# TP : Analyse de donnée

# ***1ere partie :***

## 1ere méthode :

#Insertion des données : colonnes par colonnes

F<-c(9,15,5,11,11,3)

M<-c(12,9,10,13,13,15)

H<-c(10,10,8,14,8,10)

# Rassembler les colonnes dans un tableau

A<-data.frame(F,M,H)

#représentation du tableau A en nuage de points en prenant les variables deux par deux :

plot(A)

#la moyenne de chaque variable :

Fbar<-mean(F)

Mbar<-mean(M)

Hbar<-mean(H)

#l’écart type de chaque variable :

sigmaF<-sd(F)

sigmaM<-sd(M)

sigmaH<-sd(H)

#le résumé de donné (donne les quantiles de chaque variable) :

summary(A)

#centrer la matrice des donnée

Fc<-F-Fbar

Mc<-M-Mbar

Hc<-H-Hbar

X<-data.frame(Fc,Mc,Hc)

#Centrer et réduire les donnée

x<-scale(A,center=T,scale=T)

X<-as.matrix(x)

X<-X[1:6,]

S<-t(X)%\*%X

#Extraire les valeurs et les vecteurs propres

eig<-eigen(S)

eigvalues<-eig$values

eigvalues<-data.frame(eigvalues)

lamda1<-eigvalues[1,1]

lamda2<-eigvalues[2,1]

lamda3<-eigvalues[3,1]

eigvectors<-eig$vectors

eigvectors<-data.frame(eigvectors)

colnames(eigvectors)<-c("U1","U2","U3")

U1<-eigvectors$U1

U2<-eigvectors$U2

U3<-eigvectors$U3

#Normé les vecteurs propres

normU1<-sum(t(U1)%\*%U1)^(1/2)

normU2<-sum(t(U2)%\*%U2)^(1/2)

W1<-U1/normU1

W2<-U2/normU2

W<-data.frame(W1,W2)

W<-as.matrix(W)

psi<-X%\*%W

V1<-(1/sqrt(lamda1))\*X%\*%W1

V1<-(1/sqrt(lamda2))\*X%\*%W2

V<-data.frame(V1,V2)

Fi1<-(sqrt(lamda1))\*W1

Fi2<-(lamda2^0.5))\*W2

Fi<-data.frame(Fi1,Fi2)

## 2ème méthode :

library("gtools")

library("gdata")

A<-read.xls('LAIB.xlsx',sheet=1)#mettre le fichier Excel dans le répertoire de travail.

A<-as.matrix(A)

t<-c(1,2,3)

names(t)<-c("a","b","c")

colnames(A)<-c(letters[1:3])

rownames(A)<-c(letters[1:6])

library("ade4")

Y<-dudi.pca(A,center=TRUE,scale=TRUE)#réalisation d’une ACP normé

Y$eig#valeurs propres

Y$li#coordonnées des individus

Y$c1

Y$l1

scatter.dudi(Y)#la présentation graphique des individus et des variables (dans les axes factoriel)

inertia.dudi(Y)#donne un tableau des inerties de chaque valeurs propres et les ratios.

s.corcircle(Y$li)#cercle de corrélation des individus

s.corcircle(Y$co)#cercle de corrélation des variable

### Comparaison entre les 2 méthodes précédente :

Dans la 1ere méthode c’est l’utilisateur qui a créé le tableau de donnée sur R directement, faire une ACP pas à pas et programmer les fonctions à l’aide de la théorie… par contre dans la 2ème méthode on a importé les données depuis Excel et on a réalisé une ACP à l’aide d’un package - R qui fournit des fonctions déjà programmées, « ade4 ».

## 3ème méthode :

U<-princomp(A)#réaliser une ACP dont les sorties sont les valeurs propres et la dimension du tableau de donnée

biplot(U)#représentation graphique des individus et variables

summary(U)#donne un résumé qui permet de choisir le nombre d’axe à retenir grâce à l’inertie accumulée

loadings(U)#la contribution des variable dans les composantes principales

plot(U)#l’histogramme des valeurs propres.

