```{r}

chooseCRANmirror(graphics=FALSE, ind=1)

knitr::opts\_chunk$set(echo = TRUE)

```

```{r}

library(summarytools)

install.packages("actuar")

library(actuar)

install.packages("contrib.url")

install.packages("Hmisc")

library(Hmisc)

install.packages("corrplot")

library(corrplot)

install.packages("sandwich")

library(sandwich)

install.packages("lmtest")

library(lmtest)

install.packages("summarytools")

library(summarytools)

install.packages("tree")

library("tree")

install.packages("AER")

library(AER)

install.packages("MASS")

library(MASS)

install.packages("pscl")

library(pscl)

install.packages("vcd")

library(vcd)

install.packages("car")

library(car)

```

```{r}

setwd("/Users/Deneux/Desktop/Jeremy&Pierre")

base\_sinistre<- read.csv(file="Jeremy&Pierre-PG\_2017\_CLAIMS\_YEAR0.csv", sep = ";")

base\_client <- read.csv(file="Jeremy&Pierre-PG\_2017\_YEAR0.csv", sep = ";")

base\_sinistrePositive <- subset(base\_sinistre, claim\_amount >=0)

```

```{r}

base\_sinistrePositive$claim\_amount <- gsub(",",".", base\_sinistrePositive$claim\_amount)

base\_sinistrePositive$claim\_amount <- as.numeric(as.character(base\_sinistrePositive$claim\_amount))

```

```{r}

library(dplyr)

montant\_sinistre\_annuel <- base\_sinistrePositive %>% group\_by(id\_policy) %>% filter(claim\_amount >= 0) %>%

summarise(claim\_amount = sum(claim\_amount))

nb\_sinistres <- base\_sinistrePositive %>% group\_by(id\_policy) %>% filter(claim\_amount >= 0) %>%

summarise(claim\_amount = n())

```

```{r}

base\_etude <- base\_client %>% full\_join(montant\_sinistre\_annuel, by="id\_policy")

base\_etude$claim\_amount[is.na(base\_etude$claim\_amount)] <- 0

names(base\_etude)[names(base\_etude)=="claim\_amount"] <- "montant\_sinistre\_annuel"

```

```{r}

base\_etude<- base\_etude%>% full\_join(nb\_sinistres, by="id\_policy")

base\_etude$claim\_amount[is.na(base\_etude$claim\_amount)] <- 0

names(base\_etude)[names(base\_etude)=="claim\_amount"] <- "nb\_sinistres"

```

```{r}

summary(base\_etude$montant\_sinistre\_annuel)

table(base\_etude$claim\_amount)/nrow(base\_etude)\*100

table(base\_etude$drv\_sex1)/nrow(base\_etude)\*100

table(base\_etude$drv\_sex2)/nrow(base\_etude)\*100

str(base\_etude)

table(base\_etude$pol\_coverage)/nrow(base\_etude)\*100

table(base\_etude$pol\_usage)/nrow(base\_etude)\*100

table(base\_etude$vh\_make)/nrow(base\_etude)\*100

table(base\_etude$vh\_fuel)/nrow(base\_etude)\*100

```

```{r}

hist(base\_etude$nb\_sinistres,freq=TRUE,plot=TRUE,xlab="Nombre sinistre", main="Repartition du nombre des sinistres", col="grey")

```

Etude de la corrélation entre les variables

```{r}

variables <-base\_etude[,c("drv\_age1","drv\_age\_lic1","drv\_age2","drv\_age\_lic2","vh\_age","vh\_din","vh\_sale\_begin","vh\_sale\_end","vh\_speed","vh\_value","vh\_weight","pol\_bonus","pol\_duration","pol\_sit\_duration","montant\_sinistre\_annuel","nb\_sinistres")]

mycor <- cor(variables)

```

Significativité de la correlation

```{r}

rcorr(as.matrix(variables[,1:16]))

```

Signe de la correlation

```{r}

symnum(mycor, abbr.colnames=FALSE)

