المدرسة الوطنية للعلوم التطبيقية -أكادير
ⵜⵉⵎⵙⵙⵓⵜ ⵜⵜⴰⵎⵓⵔⵜ ⵏ ⵜⵙⵙⵏⴰⵡⵉⵏ ⵜⵜⵓⵜⵜⵓⵔⵉⵏ - ⴰⴳⴰⴷⵉⵔ.
ÉCOLE NATIONALE DES SCIENCES APPLIQUÉES - AGADIR

UNIVERSITÉ IBN ZOHR - AGADIR

ENSA Agadir

PROJET D'ANALYSE DE DONNEES

Classification hiérarchique, régression logistique et analyse discriminante sur R

Vaneck Fokou Penanjo
ENSA AGADIR, FID1

PLAN

I. Chapitre 1: Classification ascendante hiérarchique (CAH)

1) Introduction
2) Mise en œuvre de la CAH sur R
3) Interprétation des résultats
4) conclusion

II. Chapitre 2: Régression logistique

1) Introduction
2) Mise en œuvre de la régression logistique su R et interprétation
3) Prédiction
4) conclusion

III. Chapitre 3: Analyse factorielle discriminante

1) Introduction
2) Mise en œuvre de l'analyse discriminante et interprétation
3) Prédiction
4) Conclusion

CHAPITRE 1 : CLASSIFICATION ASCENDANTE HIERARCHIQUE (CAH)

I. INTRODUCTION :

Il existe de nombreuses techniques statistiques visant à partitionner une population en différentes classes ou sous-groupes. La classification ascendante hiérarchique est l'une d'entre elle. On cherche à ce que les individus regroupés au sein d'une même classe (homogénéité intra-classe) soient le plus semblable possible tandis que les classes soient le plus dissemblables (hétérogénéité interclasse). Le principe de la CAH est de rassembler des individus de manière itérative selon un critère de ressemblance défini au préalable pour produire des classes ou groupe de plus en plus vaste.

Dans notre devoir, il sera question pour nous de réaliser une classification sur des résultats de 35 étudiants d'une école sur 12 matières à l'aide du logiciel R.

II. MISE EN ŒUVRE DE LA C.A.H SUR R.

Notre classification a été effectuée sur R grâce au packages « FactoMineR » et« Factoshiny ». Sur R, notre jeu de donnée est importé sur R à l'aide de la syntaxe :
>data = read.csv(file.choose(), header = TRUE, sep = ";", row.names = 1, dec = ",")

Pour mener à bien notre classification, nous avons au préalable effectué une analyse en composantes principales sur notre jeu de données pour identifier la structure de la population et éventuellement de déterminer le nombre de groupes à construire.

Le code permettant d'ouvrir l'interface Factoshiny :
>library(Factoshiny)
>Factoshiny(data)
Après ouverture de l'interface, on lance l'analyse en composantes principales ; puis dans le paramétrage de l'ACP, on coche l'option classification après avoir quitté l'appli d'ACP. Nous effectuons donc ainsi notre classification. La métrique utilisée est la distance euclidienne. Le code ayant permis d'effectuer cette classification et d'obtenir les graphes est le suivant :

```
res.PCA<-PCA(data,graph=FALSE)

res.HCPC<-HCPC(res.PCA,nb.clust=3,consol=FALSE,graph=FALSE)

plot.HCPC(res.HCPC,choice='tree',title='Arbre hiérarchique')

plot.HCPC(res.HCPC,choice='map',draw.tree=FALSE,title='Plan factoriel')

plot.HCPC(res.HCPC,choice='3D.map',ind.names=FALSE,centers.plot=FALSE,angle=60,tit
le='Arbre hiérarchique sur le plan factoriel')

res.HCPC<-HCPC(res.PCA,nb.clust=3,consol=FALSE,graph=FALSE)

summary(res.HCPC)
```

III. INTERPRETATION DES RESULTATS

1) L'arbre hiérarchique.