# ***2ème partie :***

>library(gdata)

>autos<-read.xls('TP.xlsx',sheet=1)

> autos<-autos[,-1]

>autos.actifs<- autos[,1:6]

>autos.illus<- autos[,7:9]

> n <- nrow(autos.actifs)

>summary(autos)

CYLINDREE PUISSANCE LONGEUR LARGEUR POIDS

Min. :1166 Min. : 55.00 Min. :393.0 Min. :157.0 Min. : 815

1st Qu.:1310 1st Qu.: 70.75 1st Qu.:424.0 1st Qu.:162.2 1st Qu.:1020

Median :1578 Median : 82.00 Median :434.5 Median :167.0 Median :1088

Mean :1632 Mean : 84.61 Mean :433.5 Mean :166.7 Mean :1079

3rd Qu.:1798 3rd Qu.: 98.00 3rd Qu.:448.0 3rd Qu.:169.8 3rd Qu.:1127

Max. :2664 Max. :128.00 Max. :469.0 Max. :177.0 Max. :1370

VITESS FINITION PRIX R.POID.PUIS

Min. :140.0 B :7 Min. :22100 Min. : 9.72

1st Qu.:151.2 M :5 1st Qu.:29843 1st Qu.:11.22

Median :160.0 TB:6 Median :33395 Median :13.18

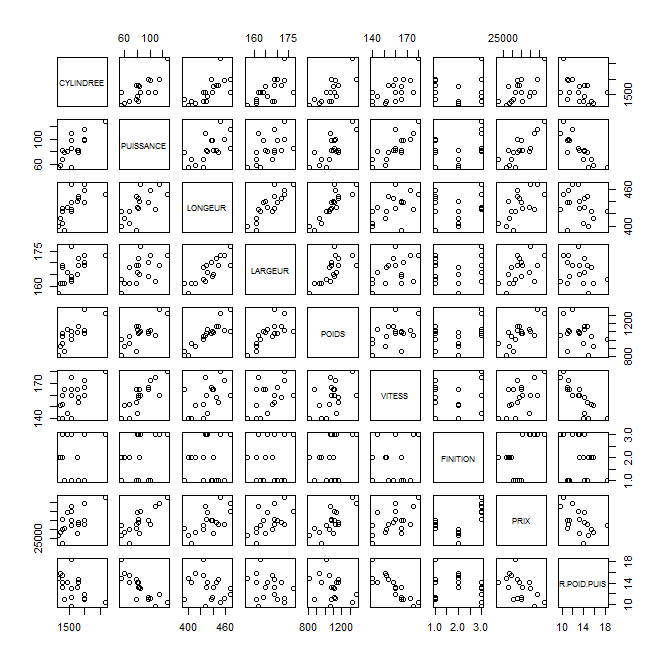
Mean :158.3 Mean :34186 Mean :13.18

3rd Qu.:165.0 3rd Qu.:38458 3rd Qu.:14.55

Max. :180.0 Max. :47700 Max. :18.36

> #nuages de points

>pairs(autos)



**FINITION est une variablequalitative. En général, sonintroduction dans ce typede graphique n’est pas trèsindiquée. Néanmoins, onremarquera qu’on peutparfois en tirer desinformations utiles : parexemple, ici, selon lafinition, les prix sontdifférents**.

print(summary(acp.autos))

Importance of components:

Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6

Standard deviation 2.1025837 0.9252363 0.61079135 0.46251712 0.30463291 0.208063146

Proportion of Variance 0.7368097 0.1426770 0.06217768 0.03565368 0.01546687 0.007215045

Cumulative Proportion 0.7368097 0.8794867 0.94166440 0.97731809 0.99278495 1.000000000

> #quelles les propriétés associées à l'objet ?

>print(attributes(acp.autos))

$names

[1] "sdev" "loadings" "center" "scale" "n.obs" "scores" "call"

$class

[1] "princomp"

>val.propres<- acp.autos$sdev^2

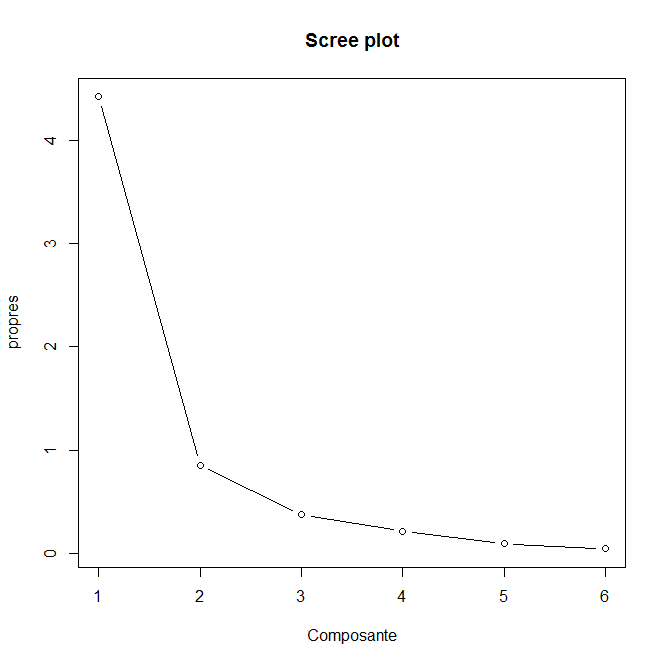
>print(val.propres)

Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6

4.42085806 0.85606229 0.37306608 0.21392209 0.09280121 0.04329027

> #scree plot (graphique des éboulis des valeurs propres)

>plot(1:6,val.propres,type="b",ylab="Valeurs propres",xlab="Composante",main="Scree plot")



**Les deux premiers axes traduisent 88%de l’information disponible. On se rendcompte ici qu’on pouvait s’en teniruniquement au premier facteur.Mais c’est moins pratique pour lesgraphiques ; on suspecte aussi un« effet taille » dans les données. On vadonc conserver les deux premiersfacteurs.**

> #intervalle de confiance des val.propres à 95% (cf.Saporta, page 172)

>val.basse<- val.propres \* exp(-1.96 \* sqrt(2.0/(n-1)))

>val.haute<- val.propres \* exp(+1.96 \* sqrt(2.0/(n-1)))

> #affichage sous forme de tableau

> tableau <- cbind(val.basse,val.propres,val.haute)

>colnames(tableau) <- c("B.Inf.","Val.","B.Sup")

>print(tableau,digits=3)

B.Inf. Val. B.Sup

Comp.1 2.2571 4.4209 8.6591

Comp.2 0.4371 0.8561 1.6768

Comp.3 0.1905 0.3731 0.7307

Comp.4 0.1092 0.2139 0.4190

Comp.5 0.0474 0.0928 0.1818

Comp.6 0.0221 0.0433 0.0848

**Les intervalles de confiance d’Andersonne sont licites que si le nuage de pointsest gaussien. On ne l’affiche donc qu’à titre indicatif**

> #\*\*\*\* corrélation variables-facteurs \*\*\*\*

> c1 <- acp.autos$loadings[,1]\*acp.autos$sdev[1]

> c2 <- acp.autos$loadings[,2]\*acp.autos$sdev[2]

> #affichage

>correlation<- cbind(c1,c2)

>print(correlation,digits=2)

c1 c2

CYLINDREE -0.89 0.11

PUISSANCE -0.89 0.38

LONGEUR -0.89 -0.38

LARGEUR -0.81 -0.41

POIDS -0.91 -0.22

VITESS -0.75 0.57

> #carrés de la corrélation (cosinus²)

>print(correlation^2,digits=2)

c1 c2

CYLINDREE 0.80 0.013

PUISSANCE 0.79 0.148

LONGEUR 0.79 0.145

LARGEUR 0.66 0.170

POIDS 0.82 0.050

VITESS 0.57 0.329

> #cumul carrés de la corrélation

>print(t(apply(correlation^2,1,cumsum)),digits=2)

c1 c2

CYLINDREE 0.80 0.81

PUISSANCE 0.79 0.93

LONGEUR 0.79 0.93

LARGEUR 0.66 0.83

POIDS 0.82 0.87

VITESS 0.57 0.90

>plot(acp.autos$scores[,1],acp.autos$scores[,2],type="n",xlab="Comp.1 -

+ 74%",ylab="Comp.2 - 14%")

>abline(h=0,v=0)

