propre,date_in, date_mort,date_out)

names(epi1)[54] <- "Ethnie" # Colonne "RACE" renommé "ETHNIE"

table(epi1$Skin_Color)

View(epi1) ----------------------------------------

setwd("/Users/33695/Desktop/R studio/SAE épidémiologie ")

epi<-read.table("Epi_Clin.txt",header = T,na.strings = """,sep="\t") ------------------------------

------------------------------------------
```

```
library(dplyr); library(ggplot2); library(tidyr)

nrow(epi)

nrow(distinct(epi, PTID, .keep_all = T))

# pas de doublon

# THEME CARACT_SOCIO-DEMOG. --> MEHDI

View(epi1)

# AGE

summary(epi1$AGE) # Analyse univariée ( repère )

sd(epi1$AGE) # Dispersion de 17 ans autour de la /x

table(mean(epi1$AGE))

library(GGall)

#------- Age après tri

summary(epi1$AGE) # Analyse Univariée

sd(epi1$AGE) # Dispersion des patients se rapprochant de la dix-septaines.

table(epi1$AGE)

hist(epi1$AGE, main = "Proportion des patients en fonction de leur âge", breaks = 7, xlab =

"Age",

ylab = "Proportion", probability = TRUE, col = "orange")

colors()

hist(epi1$SURV2MD1,main= "Probabilité estimée de survie à 2 mois", breaks = 7, xlab =

"Probababilités", ylab = "Proportion en fonction des probabilités", probability = TRUE, col =

"red3")

# ------------- SEX après tri

summary(epi1$SEX)

table(epi1$SEX)

round (table(epi1$SEX) / nrow(epi1) * 100, 2) # Analyse unviariée (en %)
```

```
pie(table(epi1$SEX))

title("Proportions des patients en fonction du genre")

pie(table(epi1$SEX))

#PROBLEME VARIABLE SEX --> valeur numérique ==> Problème

#--------------------

#-----------------EDU après TRI

summary(epi1$EDU) # Etude de medicine varie entre 9 à 12 ans en étude et au dela pour les
séniors

#------------------------------------------------------------

boxplot(epi1$EDU, col = grey(0.8), main = "Etudes des patients selon leur pathologies
jusqu'à 30
ans", ylab = "Année d'étude")

abline(h = median(epi_propre$EDU, na.rm = TRUE), col = "navy", lty = 2)

text(1.35, median(epi_propre$EDU, na.rm = TRUE) + 0.15, "Médiane", col = "grey")

Q1 <- quantile(epi_propre$EDU, probs = 0.25, na.rm = TRUE)

abline(h = Q1, col = "darkred")

text(1.35, Q1 + 0.15, "Q1 : 10 ans", col = "darkred", lty = 2)

Q3 <- quantile(epi_propre$EDU, probs = 0.75, na.rm = TRUE)

abline(h = Q3, col = "darkred")

text(1.35, Q3 + 0.15, "Q3 : 13 ans", col = "darkred", lty = 2)

arrows(x0 = 0.7, y0 = quantile(epi_propre$EDU, probs = 0.75, na.rm = TRUE), x1 = 0.7, y1
=
quantile(d$heures.tv, probs = 0.25, na.rm = TRUE), length = 0.1, code = 3)

text(0.7, Q1 + (Q3 - Q1) / 2 + 0.15, "Me(12)", pos = 2)

mtext("L'écart inter-quartile h contient 30 % des patients", side = 1)

abline(h = Q1 - 1.5 * (Q3 - Q1), col = "darkgreen")
```

```
text(1.35, Q1 - 1.5 * (Q3 - Q1) + 0.15, "Borne.inf : 5.5 ans ", col = "darkgreen", lty = 2)

abline(h = Q3 + 1.5 * (Q3 - Q1), col = "darkgreen")

text(1.35, Q3 + 1.5 * (Q3 - Q1) + 0.15, "Borne.sup : 17.5 ans ", col = "darkgreen", lty = 2)

#---------------------------------------------------------------------------

# RACE # Changer le nom de colonne ?

summary(epi1$RACE) # Etude de médecine varie entre 9 à 12 ans en étude et au dela pour les
séniors.

round (table(epi1$RACE) / nrow(epi) * 100, 2) # Analyse unviariée (en %)

table(epi1$RACE)

table(epi1$RACE) " Changer le nom du filtre colonen car autre = Péjorative"

