Mathematics for Data Science

Examen TP 2022-2023

M1 APP Big Data and Machine Learning

- Da Cruz Mathis (20220090)
- Sivananthan Sarankan (20221430)

Table des matières

1. Introduction

L'étude porte sur le cancer de la prostate. On s'intéresse ici à des données contenant des informations sur des personnes atteintes du cancer de la prostate.

Pour dresser le contexte, la prostate est une glande du système reproducteur masculin. Son cancer se développe à partir des tissus de la prostate, lorsque ses cellules mutent et se propagent de manière incontrôlée. Celles-ci peuvent ensuite atteindre d'autres parties du corps.

L'antigène de la prostate est une protéine sécrétée par les cellules de la prostate. Normalement, cet antigène est voué à être détruit par le système immunitaire qui va produire un anticorps une fois que cet antigène aura été détecté. Or dans le cas d'une personne atteinte d'un cancer, cet antigène est produit en grande quantité et les cellules de la prostate en décrètent 10 fois plus que la normale. De plus, certains facteurs peuvent être responsables de cette augmentation (volume de la prostate, infection, inflammation), et d'autres peuvent être responsables de sa diminution.

L'objectif de cette étude est alors de mettre en lumière les facteurs influençant le taux d'antigène produit par la prostate. Les données qu'on utilise pour cette étude, proviennent de patients (hommes) atteints d'un cancer de la prostate et ayant subi une prostatectomie radicale (ablation chirurgicale totale de la prostate.

2. Analyse des données

2.1. Description du jeu de données

Pour notre étude, nous disposons des données suivantes :

- Volume du cancer (vol).
- Poids de la prostate (wwt).
- Age du patient (âge).
- Quantité d'hyperplasie bénigne (bh). L'hyperplasie bénigne est une tumeur bénigne de la prostate.
- Pénétration capsulaire (pc). Plus la variable est élevée, plus le cancer traverse la capsule entourant la prostate pour atteindre les structures voisines.
- Taux spécifique d'antigène de la prostate (psa).

2.2. Première analyse : description du jeu de données

1. *Combien y a-t-il d'observations ? Présentez les données et l'objectif de l'étude statistique.*

L'étude porte sur 80 observations (patients atteint du cancer de la prostate) et 6 variables (volume du cancer, poids de la prostate, âge du patient, quantité d'hyperplasie bénigne, pénétration capsulaire, taux spécifique d'antigène de la prostate) dans notre dataset.

Le nombre d'observations avant et après suppression des valeurs NA reste le même donc il n'y a pas de valeurs nulles.

L'objectif de cette étude est de mieux comprendre les facteurs influençant le taux d'antigène prostatique spécifique.

2. *Utilisez la commande summary() pour calculer des statistiques descriptives pour la variable cible psa. Interprétez les résultats.*

Interprétons maintenant les statistiques de la variable psa, représentant le taux spécifique d'antigène de la prostate d'une personne.

Dans notre population étudiée, le taux minimum de psa est de 0.60 et le taux maximum est de 265.850.

En ce qui concerne les quartiles, le taux spécifique d'antigène de la prostate au-dessous duquel :

Se situe 25% des taux est de 6.125 (1st Qu)

Se situe 50% des taux est de 14.400 (Median)

Se situe 75% des taux est de 21.350 (3rd Qu)

On voit que le maximum, par rapport au troisième quartile, est bien supérieur. On peut donc le considérer comme une valeur suprême.

3. *Effectuer une première analyse de la variable psa basée sur les autres en calculant le coefficient corrélation entre psa et chacune des autres variables. Laquelle est la plus corrélée avec psa ?*

Le coefficient de corrélation entre psa et age est de 0.01304884. Il est très proche de 0 donc la variation de l'une a peu d'impact sur celle de l'autre et inversement (corrélation positive faible).

Le coefficient de corrélation entre psa et bh est de -0.02203714. Il est négatif et proche de 0 donc quand les valeurs d'une variable augmentent les valeurs de l'autre variable diminuent, cependant il est proche de 0, donc la variation reste minime (corrélation négative faible)

Le coefficient de corrélation entre psa et pc est de 0.5957111. Il est proche de 1 donc quand les valeurs d'une variable augmentent alors celles de l'autre aussi.

La variable volume est donc celle la plus corrélée avec psa.

On peut supposer que les variables psa, pc et vol (coefficients de corrélation les plus élevés) permettent de représenter un maximum d'informations ? (À vérifier avec l'ACP)

En traçant le plot(prostate) (cf Figure 1 - Plot prostate), on remarque une grande accumulation de poids sur une même zone ou ligne, ce problème peut être dû à la différence d'échelle entre chaque variable.

