Rapport Examen TP

Professeur : RACHED, Imen

Module : ADIF72

Table des matières

Introduction

Dans le domaine émergent et dynamique de la science des données, l'analyse et l'interprétation des ensembles de données complexes et volumineux sont cruciales pour extraire des connaissances précieuses. Le présent TP s'inscrit dans cette démarche, en se focalisant sur l'analyse d'un ensemble de données spécifique, "prostate.txt", qui compile des informations détaillées sur des patients atteints de cancer de la prostate.

Le cancer de la prostate, caractérisé par la mutation incontrôlable des cellules de la prostate, est un sujet de préoccupation majeur dans le domaine médical et la recherche clinique. Une compréhension approfondie des facteurs qui influent sur le taux d'antigène prostatique spécifique (PSA) peut être déterminante dans le diagnostic précoce et la gestion efficace de cette maladie.

Ce TP vise à naviguer à travers l'ensemble de données, à analyser et à interpréter les variables pour offrir des insights et des informations précieuses concernant les facteurs qui influent sur le taux de PSA. Nous commencerons par une première analyse où une description détaillée du jeu de données. Ensuite, nous réaliserons une analyse en composantes principales (ACP). Enfin, nous finirons par

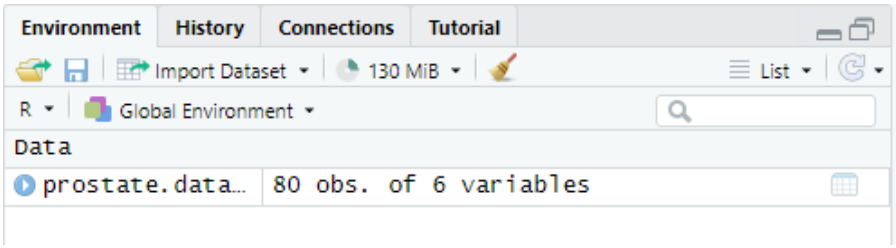
Dans le cadre de ce travail pratique, nous allons nous intéresser à l'analyse d'un jeu de données relatif aux niveaux d'antigène spécifique de la prostate (PSA). L'objectif principal est d'explorer et de comprendre les relations entre différentes variables et la manière dont elles influencent le taux de PSA. Pour ce faire, nous avons adopté une approche méthodique en suivant plusieurs étapes clés d'analyse statistique.

Nous commencerons par une première analyse où une description détaillée du jeu de données sera effectuée pour comprendre sa structure, sa composition et ses variables, notamment le volume du cancer, le poids de la prostate, l'âge du patient, la quantité d'hyperplasie bénigne, la propagation dans les vésicules séminales et la pénétration capsulaire. Ensuite, une analyse en composantes principales (ACP) sera réalisée pour réduire la dimensionnalité de l'ensemble de données et identifier les composantes principales qui capturent la majeure partie de la variance au sein des données. Enfin, nous finirons par une analyse de régression linéaire pour établir des relations prédictives entre les différentes variables. L'objectif sera d'évaluer comment les variables indépendantes, telles que le volume du cancer, influent sur la variable dépendante, à savoir le taux de PSA.

I. Première analyse : description du jeu de données

Importez le jeu de données prostate.txt :

Pour importer notre jeu de données : prostate dataset.txt, il nous suffit d'aller dans environnement, de clicker sur import Dataset, puis de cliquer From Text(base)... puis de sélectionner notre fichier. Une fois faits-nous allons pouvoir retrouver notre jeu de donnée dans notre environnement.



TP_Noté_SENECHAL_Morgan_MathFor... *
prostate.dataset
Filter

	vol	wht	age	bh	pc	psa
1	0.56	15.95	50	0.25	0.25	0.65
2	0.37	27.65	58	0.25	0.25	0.85
3	0.60	14.75	74	0.25	0.25	0.85
4	0.30	26.65	58	0.25	0.25	0.85
5	2.12	30.95	62	0.25	0.25	1.45
6	0.35	25.25	50	0.25	0.25	2.15
7	2.09	32.25	64	1.85	0.25	2.15
8	2.00	34.45	58	4.65	0.25	2.35
9	0.46	34.45	47	0.25	0.25	2.85
10	1.25	25.65	63	0.25	0.25	2.85
11	1.29	36.75	65	0.25	0.25	3.55
12	3.34	31.25	57	0.25	0.65	4.05
13	0.66	33.65	70	3.47	0.55	4.35
14	0.57	26.25	41	0.25	0.25	4.75
15	1.20	45.85	70	5.25	0.25	4.95