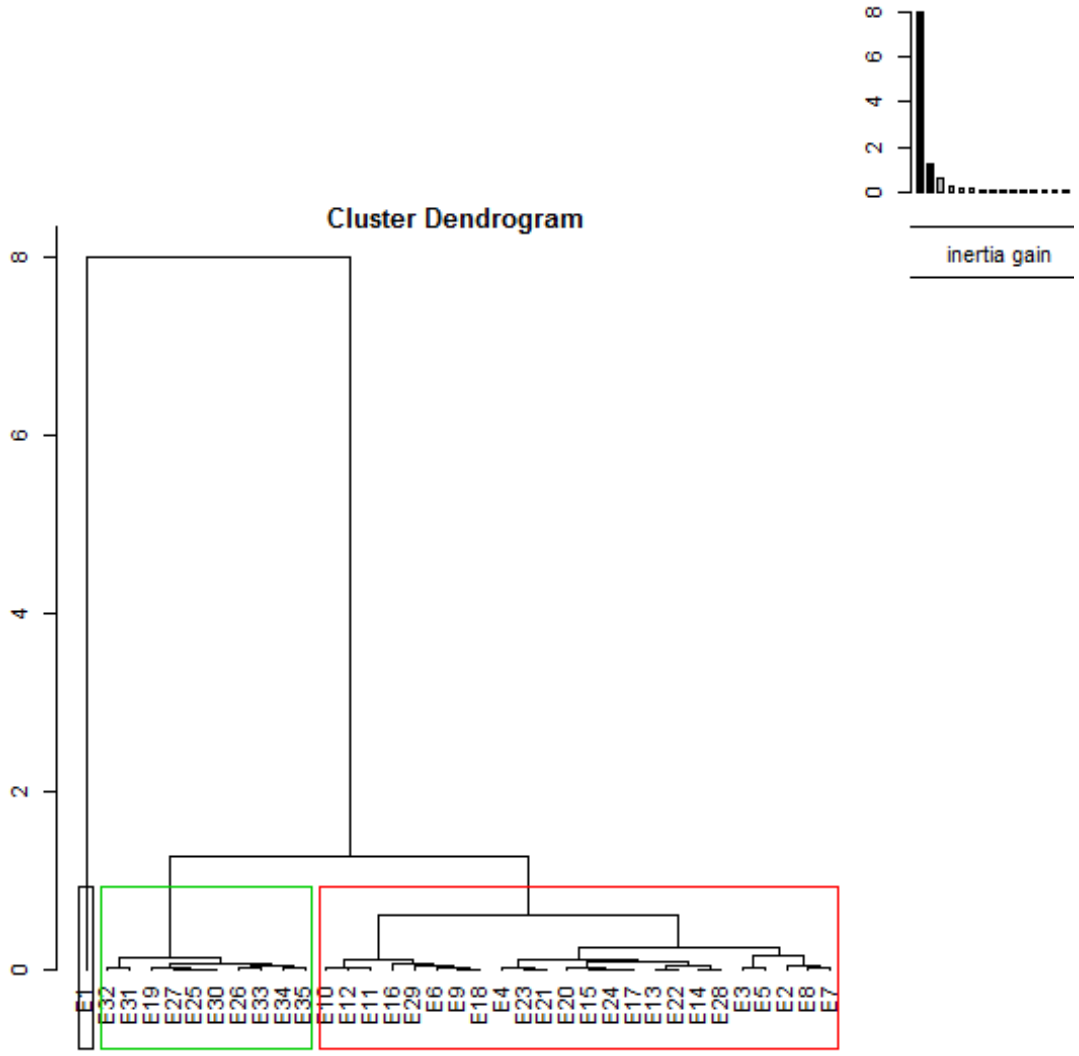


Figure 1.1 - Arbre hiérarchique.

La classification réalisée sur les individus fait apparaître 3 classes. Le graphe des gains d'inertie nous montre qu'il y'a une grosse perte d'inertie lorsqu'on passe de 3 à 2 classe. Donc on va conserver les 3 classes. Le dendrogramme est plat pour tous les 3 groupes. Donc chaque groupe est homogène. On remarque cependant qu'il est moins plat pour le deuxième groupe(en rouge). Donc celui-ci est moins homogène que les autres groupes.

2) Représentation des individus dans le plan.