4. *Effectuez une transformation logarithmique de tous les ensembles de données de variables à l'exception de l'âge et modifiez les noms des variables transformées en faisant précéder le nom de la lettre l (exemple lvol au lieu de vol). Visualisez à nouveau le Nuages de points et interpréter les résultats.*

En traçant les nuages de point de la variable psa et des autres points, on voit que les graphes affichent des grandes concentrations de points dans des petites zones. Pour remédier à cela, passer à une échelle logarithmique va nous permettre de "zoomer" sur ces zones d'accumulation. En d'autres termes, l'échelle logarithmique va nous permettre d'observer une large gamme de valeurs sur une petite zone.

Une fois cette transformation logarithmique faite (cf Figure 2 - Plot lprostate) (sauf sur la variable âge), on observe bien une corrélation positive entre les variables psa/vol et vol/pc. On observe également une corrélation entre psa et pc, mais elle est moins visible.

3. Analyse en composantes principales (ACP)

3.1. Question théorique : Si deux variables sont parfaitement corrélées dans le jeu de données, serait-il approprié de les inclure toutes les deux dans l'analyse lors de la réalisation de l'ACP? Justifiez votre réponse.

Oui car à eux deux, elles représentent 100% de l'information. En effet, lorsque le coefficient de corrélation r = 1, les deux variables sont parfaitement corrélées positivement. Cela signifie qu'une augmentation d'une unité d'une variable se traduira par une augmentation d'une unité de l'autre variable et inversement.

3.2. Application pratique : La fonction apply() permet d'appliquer une fonction à chaque ligne ou colonne du jeu de données. Pour exemple, la commande apply(prostate, 2, mean) permet de calculer la moyenne empirique de chaque variable. Calculez la variance de chaque variable et interprétez les résultats.

Étudions à présent la variance de ces variables après transformation logarithmique. On observe que toutes les variables, à l'exception de la variable âge à laquelle on n'a pas appliqué de log, ont une petite variance (inférieure à 1). Cela témoigne du fait qu'elles ne sont pas très dispersées autour de leur moyenne.

On peut comparer ces variances, aux variances avant de passer à l'échelle logarithmique : la variance des variables avant le passage au logarithme est bien supérieure, et donc ces variables seraient sans le logarithme, bien plus dispersées autour de leur moyenne.

3.3. Pensez-vous qu'il est-il nécessaire de normaliser les variables avant d'effectuer l'ACP pour ce jeu de données ? Pourquoi ? Pour réaliser l'ACP, la fonction PCA peut être utilisée par lignes de commandes à l'aide du package FactoMineR ou grâce à une interface graphique via le package Factoshiny.

Oui, il est nécessaire de normaliser les variables avant d'effectuer l'ACP, afin de réduire la variance de l'âge (variance trop élevée comparée aux autres variables). On doit faire en sorte que l'âge soit comparable aux autres variables (uniformisation des variables).

3.4. Effectuez l'ACP à l'aide de la fonction PCA() avec les arguments et options appropriés en tenant compte de votre analyse précédente. Analysez la sortie de cette fonction.

Analyse des graphes :

Cercle des corrélations (cf Figure 4 - Cercle des correlations):

Le cercle de corrélation permet de voir la liaison entre les variables. C'est-à-dire que plus l'angle entre les vecteurs est faible (proche de 0) plus la corrélation est forte (proche de 1). Si l'angle est presque

droit (~90°), cela signifie que ces deux variables ne sont pas liées. Et enfin, si l'angle formé est plat (~180°), la corrélation est proche de -1, i.e corrélation négative.

Dans notre cas, on observe deux groupes de vecteurs formant des angles faibles entre eux : on distingue un groupe composé des variables lbh, age, lwht et le groupe composé de lpsa, lvol, lpc. On peut interpréter cela par le fait que lpsa (taux spécifique d'antigène de la prostate), lvol (volume du cancer) et lpc (pénétration capsulaire) sont corrélées positivement. Pareillement, les variables lbh, age, lwht sont aussi corrélées positivement.

On peut également dire que lbh et lpc ne sont pas corrélées (angle droit entre les deux vecteurs).

Ce graphe nous indique également qu'avec ces deux axes, dans une représentation donc, on arrive à représenter 46.47 + 22.99 = 69.46% de l'information contenue dans notre jeu de données initial.

PCA graphe des individus (cf Figure 3 - Graphe des individus):

Ce graphe représente la contribution de chaque individu en fonction de chaque axe.

L'ensemble d'individus est une masse homogène, cependant il existe certains extrêmes.

Les individus 1 et 80 contribuent beaucoup à l'axe 1 alors que les individus 50, 77 et 78 contribuent beaucoup à l'axe 2.

3.5. Interpréter les valeurs des deux premiers vecteurs de chargement des composantes principales? Plusieurs solutions existent pour déterminer le nombre d'axes à analyser en ACP. La plus courante consiste à représenter le diagramme en barres des valeurs propres ou des inerties associées à chaque axe Grâce à la fonction barplot.