```

```{r}

base\_etude$nombre\_sinistre <- as.character(base\_etude$nb\_sinistres)

base\_etude$nombre\_sinistre[base\_etude$claim\_nb == 0] <- "0"

base\_etude$nombre\_sinistre[base\_etude$claim\_nb == 1 ] <- "1"

base\_etude$nombre\_sinistre[base\_etude$claim\_nb == 2 ] <- "2"

base\_etude$nombre\_sinistre[base\_etude$claim\_nb == 3 ] <- "3"

base\_etude$nombre\_sinistre[base\_etude$claim\_nb == 4 ] <- "4"

base\_etude$nombre\_sinistre[base\_etude$claim\_nb == 5 ] <- "5"

base\_etude$nombre\_sinistre[base\_etude$claim\_nb == 6 ] <- "6"

ggplot(base\_etude) +

aes(x = vh\_age, fill = nombre\_sinistre) +

geom\_bar() +

xlab("Age du vehicule") +

ylab("Effectifs") +

labs(fill = "Nombre des sinistres")

plot(base\_etude$vh\_age ~ base\_etude$nb\_sinistres)

```

```{r}

library(scales)

ggplot(base\_etude) +

aes(x = vh\_age, fill = nombre\_sinistre) +

geom\_bar(position = "fill") +

xlab("Age du vehicule") +

ylab("Proportion") +

labs(fill = "Nombre des sinistres") +

scale\_y\_continuous(labels = percent)

```

```{r}

ggplot(base\_etude) +

aes(x = drv\_age1, fill = nombre\_sinistre) +

geom\_bar() +

xlab("Age du conducteur") +

ylab("Effectifs") +

labs(fill = "Nombre des sinistres")

ggplot(base\_etude) +

aes(x = drv\_age1, fill = nombre\_sinistre) +

geom\_bar(position = "fill") +

xlab("Age du conducteur") +

ylab("Proportion") +

labs(fill = "Nombre des sinistres") +

scale\_y\_continuous(labels = percent)

```

```{r}

library(emmeans)

moyenne <- mean(base\_etude$nb\_sinistres)

variance <- var(base\_etude$nb\_sinistres)

summary(variance)

summary(moyenne)

```

```{r}

base\_etude$Dummy\_sin[base\_etude$nb\_sinistres >=1] <- 1

base\_etude$Dummy\_sin[is.na(base\_etude$Dummy\_sin)] <- 0

sum(base\_etude$Dummy\_sin)/nrow(base\_etude)

by(base\_etude[,c("pol\_bonus","pol\_duration","pol\_sit\_duration","drv\_age1","drv\_age2",

"drv\_age\_lic1","drv\_age\_lic2","vh\_age","vh\_cyl","vh\_din","vh\_sale\_begin",

"vh\_sale\_end","vh\_speed","vh\_value","vh\_weight")],list(nbr\_sin=base\_etude$nb\_sinistres),summary) #important

by(base\_etude[,c("pol\_coverage","pol\_pay\_freq","pol\_payd","pol\_insee\_code","drv\_drv2",

"drv\_sex1","drv\_sex2","vh\_fuel","vh\_make","vh\_model","vh\_type")],list(nbr\_sin=base\_etude$nb\_sinistres),summary)

```

```{r}

library(Hmisc)

library(forcats)

#pol\_usage

base\_etude$pol\_usagee <- fct\_recode(base\_etude$pol\_usage,

"Professional" = "AllTrips")

#vh\_make

base\_etude$vh\_make2 <- as.character(base\_etude$vh\_make)

base\_etude$vh\_make2[!(base\_etude$vh\_make %in% c("RENAULT","PEUGEOT","CITROEN","VOLKSWAGEN","FORD","OPEL","TOYOTA","MERCEDES BENZ","FIAT"))] <- "AUTRE"

base\_etude$vh\_make2 <- as.factor(base\_etude$vh\_make2)

base\_etude$pol\_duration2 <- cut(base\_etude$pol\_duration, c(-Inf,21,Inf))

base\_etude$pol\_usage1.2 <- fct\_recode(base\_etude$pol\_usage,"Professional" = "AllTrips")

base\_etude$drv\_age1.2 <- cut(base\_etude$drv\_age1, c(-Inf, 35, 41,46,50,54,58,63,68,75,Inf))

base\_etude$vh\_fuel2 <- fct\_recode(base\_etude$vh\_fuel,"Diesel" = "Gasoline")

base\_etude$vh\_age2 <- cut(base\_etude$vh\_age, c(-Inf,17,Inf))

names(base\_etude)

nom\_variable <- c("vh\_age2","drv\_drv2","vh\_dinn","vh\_valuee","pol\_coverage",

"pol\_payd","vh\_fuel2","vh\_weightt","pol\_usage1.2",

"pol\_usage1.2","drv\_age1.2","vh\_agee","vh\_make2","vh\_speedd"

,"drv\_sex2","vh\_cyll","vh\_model","vh\_type"

,"pol\_pay\_freq")