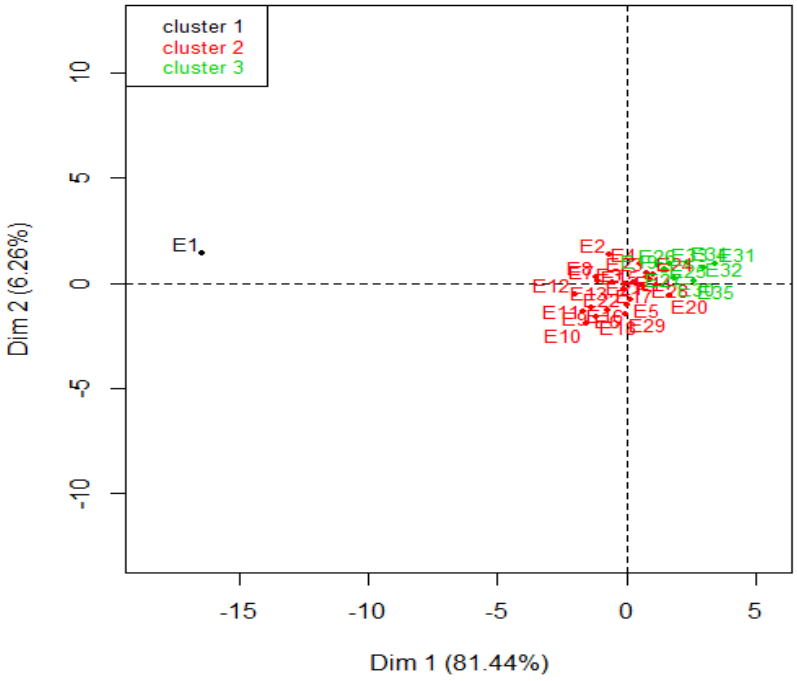


Figure 1.2 - Classification Ascendante Hiérarchique des individus.

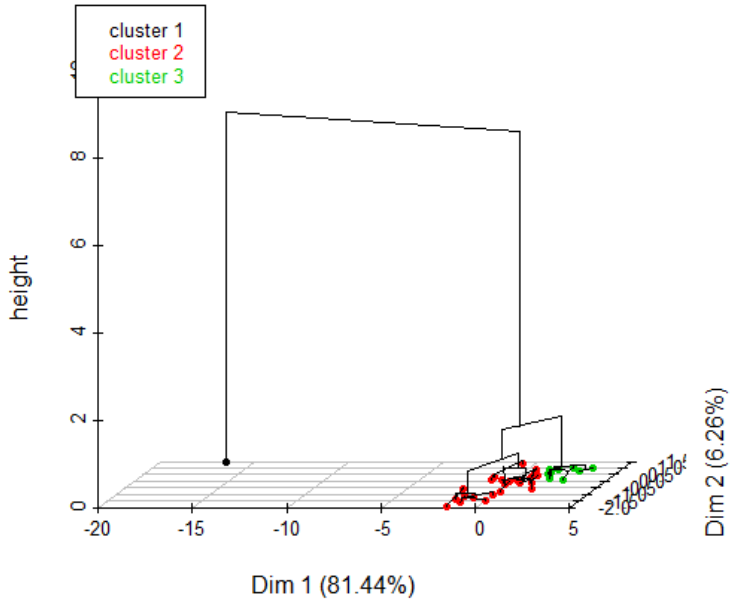


Figure 1.3 - Arbre hiérarchique sur le plan factoriel.

La représentation des individus dans le plan factoriel fait ressortir la présence d'un individu extrême, constituant à lui seul le premier groupe, très éloigné des autres. Ce qui laisse penser que cet individu pourrait être dû à une erreur ou une aberration. Les deux autres groupes sont relativement proches et sont regroupé au niveau du centre du centre du plan.

3) Caractérisation des classes

	Classe 1	Classe 2	Classe 3
M1	-4.54	-1.15	2.86
M10	-5.28	0.173	1.77
M11	-4.85	-0.0453	1.83
M12	-5.66	0.981	1.08
M2	-3.55	-1.79	3.15
M3	-5.05	0.127	1.73
M4	-4.47	-1.6	3.29
M5	-5.02	-0.242	2.1
M6	-5.01	0.395	1.44
M7	-4.09	-1.92	3.48
M8	-3.58	-1.04	2.39
M9	-5.53	0.519	1.5

Ce tableau nous montre que la classe 1 regroupe les individus ayant de très faible valeurs pour l'ensemble des 12 variables, la classe 2 regroupe les individus ayant des valeurs moyennes et la classe 3 les individus ayant de très grandes valeurs.