# Desciption brutes des données

#Vérification si la fonction summarytools est bien téléchargé

library(summarytools)

# Descriptions brutes sur l'ensembles des variables

summarytools::descr(epi1)

summary(epi$AGE)

# Description en fréquences des variables

summarytools::freq(epi1)

summarytools::ctable(epi1)

summarytools::dfSummary(epi1)

summary(epi1$NINCLAS) ---------------------------------

" Partie 3.a"

print(epi1$SEX)

epi$SEX_new = ifelse(SEX.length == "Male", "H","F" )

contingence = table(epi1$SWANG1,epi1$DEATH)  #?????? RHC (X) / Décès suivi (Y)

print(contingence)
```

```
# distributions conditionnelles du RHC/ Décès.s en %

prop.table(contingence, 1)*100

# distributions conditionnelles du Décès.s/RHC en %

prop.table(contingence,2)*100

" Partie 3.b"

print(epi1$NINSCLAS)

contingence = table(epi1$NINSCLAS,epi1$DEATH)  #?????? Med.Insurance (X) / Décès

suivi (Y)

print(contingence)

# distributions conditionnelles du MED.Insurrance/Décès.s en %

prop.table(contingence, 1)*100

# distributions conditionnelles Décès.s/Med.Insurrance en %

prop.table(contingence,2)*100

" Partie 3.c"

print(epi1$INCOME)

contingence = table(epi1$INCOME,epi1$DEATH)  #?????? REVENU (x)/ Décès suivi (Y)

print(contingence)

# distributions conditionnelles du Revenu/Décès.s en %

prop.table(contingence, 1)*100

# distributions conditionnelles du Décès.s/Revenu en %

prop.table(contingence,2)*100 ----------------------------

install.packages("GGally")

library(GGally)

GGally::ggpairs(iris, columns = 1:4,

ggplot2::aes(colour = Species))

# Diagramme des proportions cumulés principal ills with seeing death fill.
```

```
library(GGally)
epi1$CAT1 <- forcats::fct_explicit_na(factor(epi1$CAT1))
ggplot(epi1) +
aes(x = CAT1, fill = DEATH) +
geom_bar(position = "fill") +
geom_text(aes(by = CAT1, cex= 1.5), stat = "prop", position = position_fill(.5)) +
xlab("Maladies principaux") +
ylab("Proportion") +
labs(fill = "Décès lors du suivi") +
scale_y_continuous(labels = scales::percent)
#---------------------------------------
# Fct répartition empirique
hist(epi[epi1$AGE == "2011" & epi1$AGE == "Dégagé", "cnt"],
breaks = 13, freq = FALSE,
main = "Densité empirique du nombre de locations entre 17h et 18h en 2011 par temps
dégagé",
xlab = "nombre de locations entre 17h et 18h", ylab = "densité")
densite <- density(epi[epi1$AGE == "2011" & epi1$AGE == "Dégagé", "cnt"])
lines(densite)
# Test ki2
#Hypothèse
"H0 : Les 2 variables sont indépendantes"
"H1 : Les 2 variables sont dépendantes"
chisq.test(epi1$SWANG1,epi1$DEATH)
"D'après notre Hypothèse H1, on remarque que p-valeur 1e-4 < 5%
```

autrement dit que la probabilité de survie de 2 mois à un risque sur le décès suivis des patients.

La proba 2 mois influence sur le décès donc dépendance"

chisq.test(epi1$NINSCLAS,epi1$DEATH)

"Idem, l'assurance maladie à un risque les patients meurt au cours du suivi

p_valeur = 2.2 e-16"

"""Proba décès 2 mois on fait l'hypothèse si Proba > certains seuil alors codé probabilité très forte

TCD 2 mois ou non 2 mois

décès 2 mois associé au femmes ?"""

#? Test regression linéaire (x= v.a indépendante \ y = v.a dépendante)

y = a + bx

data.x.y <- subset(epi1,select = c("SURV2MD1","AGE"))

summary(data.x.y)

cor(data.x.y)

plot(data.x.y, xlab = "Probabilité de survie", ylab= "Age des patients",

main= "Probabilité de survie en fonction de l'AGE")

sat.mod1 <- lm(AGE ~ SURV2MD1, # Regression formule

data = epi1) # data set

summary(sat.mod1) # Montre la table des coefficients de régression

"Le modèle est singificativmeent significative

SURV2MD1 coefficient relié à notre variable indépendante

On constate que les 2 coefficients sont significatifs

La variable estimée par la variable de probabilité estimée à 2 mois est < 0 qui est surprenant

ainsi, on se demande pourquoi il ne faudrait pas mieux introduire d'autres variables

dans le modèle pour pouvoir exprimer réelement la variable SURV2MD1 choisi