Avec les deux premiers axes (psa et vol) on arrive à représenter 69.46201% de l'information.

Dans le barplot (cf Figure 5 - Graphe des valeurs propres et Figure 6 - Barplot) on voit que les 3 premiers axes permettent de représenter un maximum d'informations et que les 3 derniers ne rajoutent pas plus d'informations par rapport aux 3 premiers. En effet, le passage du 2eme au 3eme axe nous fait gagner encore 10% d'informations pour monter donc jusqu'à 81.18177% de l'information alors que les 4,5,6 nous rajoutent seulement quelques pourcents de l'information.

3.6. Tracez le PVE expliqué par chaque composant, ainsi que le PVE cumulé. Calculer le pourcentage de variance expliquée (PVE) par chaque composant ?

Le premier graphe (cf Figure 7 - % de variance expliquée par chaque composant) représente le % de variance expliquée par chaque composant alors que le second (cf Figure 8 - PVE cumulé) représente le PVE cumulée. Et on voit bien qu'avec les 3 premières composantes, on arrive à représenter beaucoup d'informations.

3.7. Combien de composants garderiez-vous ? Pourquoi ?

Comme vu dans la 3.5, les 2 premiers axes représentent 69% de l'information contre 81% de l'information pour les 3 premiers axes. Afin de passer le pallier des 70% considéré comme un bon pourcentage de représentation de l'information, nous opterons pour une représentation avec les 3 premiers composants.

4.Régression linéaire

4.1. Question théorique : Supposons que l'on fait un modèlede régression linéaire simple pour expliquer Y comme une fonction linéaire de X. Quelle est la relation entre, la corrélationcoefficient entre ces deux variables r (X; Y) et le coefficientde détermination R2 obtenu paradapter le modèle ? Quelle est la plage de valeurs que peut prendre r ?

Si on suppose qu'on fait un modèle de régression linéaire simple pour expliquer une variable Y en fonction d'une variable X, la relation entre le coefficient de corrélation r(X,Y) et le coefficient de détermination R^2 est: R^2 = r(X,Y)^2.

Le coefficient de corrélation est compris entre -1 et 1. Donc le coefficient de détermination est compris entre 0 et 1.

4.2. Application pratique : Calculez la corrélation entre la variable lpsa et les autres variables existant dans le jeu de données. Notons X la variable la plus corrélée avec lpsa et considérons le modèle de régression linéaire simple suivante : (1). Ajuster le modèle donné en (1) et répondre aux questions suivantes :

Le coefficient de corrélation entre lpsa et lvol est de 0.7858116. Il est proche de 1 donc quand les valeurs d'une variable augmentent alors celles de l'autre aussi.

Le coefficient de corrélation entre lpsa et lwht est de 0.4558655. Il a une corrélation positive.

Le coefficient de corrélation entre lpsa et ages est de 0.1755921. Il est proche de 0 donc la variation de age a peu d'impact sur celle de lpsa et inversement (corrélation positive faible).

Le coefficient de corrélation entre lpsa et lbh est de 0.1927745. Il est proche de 0 donc la variation de lbh a peu d'impact sur celle de lpsa et inversement (corrélation positive faible).

Le coefficient de corrélation entre lpsa et lpc est de 0.5545791. Il est proche de 1 donc quand les valeurs d'une variable augmentent alors celles de l'autre aussi.

(cf Figure 9 - Graphique de la régression linéaire)

4.3. Quelles sont les estimations des coefficients ? Interpréter l'estimation du coefficient

L'équation de la droite de régression est : b1^ * X + b0^ avec b1^ = 0.79560 représentant l'estimation du coefficient directeur et b0^ = 0.64360 qui est l'estimation de l'ordonnée à l'origine de notre modèle.

On interprète donc b1^ par le fait qu'à chaque fois que le volume de cancer augmente de 1 (cm3?) la quantité de psa augmente de 0.79560 (ng/ml?)

De plus quand le volume de cancer est de 0, i.e. quand on a pas de cancer, le taux normal de psa est de 0.64360 (ng/ml?).

4.4. Élaborer le test d'hypothèse de pente nulle pour le coefficientβ1 et conclure s'il y a

Considérons les hypothèses de tests suivantes, pour un seuil d'erreur alpha = 5%:

H0: b1^ = 0

H1: b1^ != 0

On obtient de plus grâce aux résultats de la régression que t_value_calculée = 11.22147. On pouvait également la calculer à la main en divisant l'estimation du coefficient directeur par son Std.error.

Pour déterminer la t_value_théorique, on utilise la table de Students, avec un degré de liberté de 78 et un taux de confiance de 97,5 % (test bilatéral). On a alors que t_value_théorique = 1.990847.

On constate que t_valeur_calculée est > t_value_théorique. On est dans la zone de rejet, donc on rejette H0.