```

```{r}

v<-base\_etude$nb\_sinistres

moyenne <-emm(v, order=1)

variance <- emm(v, order=2)

```

```{r}

goodfit\_pois<- goodfit(base\_etude$nb\_sinistres, type = c( "poisson"),

method = c("ML", "MinChisq"), par = NULL)

plot(goodfit\_pois) #loi de poisson

goodfit\_NB <- goodfit(base\_etude$nb\_sinistres, type = c( "nbinomial"),

method = c("ML", "MinChisq"), par = NULL)

plot(goodfit\_NB)

```

```{r}

fpois <- glm(nb\_sinistres ~ drv\_sex1 + drv\_drv2 + vh\_din + vh\_sale\_begin + vh\_value +

pol\_coverage + pol\_payd + drv\_age2 + vh\_fuel + vh\_sale\_end +

vh\_weight + pol\_duration + pol\_usage + drv\_age1 + drv\_age\_lic2 +

vh\_age + vh\_speed + pol\_sit\_duration + drv\_age\_lic1 +

drv\_sex2 + vh\_cyl + vh\_type + pol\_bonus + pol\_pay\_freq, family=poisson("log"), data=base\_etude)

summary(fpois)

stepwise <- step(fpois)

step(fpois)

```

```{r}

library(AER)

dispersiontest(fpois)

fpois2 <- glm(formula = nb\_sinistres ~ pol\_coverage + pol\_duration2 +

drv\_age1.2 + vh\_age2 , family = poisson("log"),

data = base\_etude)

summary(fpois2)

cov.fpois2 <- vcovHC(fpois2, type="HC0")

std.err <- sqrt(diag(cov.fpois2))

r.est <- cbind(Estimate= coef(fpois2), "Robust SE" = std.err,

"Pr(>|z|)" = 2 \* pnorm(abs(coef(fpois2)/std.err), lower.tail=FALSE),

LL = coef(fpois2) - 1.96 \* std.err,

UL = coef(fpois2) + 1.96 \* std.err)

```

```{r}

phi <- 0.9749978

regbn<- glm(formula=nb\_sinistres ~ pol\_coverage + pol\_duration2 + drv\_age1.2 + vh\_age2,family=negative.binomial(phi), data=base\_etude)

summary(regbn)

linearHypothesis(regbn,c("drv\_age1.2(46,50]=0","drv\_age1.2(50,54]=0"))

base\_etude$drv\_age1.2\_bis <- cut(base\_etude$drv\_age1, c(-Inf, 35, 41,46,54,58,63,68,75,Inf))

#Nouvelle regression NB apres le regroupement d'age

phi2 <- 0.975012

regbn.2 <- glm(formula=nb\_sinistres ~ pol\_coverage + pol\_duration2 + drv\_age1.2\_bis +

vh\_age2,family=negative.binomial(phi2), data=base\_etude)

summary(regbn.2)

```

```{r}

NB1 <- glm.nb(formula=nb\_sinistres ~ pol\_coverage + pol\_duration2 + drv\_age1.2 + vh\_age2, data=base\_etude)

summary(NB1)

NB2 <- glm.nb(formula=nb\_sinistres ~ pol\_coverage + pol\_duration2 + drv\_age1.2\_bis + vh\_age2, data=base\_etude)

summary(NB2)

```

```{r}

cbind(regbn$coefficients,fpois2$coefficients) # coefficient de poisson et coefficient binomiaux

#Graphiques des residus pour Poisson et Negative Binomiale par(mfrow = c(1, 2)) plot(regbnresiduals,ylab="ResidusNB")plot(fpois2residuals,ylab= “Residus Poisson”) #les résidus des modèles poissons et binomiaux

```

Comparaison des Residus

```{r}

plot(regbn$residuals,fpois2$residuals,xlab = "Residus BN", ylab = "Residus Poisson")

abline(a=0,b=1,lty=2,col="blue")

cbind(regbn$residuals,fpois2$residuals)