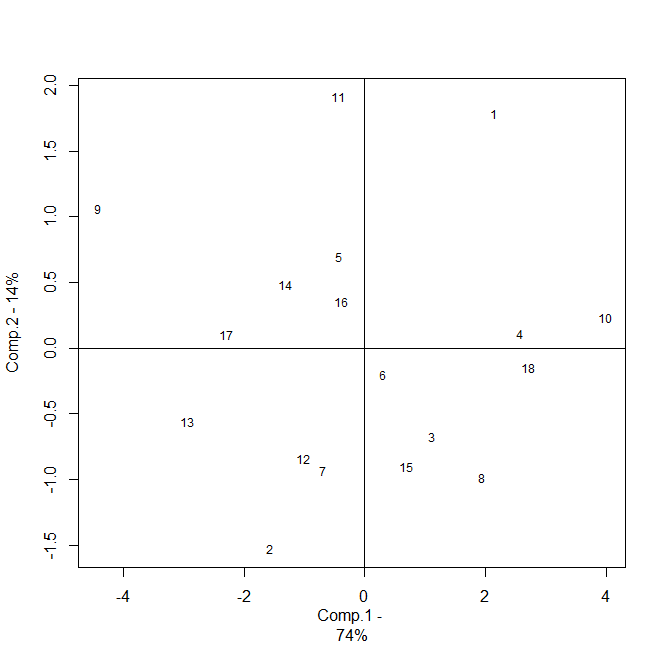
>

>plot(acp.autos$scores[,1],acp.autos$scores[,2],type="n",xlab="Comp.1 -

+ 74%",ylab="Comp.2 - 14%")

>abline(h=0,v=0)

>text(acp.autos$scores[,1],acp.autos$scores[,2],labels=rownames(autos.actifs),cex=0.75)



biplot(acp.autos,cex=0.75)

>

> #calcul du carré de la distance d'un individu au center de gravité

> d2 <- function(x){return(sum(((x-acp.autos$center)/acp.autos$scale)^2))}

> #appliquer à l'ensemble des observations

> all.d2 <- apply(autos.actifs,1,d2)

> #cosinus^2 pour une composante

> cos2 <- function(x){return(x^2/all.d2)}

> #cosinus^2 pour les composantes retenues (les 2 premières)

> all.cos2 <- apply(acp.autos$scores[,1:2],2,cos2)

>print(all.cos2)

Comp.1 Comp.2

[1,] 0.55621843 0.387670355

[2,] 0.36533449 0.349406285

[3,] 0.58028427 0.210694005

[4,] 0.97699183 0.001879434

[5,] 0.15657878 0.413825852

[6,] 0.08155471 0.033899586

[7,] 0.30920173 0.575488494

[8,] 0.67353923 0.170535407

[9,] 0.89243070 0.051919840

[10,] 0.97521929 0.003425970

[11,] 0.04297836 0.820652103

[12,] 0.53094720 0.362855147

[13,] 0.77838982 0.028137029

[14,] 0.70481879 0.096496323

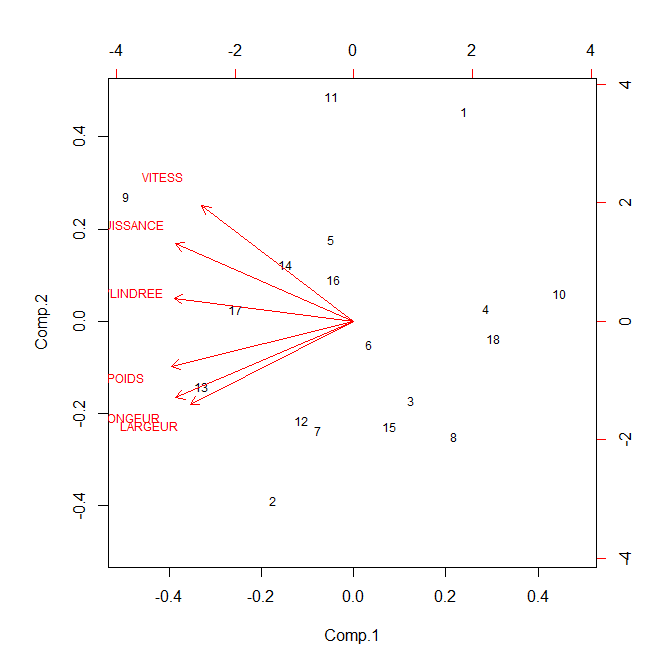
[15,] 0.24327262 0.410468952

[16,] 0.21733631 0.185337481

[17,] 0.86189979 0.001789839

[18,] 0.92605245 0.002606517

**La somme pour chaque ligne(individu) vaut 1 si l’on prendl’ensemble des composantes (les6 composantes).**



> #contributions à une composante - calcul pour les 2 premières composantes

>all.ctr<- NULL

>for (k in 1:2){all.ctr<-

+ cbind(all.ctr,100.0\*(1.0/n)\*(acp.autos$scores[,k]^2)/

+ (acp.autos$sdev[k]^2))}

>print(all.ctr)

[,1] [,2]