4.5. Une relation entre lpsa et X. β1 est-il significativement non nul ?

Comme vu dans la 4.4, on rejette H0 : b1^ = 0, donc une relation entre lpsa et X.b1^ sera significativement non nul. Donc une relation entre lpsa et X.b1_estimate n'est pas significativement nul avec un risque d'erreur de 5%

4.6. Quelle est la valeur du coefficient de détermination R2 ? Interprétez ce résultat. Ce modèle est-il adapté pour prédire le taux d'antigène spécifique de la prostate

R^2 = 0.6126, donc 61% de la variance est expliquée par le modèle de régression linéaire, le reste ne peut pas être expliqué par le modèle.

Il est plus proche de 0.5(faible) que de 0.9(fort). Notre modèle est donc assez moyen.

On peut supposer que le R^2 pourrait être plus fort avec lbh...

5. Annexe des graphes et des codes

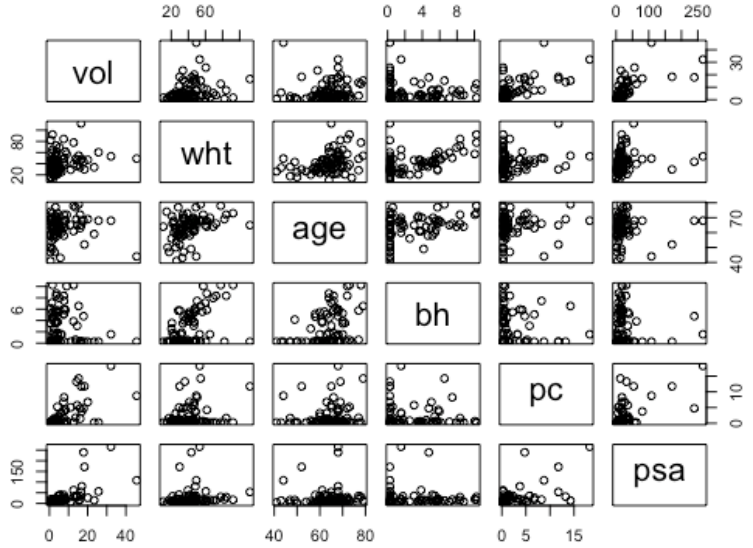

Figure 1 - Plot prostate

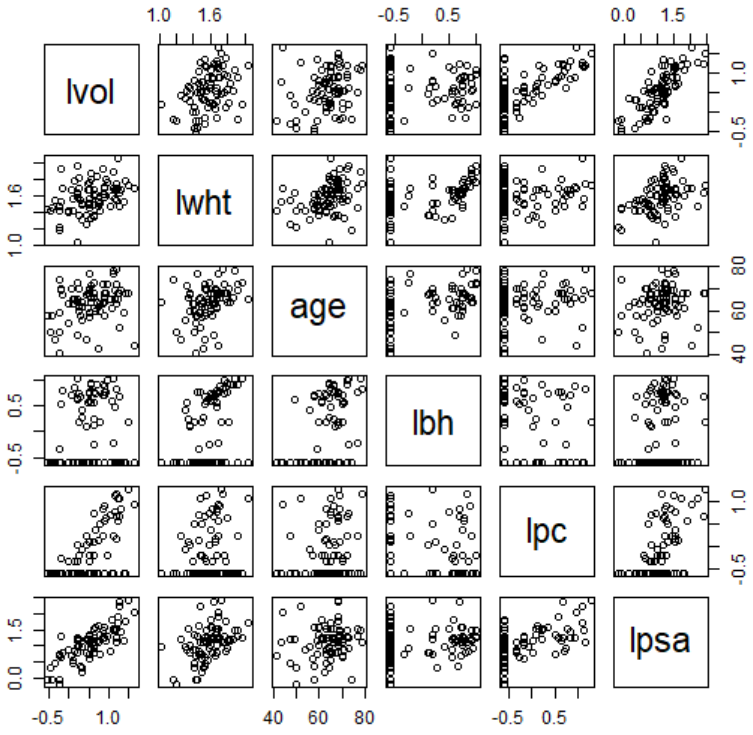

Figure 2 - Plot lprostate

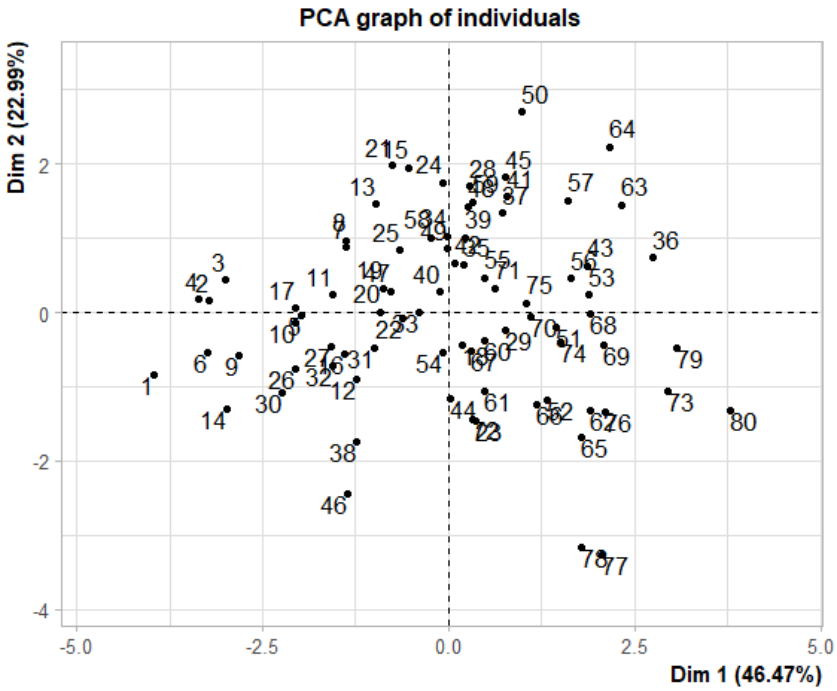

Figure 3 - Graphe des individus

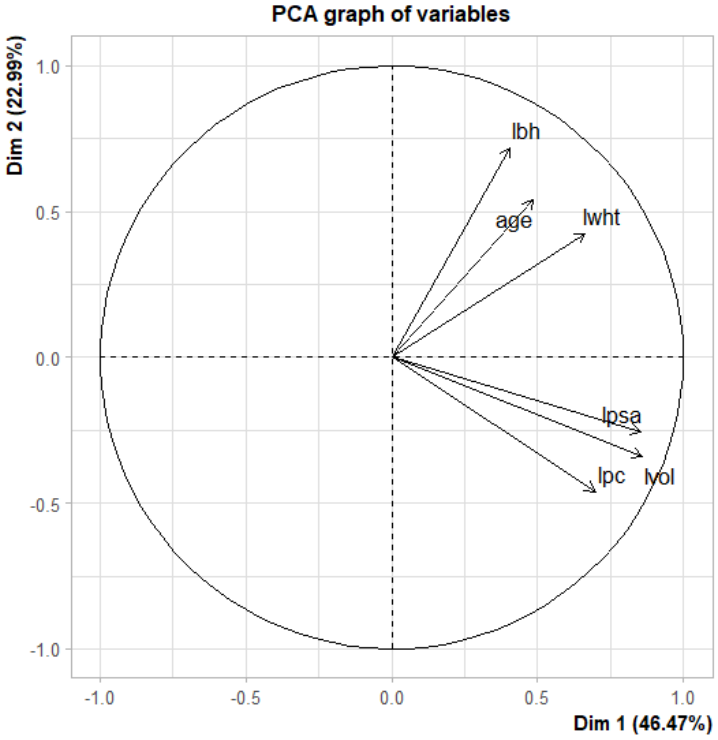

Figure 4 - Cercle des correlations

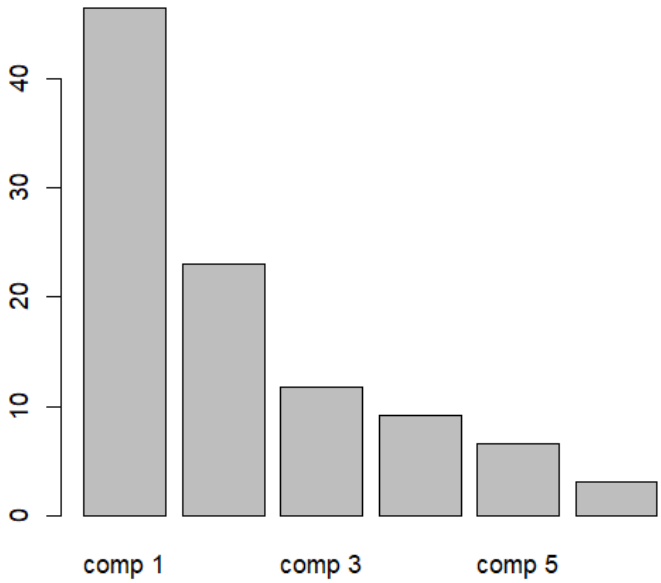

Figure 5 - Graphe des valeurs propres

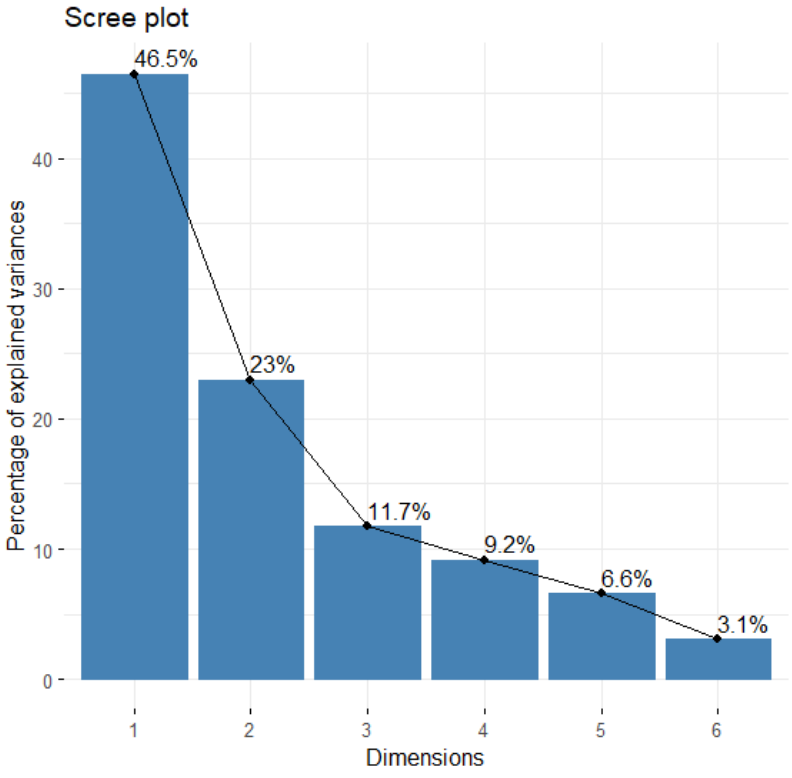