```

```{r}

quasipois <- glm(formula=nb\_sinistres ~ pol\_coverage + pol\_duration2 + drv\_age1.2 +

vh\_age2 ,

family=quasipoisson("log"), data=base\_etude)

summary(quasipois)

```

```{r}

quasipois2 <- glm(nb\_sinistres ~ pol\_coverage + pol\_duration2 + drv\_age1.2 +

vh\_age2 ,

family=quasipoisson("log"), data=base\_etude)

summary(quasipois2)

```

```{r}

regZI<- zeroinfl(nb\_sinistres ~pol\_coverage + drv\_age1.2 + vh\_age2 + pol\_duration2 |pol\_bonus ,data = base\_etude,dist = "poisson",link="logit")

summary(regZI)

```

````{r}

vuong(regZI,fpois2)

```

```{r}

regZI2<- zeroinfl(nb\_sinistres ~pol\_coverage + drv\_age1.2\_bis + vh\_age2 + pol\_duration2 |pol\_bonus ,data = base\_etude,dist = "poisson",link="logit")

summary(regZI2)

```

```{r}

regZINB<- zeroinfl(nb\_sinistres ~ pol\_coverage +drv\_age1.2 + vh\_age2 + pol\_duration2 | pol\_bonus,data = base\_etude,dist = "negbin")

summary(regZINB)

plot(regZINB$residuals)

```

````{r}

vuong(regZINB,NB1)

```

```{r}

regZINB2<- zeroinfl(nb\_sinistres ~pol\_coverage + pol\_duration2 + drv\_age1.2\_bis + vh\_age2 | pol\_bonus,data = base\_etude,dist = "negbin")

summary(regZINB2)

```

```{r}

logLik(NB2)

logLik(regbn.2)

logLik(fpois2)

logLik(regZINB2)

logLik(regZI2)

AIC(NB2)

AIC(regbn.2)

AIC(fpois2)

AIC(regZINB2)

AIC(regZI2)

deviance(NB2)

deviance(regbn.2)

deviance(fpois2)

deviance(regZINB2)

deviance(regZI2)

```

```{r}

plot(regZINB2$residuals,regZI2$residuals,xlab = "Coefficients ZIBN", ylab = "Coefficients ZIP")

abline(a=0,b=1,lty=2,col="red")

cbind(regZINB2$residuals,regZI2$residuals)

```

```{r}

par(mfrow = c(2, 2))

plot(fpois2$residuals,ylab= "Residus Poisson")

plot(regZI2$residuals,ylab = "Residus ZIP")

plot(regZINB2$residuals,ylab = "Residus ZINB")

plot(regbn$residuals,ylab = "Residus NB")

```

```{r}

setwd("/Users/Deneux/Desktop/Jeremy&Pierre")

base\_client1 <- read.csv(file="Jeremy&Pierre-PG\_2017\_YEAR1.csv")

```

```{r}

base\_client1$vh\_age2 <- cut(base\_client1$vh\_age, c(-Inf,17,Inf))

base\_client1$pol\_duration2 <- cut(base\_client1$pol\_duration, c(-Inf,21,Inf))

base\_client1$drv\_age1.2 <- cut(base\_client1$drv\_age1, c(-Inf, 35, 41,46,50,54,58,63,68,75,Inf))

base\_client1$drv\_age1.2\_bis <- cut(base\_client1$drv\_age1, c(-Inf, 35, 41,46,54,58,63,68,75,Inf))

```

```{r}

Prediction\_frequence\_nb2 <- predict.glm(NB2, base\_client1, type="response")

summary(Prediction\_frequence\_nb2)

```

```{r}

Prediction\_frequence\_ZINB <- predict(regZINB2, base\_client1)

summary(Prediction\_frequence\_ZINB)

```

```{r}

base\_severite <- subset(base\_etude , montant\_sinistre\_annuel>0 )

base\_severite$vh\_age <- cut(base\_severite$vh\_age, c(-Inf,17,Inf))

base\_severite$pol\_duration<- cut(base\_severite$pol\_duration, c(-Inf,21,Inf))

base\_severite$drv\_age1.2 <- cut(base\_severite$drv\_age1, c(-Inf, 35, 41,46,50,54,58,63,68,75,Inf))