Showing 1 to 16 of 80 entries, 6 total columns

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

Dans un premier temps, nous allons afficher les premières lignes du jeu de donnée :

```
> # Afficher les premières lignes du jeu de données
> head(prostate.dataset)
   vol   wht age   bh   pc  psa
1 0.56 15.95  50 0.25 0.25 0.65
2 0.37 27.65  58 0.25 0.25 0.85
3 0.60 14.75  74 0.25 0.25 0.85
4 0.30 26.65  58 0.25 0.25 0.85
5 2.12 30.95  62 0.25 0.25 1.45
6 0.35 25.25  50 0.25 0.25 2.15
```

Ensuite nous allons afficher le nombre total d'observations :

```
> # Afficher le nombre total d'observations
> cat("Nombre d'observations :", nrow(prostate.dataset), "\n")
Nombre d'observations : 80
```

Enfin affichons le résumé statistique des variables de notre jeu de données :

```
> # Afficher un résumé statistique des variables
> cat("Résumé statistique des variables :\n")
Résumé statistique des variables :
> print(summary(prostate.dataset))
      vol              wht              age              bh               pc              psa        
 Min.   : 0.300   Min.   : 10.75   Min.   :41.00   Min.   : 0.250   Min.   : 0.250   Min.   :  0.650  
 1st Qu.: 1.650   1st Qu.: 29.20   1st Qu.:60.00   1st Qu.: 0.250   1st Qu.: 0.250   1st Qu.:  6.125  
 Median : 3.565   Median : 38.30   Median :65.00   Median : 1.300   Median : 0.450   Median : 14.400  
 Mean   : 6.771   Mean   : 41.58   Mean   :63.61   Mean   : 2.692   Mean   : 2.189   Mean   : 25.473  
 3rd Qu.: 8.060   3rd Qu.: 48.48   3rd Qu.:68.00   3rd Qu.: 5.075   3rd Qu.: 1.875   3rd Qu.: 21.350  
 Max.   :45.650   Max.   :111.95   Max.   :79.00   Max.   :10.240   Max.   :18.250   Max.   :265.850  
```

Objectif de l'étude statistique

L'objectif de cette étude statistique, est de mieux comprendre les facteurs influençant le taux d'antigène prostatique spécifique (PSA). Le taux de PSA est souvent utilisé comme un indicateur pour le diagnostic du cancer de la prostate. Cependant, de nombreux facteurs peuvent affecter le taux de PSA, tels que le volume de la prostate, les infections, l'inflammation et certains traitements.

En analysant le jeu de données, qui comprend des variables telles que le volume du cancer, le poids de la prostate, l'âge du patient, la quantité d'hyperplasie bénigne, la propagation dans les vésicules séminales et la pénétration capsulaire, les chercheurs peuvent identifier des modèles et des relations entre ces variables et le taux de PSA. Cette connaissance peut ensuite être utilisée pour améliorer la précision du diagnostic et le traitement du cancer de la prostate.

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

Utilisation de la commande summary() pour les statistiques descriptives pour la variable TARGET psa :

```
> # Calcule des statistiques descriptives pour la variable cible psa
> cat("Statistiques descriptives pour la variable cible psa :\n")
Statistiques descriptives pour la variable cible psa :
> print(summary(prostate.dataset$psa))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  0.650   6.125  14.400  25.473  21.350 265.850 
```

Interprétation des résultats :

Les résultats de la commande summary() nous donne une idée claire de la distribution des valeurs de la variable psa :

- **Min :** La valeur minimale du PSA dans l'ensemble de données.
- **1er Qu. :** Le premier quartile indique que 25% des patients ont un niveau de PSA inférieur à cette valeur. C'est également la médiane du premier demi-ensemble de données.
- **Médiane :** La médiane sépare l'ensemble de données en deux parties égales. 50% des valeurs de PSA sont inférieures à la médiane et 50% sont supérieures. Cela peut aussi être interprété comme le deuxième quartile.
- **Moyenne :** La moyenne de PSA, calculée en additionnant tous les niveaux de PSA et en divisant par le nombre total d'observations.
- **3e Qu. :** Le troisième quartile indique que 75% des patients ont un niveau de PSA inférieur à cette valeur. C'est également la médiane du second demi-ensemble de données.
- **Max :** La valeur maximale du PSA dans l'ensemble de données.