Figure 6 - Barplot

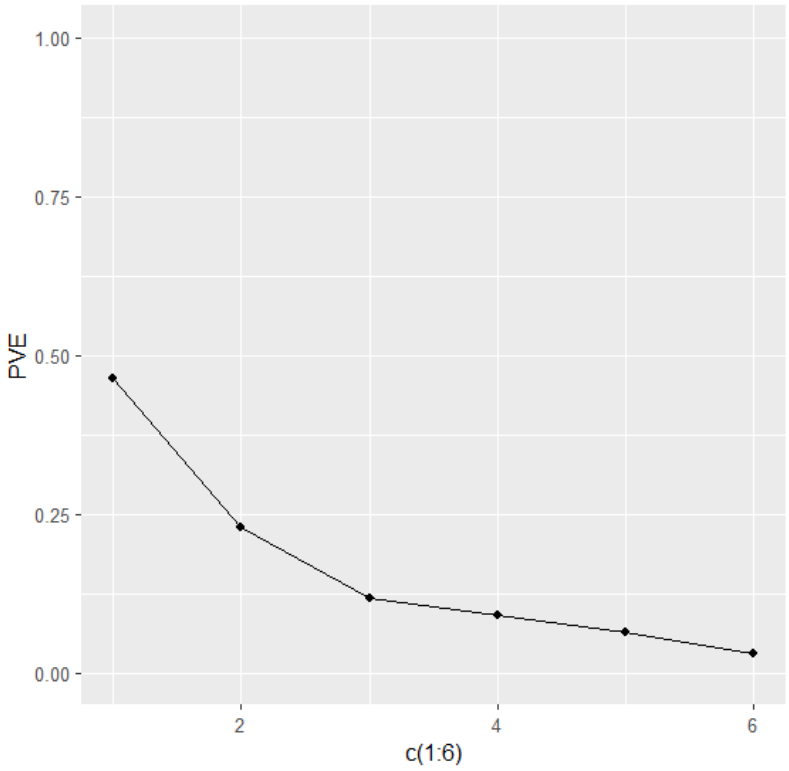

Figure 7 - % de variance expliquée par chaque composant

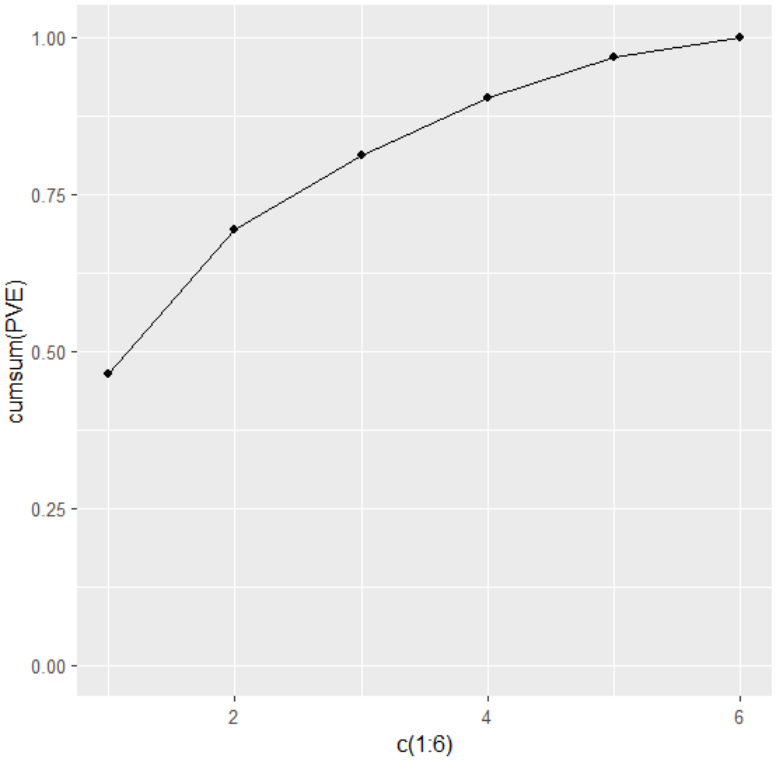

Figure 8 - PVE cumulé

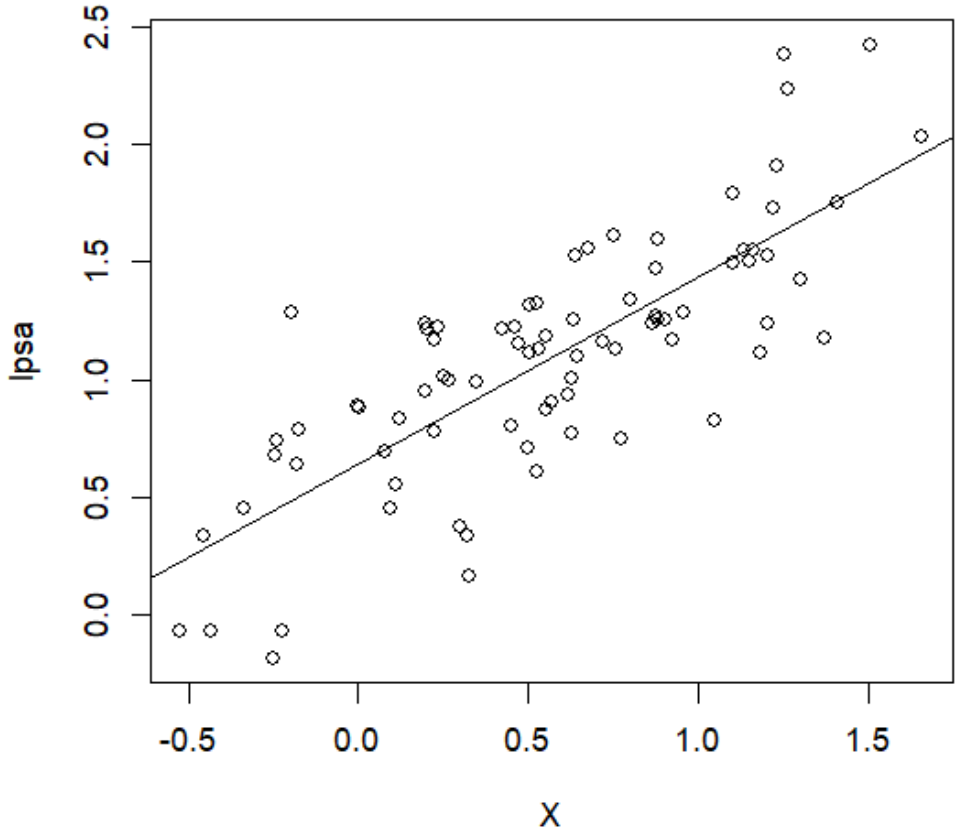

Figure 9 - Graphique de la régression linéaire

```r
library(FactoMineR)
library(factoextra)

#Importation csv dans var prostate

prostate <- read.table(file = "/Users/sarankansivananthan/Desktop/M1/Maths/TP_R/prostate.txt", header = TRUE , sep = " ", dec = ".")

#Premieres stats sur prostate

#2.2

#1

#Nb rows

nrow(prostate)
nrow(na.omit(prostate))

#nb lignes égales avant et après suppression des valeurs na => pas de valeurs na''
```

Figure 10 - Code 2.1

```r
#2

summary(prostate["psa"])

# le taux spécifique d'antigène de la prostate au-dessous duquel se situe 25% des taux est de 6.125 (1st Qu)
# le taux spécifique d'antigène de la prostate au-dessous duquel se situe 50% des taux est de 14.400 (Median)
# le taux spécifique d'antigène de la prostate au-dessous duquel se situe 75% des taux est de 21.350 (3rd Qu)
# le taux spécifique d'antigène de la prostate moyen est de 25.473 (Mean)
#le max est bien supérieur au troisième quertile. On peut le considérer comme une valeur extrême
```