```
R-squared:  0.07597"

# Partie 2 : Nuage de point et droite de régression

install.packages("scatterplot")

??scatterplot

Model <- lm(formula= AGE ~ SURV2MD1, data= epi1)

summary(Model)

confint.default(Model)

attributes(Model)

Model.epi1$residuals

Model$coefficients

?scatterplot(AGE ~ SURV2MD1, data=epi1, xlab = "Age",

ylab= "Probabilité de survie de 2 mois", main= "Regression Probabilité de survie en

fonction de l'AGE des patients",

regline=TRUE, ellipse= FALSE, smooth= FALSE, grid = TRUE, boxplots=FALSE)

summary(Model)

confint.default(Model)

# Affichages :)

g = ggplot(diamonds, aes(carat, price))+geom_point() +geom_smooth(method = "lm",

se = FALSE)

##################################### Modèle 1

#################################

reg1 <- lm( AGE ~ SURV2MD1, data= epi1)

plot(reg1)

AIC(reg1)  # AIC =>

##################################### Modèle 2

#################################
```

```
reg2 <- lm( AGE ~ SURV2MD1, data= epi1)

plot(AIC)

AIC(reg2)

##################################### Modèle 3

#################################
reg3 <- lm( AGE ~ SURV2MD1, data= epi1)

plot(AIC)

AIC(reg3)

# Réglage supplémentarie de la qualité du modele :

# Critère 1 : Corbe residuels vs valeurs predits par modele

plot(Model)

# Critère 2 : QQ-plot

plot(Model)

# 4x4 Matrice de corrélation

library(GGally)

ggpairs(rp99[, c("hlm", "locataire", "jeux","SURVIE")], aes(colour = SURVIE))

#

ggpairs(epi1, columns = 3:7,

title = "Analyse bivariée des dépenses de recettes par le ménage britannique",

upper = list(continuous = wrap("cor", taille = 3),

mappage = aes(color = SWANG1, shape = DEATH) ,

alpha = 0,3, taille = 0,1))

# Matrice de corrélation

library(ggplot2)

ggpairs(epi, columns = 5:7,aes(color = DEATH, shape = SEX))

# Horloge répartition de l'âge
```

```
x <- c(15, 9, 75, 90, 1, 1, 11, 5, 9, 8, 33, 11, 11,
20, 14, 13, 10, 28, 33, 21, 24, 25, 11, 33)
# j'ai essayé pendant près d'une demi-heure d'utiliser la commande stars
# pour faire ça, mais sans succès.
clock.plot <- function (epi1, col=rainbow(n), ...) {
if( min(x)<0 ) x <- x - min(x)
if( max(x)>1 ) x <- x/max(x)
n <- length(x)
if(is.null(names(x))) names(x) <- 0:(n-1)
m <- 1.05
plot(0, type='n', xlim=c(-m,m), ylim=c(-m,m),
axes=F, xlab='', ylab='', ...)
a <- pi/2 - 2*pi/200*0:200
polygon( cos(a), sin(a) )
v <- .02
a <- pi/2 - 2*pi/n*0:n
segments( (1+v)*cos(a), (1+v)*sin(a), (1-v)*cos(a), (1-v)*sin(a) )
segments( cos(a), sin(a), 0, 0, col='light grey', lty=3)
ca <- -2*pi/n*(0:50)/50
for (i in 1:n) {
a <- pi/2 - 2*pi/n*(i-1)
b <- pi/2 - 2*pi/n*i
polygon( c(0, x[i]*cos(a+ca), 0),
c(0, x[i]*sin(a+ca), 0),
col=col[i] )
v <- .1
```

```
text((1+v)*cos(a), (1+v)*sin(a), names(x)[i])

}

}

clock.plot(epi1$AGE, main="Affluence d'un site Web en fonction de l'heure")
```