[1,] 5.7492537 20.69330670

[2,] 3.0639513 15.13293304

[3,] 1.5746358 2.95251897

[4,] 8.3243598 0.08269669

[5,] 0.2300459 3.13978898

[6,] 0.1163180 0.24968553

[7,] 0.5878172 5.64986748

[8,] 4.7710990 6.23837245

[9,] 24.4368837 7.34185620

[10,] 19.9640247 0.36218506

[11,] 0.2407083 23.73566876

[12,] 1.3027651 4.59779144

[13,] 10.8701289 2.02916325

[14,] 2.1726682 1.53612980

[15,] 0.6002290 5.23004324

[16,] 0.1869563 0.82332696

[17,] 6.5887644 0.07065818

[18,] 9.2193906 0.13400728

**La somme pour chaque colonne(composante) vaut 100.**

**Remarque : toutes lesobservations ont le même poidsdans notre exemple.**

> #\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

> #que faut-il penser de PRIX et R.POIS.PUIS ?

> #\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

> #corrélation de chaque var. illustrative avec le premier axe

> ma\_cor\_1 <- function(x){return(cor(x,acp.autos$scores[,1]))}

> s1 <- sapply(autos.illus[,2:3],ma\_cor\_1)

> #corrélation de chaque variable illustrative avec le second axe

> ma\_cor\_2 <- function(x){return(cor(x,acp.autos$scores[,2]))}

> s2 <- sapply(autos.illus[,2:3],ma\_cor\_2)

> #position sur le cercle

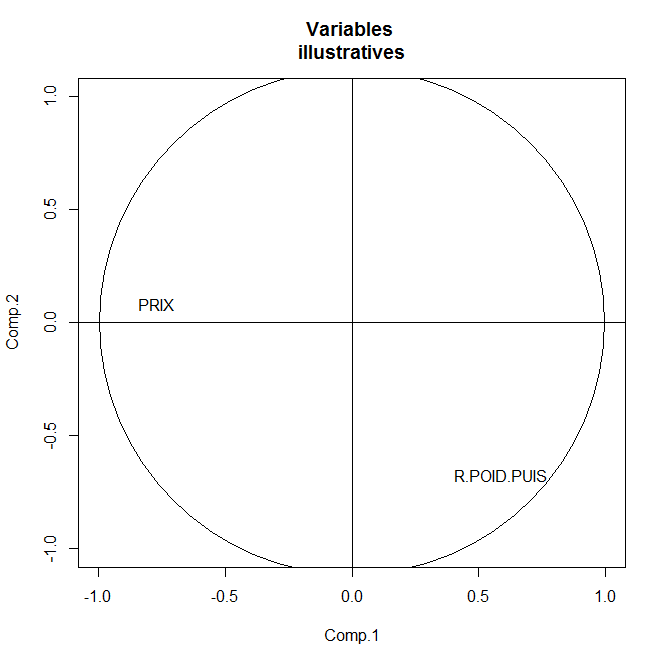
> plot(s1,s2,xlim=c(-1,+1),ylim=c(-1,+1),type="n",main="Variables

+ illustratives",xlab="Comp.1",ylab="Comp.2")

>abline(h=0,v=0)

>text(s1,s2,labels=colnames(autos.illus[2:3]),cex=1.0)

>symbols(0,0,circles=1,inches=F,add=T)



**La représentation simultanée estautrement plus instructive.**

**La forte corrélation entre unevariable illustrative et un axe estd’autant plus intéressante que lavariable n’a pas participé à laconstruction de l’axe.**

> #\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

> #que faut-il penser de PRIX et R.POIS.PUIS ?

> #\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

> #corrélation de chaque var. illustrative avec le premier axe

> ma\_cor\_1 <- function(x){return(cor(x,acp.autos$scores[,1]))}

> s1 <- sapply(autos.illus[,2:3],ma\_cor\_1)

> #corrélation de chaque variable illustrative avec le second axe

> ma\_cor\_2 <- function(x){return(cor(x,acp.autos$scores[,2]))}

> s2 <- sapply(autos.illus[,2:3],ma\_cor\_2)

> #position sur le cercle

> plot(s1,s2,xlim=c(-1,+1),ylim=c(-1,+1),type="n",main="Variables

+ illustratives",xlab="Comp.1",ylab="Comp.2")

>abline(h=0,v=0)

>text(s1,s2,labels=colnames(autos.illus[2:3]),cex=1.0)

>symbols(0,0,circles=1,inches=F,add=T)

> #représentation simultanée (avec les variables actives)