Figure 11 - Code 2.2

```r
#3

cor(prostate["psa"] , prostate["vol"]) # 0.6664723 => coef proche de 1 ie quand les valeurs d'une var augmentent alors celles de l'autre aussi
cor(prostate["psa"] , prostate["wht"]) # 0.1662757 => coef proche de 0 ie corrélation positive faible
cor(prostate["psa"] , prostate["age"]) # 0.01304884 => coef proche de 0 ie corrélation  positive faible A CHANGER
cor(prostate["psa"] , prostate["bh"]) # -0.02203714 => corr nég ie quand les val d'une var augmentent les valeurs de l'autre var diminuent A CHANGER
cor(prostate["psa"] , prostate["pc"]) # 0.5957111 => coef proche de 1 ie quand les valeurs d'une var augmentent alors celles de l'autre aussi

# => vol est plus corrélée avec psa
# => on peut supposer que les var psa, vol, pc permettraient de representer un max d'information? (à vérifier avec l'ACP)

plot(prostate) # graphes en fonction de chaques variables
```

Figure 12 - Code 2.3

```r
#4
attach(prostate)

lvol = log10(vol)
lwht = log10(wht)
age = age
lbh = log10(bh)
lpc = log10(pc)
lpsa = log10(psa)

lprostate = data.frame("lvol" = lvol , "lwht" = lwht , age , "lbh" = lbh , "lpc" = lpc , "lpsa" = lpsa)

plot(lprostate) #graphes en fonction de chaques varibles

# Interpretations
  # corr linéaire positive entre vol et psa
  # aussi entre vol et pc
  # moins visible pour psa / pc
```

Figure 13 - Code 2.4

```r
#3.2

apply(lprostate , 2 ,var) # plus variance est petite moins la variable a des valeurs dispersées ie moins elle varie par rapport à la moyenne

apply(prostate , 2 ,var)

# on voit bien que le log a permis de réduire la dispersion, ex: psa: 1963 -> 1,44...
```

Figure 14 - Code 3.2

```r
#3.3

#oui pour réduire la variance de l'age (variance trop élevée), faire en sorte que l'age soit comparable aux autres variables

prostate.cr = scale(lprostate , center = TRUE , scale = TRUE)
```

Figure 15 - Code 3.3

```r
#3.4
resultat.PCA = PCA(prostate.cr , graph = T)

names(resultat.PCA)

resultat.PCA$eig

summary(resultat.PCA)
```

Figure 16 - Code 3.4

```r
#3.5

# avec les deux premiers axes on arrive à représenter 69.462% de l'information
resultat.PCA$eig


barplot(resultat.PCA$eig[,2])
fviz_eig(resultat.PCA , addlabels=TRUE)

# On voit que les 3 premiers axes permettent de représenter le max d'info et que les 3 derniers ne rajooutent pas plus d'infos par rapport aux 3 premiers
```

Figure 17 - Code 3.5

```r
#3.6
PVE <- resultat.PCA$eig[,2]/ sum(resultat.PCA$eig[,2])

qplot(c(1:6), PVE) + geom_line() + ylim(0, 1)
qplot(c(1:6), cumsum(PVE)) + geom_line() + ylim(0, 1)
```

Figure 18 - Code 3.6

```r
#4.2


cor(lpsa , lvol) # 0.7858116 => coef proche de 1 ie quand les valeurs d'une var augmentent alors celles de l'autre aussi
cor(lpsa , lwht) # 0.4558655 => coef proche de 0 ie corrélation positive faible
cor(lpsa , prostate["age"]) # 0.1755921 => coef proche de 0 ie corrélation  positive faible A CHANGER
cor(lpsa , lbh) # 0.1927745 => corr nég ie quand les val d'une var augmentent les valeurs de l'autre var diminuent A CHANGER
cor(lpsa , lpc) # 0.5545791 => coef proche de 1 ie quand les valeurs d'une var augmentent alors celles de l'autre aussi


X = lvol

model = lm(lpsa ~ X , data = lprostate)

summary(model)

plot(X, lpsa)
abline(model)

#61% dela variance est expliqué par le modèle de régression
#linéaire, le reste ne peut pas être expliqué
```

Figure 19 - Code 4.2

```r
#4.3
model$coefficients

# le coeff directeur dit qu'il y  aune correlation positive faible
# Quand on a pas de cancer, le taux normal de psa est de ...
# Quand le volume de cancer augmente de 1%, la quantité de psa augmente de 0,8%


sum = summary(model)

sum$coefficients[c(3)]
```

Figure 20 - Code 4.3

```r
#4.4
t.val.calc = sum$coefficients[c(6)] # = sum$coefficients[c(2)] / sum$coefficients[c(4)]

t.val.theo = qt(0.975, df = 78) # niveau d'erreur de 5%. Test de Student bilatéral => 2.5% de part et d'autres
```

Figure 21 - Code 4.4