```

```{r}

mean(base\_severite$montant\_sinistre\_annuel)

quantile(base\_severite$claim\_amount,prob=c(.5,.9,.95,.99))

aggregate(base\_severite$vh\_din ,list(base\_severite$drv\_age1.2), mean)

aggregate(base\_severite$vh\_speed ,list(base\_severite$drv\_age1.2), mean)

aggregate(base\_severite$vh\_cyl ,list(base\_severite$drv\_age1.2), mean)

aggregate(base\_severite$vh\_value ,list(base\_severite$drv\_age1.2), mean)

table(base\_severite$drv\_age1.2,base\_severite$vh\_fuel)

```

```{r}

log\_normale <- lm(log(montant\_sinistre\_annuel) ~ drv\_sex1 + drv\_drv2 + vh\_din + vh\_sale\_begin + vh\_value +

pol\_coverage + pol\_payd + drv\_age2 + vh\_fuel + vh\_sale\_end +

vh\_weight + pol\_duration + pol\_usage + drv\_age1 + drv\_age\_lic2 +

vh\_age + vh\_speed + pol\_sit\_duration + drv\_age\_lic1 +

drv\_sex2 + vh\_cyl + vh\_type + pol\_bonus + pol\_pay\_freq, data=base\_severite)

stepwise\_LN <-step(log\_normale)

log\_normale2 <- lm(log(montant\_sinistre\_annuel) ~ vh\_fuel+ drv\_age1 , data=base\_severite)

summary(log\_normale2)

```

````{r}

loglikg<-logLik(log\_normale2)

akaike<-function(npar,loglik,k){-2\*loglik+k\*npar}

akaike(9,loglikg[1],2)

```

````{r}

AIC(log\_normale2)

deviance(log\_normale2)

logLik(log\_normale2)

plot(log\_normale2$residuals)

````

```{r}

gamma1 <- glm(montant\_sinistre\_annuel ~ drv\_sex1 + drv\_drv2 + vh\_din + vh\_sale\_begin + vh\_value +

pol\_coverage + pol\_payd + drv\_age2 + vh\_fuel + vh\_sale\_end +

vh\_weight + pol\_duration + pol\_usage + drv\_age1 + drv\_age\_lic2 +

vh\_age + vh\_speed + pol\_sit\_duration + drv\_age\_lic1 +

drv\_sex2 + vh\_cyl + vh\_type + pol\_bonus + pol\_pay\_freq, family=Gamma(link="log"), data=base\_severite)

summary(gamma1)

stepwise\_gamma <- step(gamma1)

summary(stepwise\_gamma)

gamma3 <- glm(montant\_sinistre\_annuel ~ vh\_fuel +drv\_age1 , family = Gamma(link = "log"), data=base\_severite )

summary(gamma3)

```

```{r}

AIC(gamma3)

deviance(gamma3)

logLik(gamma3)

plot(gamma3$residuals)

```

```{r}

base\_client1$Prediction\_severite <- predict.glm(gamma3, base\_client1, type="response")

summary(base\_client1$Prediction\_severite)

```

```{r}

aggregate(base\_client1$Prediction\_severite ,list(base\_client1$vh\_fuel), mean)

aggregate(base\_client1$vh\_value ,list(base\_client1$vh\_fuel), mean)

```

```{r}

prediction\_sev <-predict(log\_normale2, base\_client1,type="response")

sigma <- summary(log\_normale2)$sigma

Pln <- exp(prediction\_sev+sigma^2/2)

mean(Pln)

min(Pln)

max(Pln)

```

```{r}

base\_client1$tarif <- Prediction\_frequence\_nb2\*Pln

summary(base\_client1$tarif)

base\_tarif <- base\_client1[,c("id\_policy","tarif")]

write.csv(base\_tarif, file = "base\_tarif.csv")

```

```{r}

M <- base\_severite[order(-base\_severite$montant\_sinistre\_annuel),c("montant\_sinistre\_annuel","nb\_sinistres","vh\_din","vh\_age","drv\_age1","vh\_type")]

M$SUM <- cumsum(M$montant\_sinistre\_annuel)/sum(M$montant\_sinistre\_annuel)\*100

head(M)

```