Tableau des individus spécifiques

Classe 1					
	E1				
Distance	16.36613				
Classe 2					
	E10	E12	E11	E18	E9
Distance	4.538397	4.269502	4.258827	4.000805	3.898765
Classe 3					
	E31	E32	E35	E34	E33
Distance	3.959971	3.497042	3.031181	3.025823	2.611292

4) Interprétation générale.

La **classe 1** est composée d'individus tels que *E1*. Ce groupe est caractérisé par : de faibles valeurs pour des variables telles que *M12*, *M9*, *M10*, *M3*, *M5*, *M6*, *M11*, *M1*, *M4* et *M7*.

La **classe 2** est composée d'individus tels que *E2*, *E4*, *E6*, *E9*, *E10*, *E11*, *E16*, *E18* et *E29*. Ce groupe est caractérisé par: des variables dont les valeurs ne diffèrent pas significativement de la moyenne.

La **classe 3** est composée d'individus tels que *E26*, *E31*, *E32*, *E33* et *E34*. Ce groupe est caractérisé par: de fortes valeurs pour les variables *M7*, *M4*, *M2*, *M1*, *M8* et *M5*.

IV. CONCLUSION

La classification que nous avons réalisée sur R nous a permis de partitionner les étudiants de cette classe en 3 groupes. Le premier groupe est constitué d'un seul individu (E1) qui a de très faibles notes sur l'ensemble des 12 matières et donc une moyenne très faible par rapport à la moyenne de la classe. Le second groupe est composé des étudiants qui ont une moyenne proche de la moyenne de la classe ; et le troisième groupe est constitué des étudiants ayant une moyenne élevée par rapport à la moyenne de la classe.

CHAPITRE 2 : REGRESSION LOGISTIQUE

I. INTRODUCTION

La régression logistique est une méthode de classification supervisée (discrimination) qui permet d'établir un lien entre une variable à expliquer qualitative et des variables explicatives quantitatives ou binaires. Son objectif est de prédire et/ou expliquer une variable catégorielle Y à partir d'une collection de descripteurs X = (X1, X2, . . . , XJ). Il s'agit en quelque sorte de mettre en évidence l'existence d'une liaison fonctionnelle sous-jacente de la forme Y = f(X, α) entre ces variables. La fonction f(.) est le modèle de prédiction ; α est le vecteur des paramètres de la fonction, on doit en estimer les valeurs à partir des données disponibles.

Dans ce chapitre, nous réaliserons une régression logistique binaire avec le logiciel R sur un jeu de données comportant 20 observations et 3 variables prédictives. L'objectif est de prédire la présence ou l'absence d'un problème cardiaque (COEUR - Y ; avec "présence" = "1" et "absence" = "0") à partir de son AGE (quantitative - X1), du TAUX MAX (quantitative - X2) et l'occurrence d'une ANGINE de poitrine (binaire - X3).

II. MISE EN ŒUVRE DE LA REGRESSION LOGISTIQUE BINAIRE SUR R

1) Etude préliminaire des données

Notre jeu de donnée est importé dans R grâce à la syntaxe :

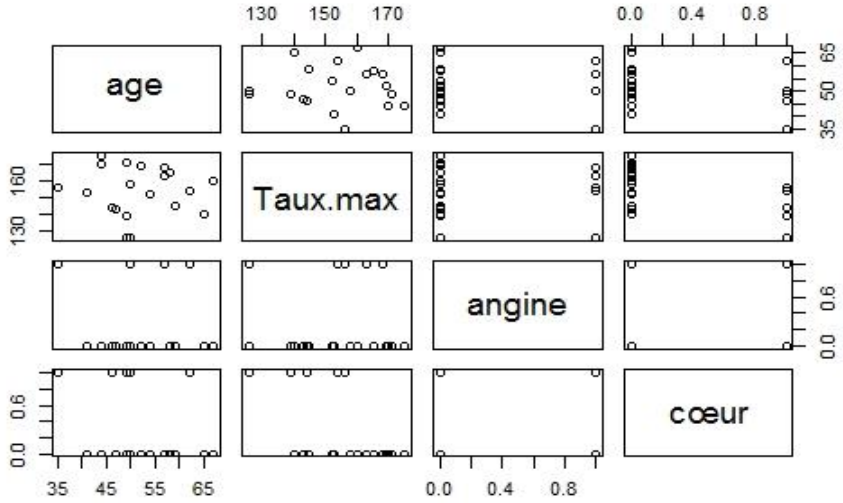
```
mydata<- read.csv(choose.files(), header = TRUE, sep = ";")
```