> plot(c(c1,s1),c(c2,s2),xlim=c(-1,+1),ylim=c(-1,+1),type="n",main="Variables

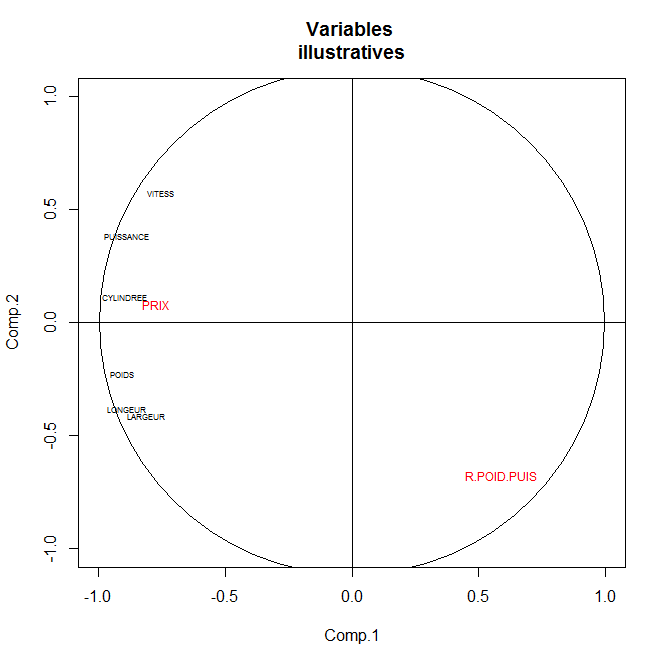
+ illustratives",xlab="Comp.1",ylab="Comp.2")

>text(c1,c2,labels=colnames(autos.actifs),cex=0.5)

>text(s1,s2,labels=colnames(autos.illus[2:3]),cex=0.75,col="red")

>abline(h=0,v=0)

>symbols(0,0,circles=1,inches=F,add=T)



> #\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

> #positionner les modalités de la variable illustrative + calcul des valeurs test

> #\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

> K <- nlevels(autos.illus[,"FINITION"])

>var.illus<- unclass(autos.illus[,"FINITION"])

> m1 <- c()

> m2 <- c()

>for (i in 1:K){m1[i] <- mean(acp.autos$scores[var.illus==i,1])}

>for (i in 1:K){m2[i] <- mean(acp.autos$scores[var.illus==i,2])}

>cond.moyenne<- cbind(m1,m2)

>rownames(cond.moyenne) <- levels(autos.illus[,"FINITION"])

>print(cond.moyenne)

m1 m2

B -0.2353131 0.04527122

M 2.0003548 -0.02257896

TB -1.3924304 -0.03400062

> #graphique

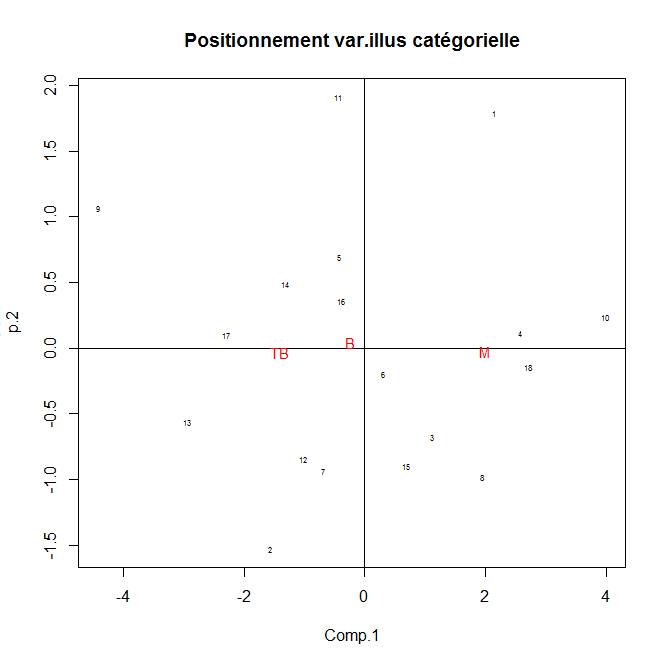
> plot(c(acp.autos$scores[,1],m1),c(acp.autos$scores[,2],m2),xlab="Comp.1",ylab="Com

+ p.2",main="Positionnement var.illuscatégorielle",type="n")

>abline(h=0,v=0)

>text(acp.autos$scores[,1],acp.autos$scores[,2],rownames(autos.actifs),cex=0.5)

>text(m1,m2,rownames(cond.moyenne),cex=0.95,col="red")



**Dans notre exemple, les véhicules se différencient véritablement par la FINITION sur le premier axe factoriel.**