La variable Coeur est codée de sorte que : presence = 1 et absence = 0.

```
mydata$cœur<- ifelse(mydata$cœur == "presence", 1, 0)
```

Puis, on observe la structure de notre jeu de données et les relations entre les différentes variables de notre jeu de données.

```
> str(mydata)
'data.frame':  20 obs. of  4 variables:
 $ age     : int  50 49 46 49 62 35 67 65 47 58 ...
 $ Taux.max: int  126 126 144 139 154 156 160 140 143 165 ...
 $ angine  : int  1 0 0 0 1 1 0 0 0 0 ...
 $ cœur    : num  1 1 1 1 1 1 0 0 0 0 ...
> plot(mydata)
```



Sur ce graphique nous pouvons voir sur ce graphique nous pouvons voir qu'il n'existe pas une véritable relation entre les problèmes cardiaque (cœur) et l'angine. Par contre, les problèmes cardiaques sont beaucoup plus liés aux variables age et Taux.max. Donc ces variables sont les plus pertinentes pour l'explication de notre modèle.

2) Développement du modèle de prédiction

Nous allons établir notre premier modèle (model1) à partir des variables explicatives age, Taux.max et angine. La syntaxe sur R est :

```
model1 <- glm(formula = cœur~age+Taux.max+angine, family = "binomial", data = mydata)

summary(model1)
```

On obtient les résultats suivant :

```
Warning messages:
1: glm.fit: l'algorithme n'a pas convergé
2: glm.fit: des probabilités ont été ajustées numériquement à 0 ou 1
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  12277.13 2356951.64   0.005    0.996
age            -92.11   17920.19  -0.005    0.996
Taux.max       -55.71   10680.02  -0.005    0.996
angine        2033.02  391994.71   0.005    0.996
```

R nous signale que l'algorithme n'a pas convergé. Et on peut constater que les p-value (pr(>|z|)) ne sont pas très significative (car proche de 1). Donc ce modèle n'est pas bon.

Nous allons construire un second modèle en omettant la variable angine car elle n'est pas pertinente.

Le deuxième modèle:

```
> model2 <- glm(formula = cœur~age+Taux.max, family = "binomial", data = my
data)
> summary(model2)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 25.65196   11.80432   2.173   0.0298 *
age         -0.10658    0.07984  -1.335   0.1819
Taux.max    -0.14036    0.06831  -2.055   0.0399 *
```

Dans ce modèle, les p-value obtenues sont significatives (proches de 0). Donc notre modèle est bon. Les paramètres estimés sont dans la colonne « Estimate » du tableau précédent.

On a donc :

$$Logit(p) = \ln(\frac{p}{1-p}) = 25{,}65196 - 0{,}10658 * age - 0{,}14036 * Taux.max$$

3) Prédire avec le modèle

Nous pouvons à présent faire des prédictions sur R avec notre modèle.

Nous allons prédire s'il y'a présence ou pas de problèmes cardiaques chez deux individus (i1 et i2) ayant les caractéristiques suivantes :
I1 : age = 55, Taux.max = 175
I2 : age = 45, Taux.max = 125
On fixe le seuil à 50%.
Le code et les résultats sur R :

```
> i1 <- data.frame(age = 55, Taux.max = 175)
> predict(model2, i1, type = 'response')
          1
0.008390094
> i2 <- data.frame(age = 45, Taux.max = 125)
> predict(model2, i2, type = 'response')
        1
0.9648161
```

Nous obtenons un résultat de 0,84% < 50% pour le premier individu contre 96,48% > 50% pour le deuxième. Donc notre modèle prévoit que l'individu i1 n'a pas de problème cardiaque et que l'individu i2 a un problème cardiaque.

III. CONCLUSION

La régression logistique binaire que nous avons réalisée nous a permis de développer un modèle nous permettant de prévoir la présence ou pas de problème cardiaque chez un individu connaissant son âge et son Taux.max.

CHAPITRE 3 : ANALYSE DISCRIMINANTE

I. INTRODUCTION

L'**analyse factorielle discriminante** (**AFD**) ou simplement **analyse discriminante** est une technique statistique qui vise à décrire, expliquer et prédire l'appartenance à des groupes prédéfinis d'un ensemble d'observations à partir d'une série de variables prédictives. Deux modèles d'Analyse Factorielle Discriminante sont possibles en fonction d'une hypothèse fondamentale : **L'Analyse Factorielle Discriminante Linéaire** : si l'on suppose que les matrices de covariance sont identiques et **L'Analyse Factorielle Discriminante Quadratique** : si l'on suppose au contraire que les matrices de covariance sont différentes pour au moins deux groupes.

Dans ce chapitre, nous effectuerons une analyse discriminante linéaire sur un jeu de donnée comportant 10 individus (clients d'une banque) repartis en 3 groupes de risque financier (haut risque, risque moyen et risque faible) et 4 variables explicatives. L'objectif est de mieux discriminer les groupes et prédire le groupe d'appartenance d'un éventuel individu à partir de son âge, son revenu, son patrimoine et le montant de son emprunt.

II. MISE EN ŒUVRE DE L'ANALYSE DISCRIMINANTE LINEAIRE SUR R

Notre jeu de données est importé sur R à l'aide de la syntaxe :

```
mydata <- read.csv(file.choose(), header = TRUE, sep = ";", dec = ",", row.names = 1)
```

Ensuite, on regarde la structure de notre jeu de donnée:

```
> str(mydata)
'data.frame':  10 obs. of  5 variables:
 $ Age       : int  45 47 38 36 29 39 27 51 32 35
 $ Revenu    : int  250000 160000 165000 175000 99000 170000 120000 160000
155000 170000
 $ Patrimoine: int  1300000 1150000 850000 770000 450000 1400000 1400000 13
00000 1500000 1400000
 $ Emprunt   : int  600000 450000 370000 250000 400000 120000 160000 320000
350000 180000
 $ Groupe    : int  3 2 1 1 1 3 2 3 2 2
```

Ensuite, on effectue l'analyse discriminante linéaire sur notre jeu de donnée à l'aide de la fonction lda (linear discriminant analysis) du package MASS.

Le code sur R est le suivant :

```
library(MASS)

linear <- lda(Groupe~., data = mydata)

linear
```

On obtient les résultats suivants :

```
Prior probabilities of groups:
  1   2   3
0.3 0.4 0.3

Group means:
       Age   Revenu Patrimoine  Emprunt
1 34.33333 146333.3     690000 340000.0
2 35.25000 151250.0    1362500 285000.0
3 45.00000 193333.3    1333333 346666.7

Coefficients of linear discriminants:
                    LD1           LD2
Age        -6.346552e-02  1.123216e-01
Revenu      2.333414e-05  1.942416e-05
Patrimoine -9.463388e-06 -1.568186e-06
Emprunt    -4.734906e-06 -3.519063e-06

Proportion of trace:
   LD1    LD2
0.9336 0.0664
```

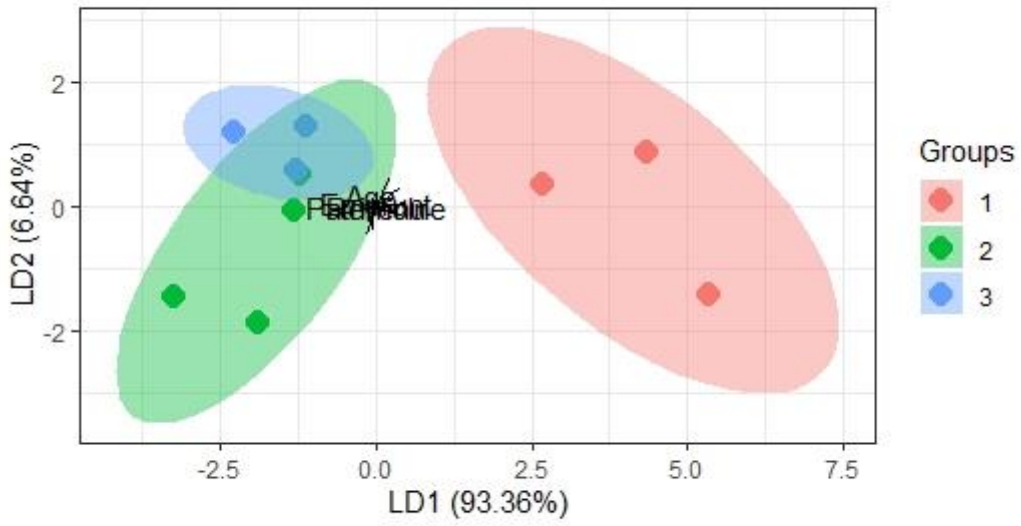
Le premier résultat est le pourcentage d'individu par groupe 30% d'individus dans le premier groupe, 40% dans le second et 30% dans le troisième. Ensuite, on a les moyennes par groupe de chaque variable explicative.
Puis, nous avons les coefficients de l'analyse discriminante linéaire pour deux modèles : LD1 et LD2. Ces modèles sont nos fonctions discriminantes, qui s'écrivent comme combinaison linéaire des variables explicatives à l'aide de ces coefficients.
Pour le modèle LD1, on a :

$$Y1 = -6{,}35.10^{-2} * age + 2{,}33.10^{-5} * Revenu - 9{,}46.10^{-6} * Patrimoine - 4{,}73.10^{-6} * Emprunt$$

Le dernier résultat est le pourcentage de discrimination apporté par chaque modèle. Ainsi, nous pouvons dire le pourcentage de discrimination apporté par la première fonction discriminante (premier modèle : LD1) est de 93,36% ; contre 6,64% pour la seconde.

Le code nous permettant d'afficher le graphe des individus discriminés dans le plan est :

```
#convertir la variable Groupe en facteur
mydata$Groupe <- as.factor(mydata$Groupe)
#le graphe
library(devtools)
library(ggord)
ggord(linear, mydata$Groupe)
```



On peut voir sur ce graphe que le premier groupe est très éloigné des deux autres qui sont quant à eux très proche, presque confondu. La première fonction discriminante est responsable 93,36% de la séparation et la deuxième fonction de 6,64% de la séparation.

Matrice de confusion et dégré de précision du modèle

Pour obtenir la matrice de confusion et le degré de précision de notre modèle, on utilise le code suivant sur R :

```
#Matrice de confusion
p1 <- predict(linear, mydata)$class
tab <- table(predicted = p1, Actual  = mydata$Groupe)
tab
#degré de précision
sum(diag(tab))/sum(tab)
```

On obtient le résultat suivant dans R:

```
> tab
         Actual
predicted 1 2 3
        1 3 0 0
        2 0 3 0
        3 0 1 3
> sum(diag(tab))/sum(tab)
[1] 0.9
```

Ainsi, notre modèle a eu une précision de 90% dans la prédiction du groupe de risque des individus de notre jeu de donnée (une seule mauvaise prédiction).

Prédiction sur un nouvel individu

Considérons un nouvel individu ayant les caractéristiques suivantes : Age = 50, Revenu = 1205000, Patrimoine = 5000000, Emprunt = 10000000. On veut déterminer son groupe de risque financier à partir de notre modèle.

Le code nous permettant de prédire son groupe sur R est :

```
i1 <- data.frame(Age = 50, Revenu = 1205000, Patrimoine = 5000000, Emprunt = 10000000)

predict(linear, i1, type = 'response')$class
```

on obtient comme résultat:

```
> predict(linear, i1, type = 'response')$class
[1] 2
Levels: 1 2 3
```

Donc cet individu appartient au deuxième groupe de risqué financier (risque moyen).

III. CONCLUSION

L'analyse discriminante linéaire que nous avons effectuée à l'aide du logiciel R nous a permis de discriminer les 3 groupes de risque financier (haut risque, risque moyen, risque faible) à l'aide de fonctions discriminantes (LD et LD2). Celle-ci nous a aussi permis de construire un modèle pouvant prédire le groupe de risque d'un individu connaissant son âge, son revenu, son patrimoine et le montant emprunté. Le degré de précision de notre modèle sur notre jeu de donnée est de 90%.