Ces statistiques nous aideront à comprendre la distribution de la variable cible psa, identifiant ainsi les tendances centrales et la dispersion des données.

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 ?

Calcule du coefficient de corrélation :

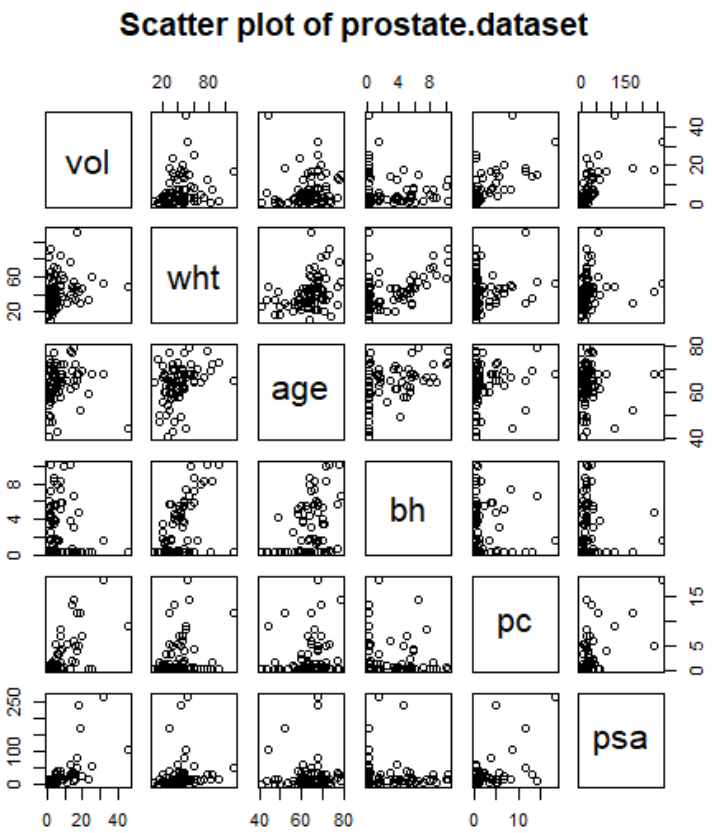
```
> # calcule du coefficient de corrélation entre psa et les autres variables
> correlations <- cor(prostate.dataset)[,"psa"]
> correlations <- correlations[-length(correlations)]  # Exclure la corrélation de psa avec elle-même
>
> # Afficher les coefficients de corrélation
> cat("Coefficients de corrélation entre psa et les autres variables :\n")
Coefficients de corrélation entre psa et les autres variables :
> print(correlations)
        vol         wht         age          bh          pc
 0.66647230  0.16627567  0.01304884 -0.02203714  0.59571112
```

```
> # Identifier la variable la plus corrélée avec psa
> max_corr_var <- names(which.max(abs(correlations)))
> cat("La variable la plus corrélée avec psa est :", max_corr_var, "avec un coefficient de corrélation de", round(correlations[max_corr_var], 3), "\n")
La variable la plus corrélée avec psa est : vol avec un coefficient de corrélation de 0.666
```

Nuage de point de prostate.dataset | Traiter les accumulations de points en utilisant le logarithme :

```
# Tracer le nuages de points
plot(prostate.dataset,main="Scatter plot of prostate.dataset")
```

Output :



Interprétation :

1. Scatter plot of prostate.dataset :

Il s'agit du graphique de l'ensemble de données original.

vol vs. wht : Il y a une tendance linéaire positive, indiquant que lorsque le volume du cancer (vol) augmente, le poids de la prostate (wht) tend également à augmenter.

vol vs. psa et wht vs. psa : Il semble y avoir une tendance exponentielle positive, montrant que des augmentations dans le volume du cancer ou le poids de la prostate sont associées à de plus grandes augmentations du PSA. Cela suggère que le PSA pourrait augmenter de manière disproportionnée avec ces facteurs.

age : La distribution de l'âge semble assez uniforme par rapport aux autres variables. Il n'y a pas de tendance linéaire claire entre l'âge et le PSA.

bh vs. psa : Encore une fois, il y a une tendance qui semble exponentielle positive entre l'hyperplasie bénigne (bh) et le PSA.

pc vs. psa : Cette relation semble également suivre une tendance exponentielle positive, suggérant que des augmentations de la pénétration capsulaire (pc) pourraient être associées à de plus grandes augmentations du PSA.

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 pointset interpréter les résultats.

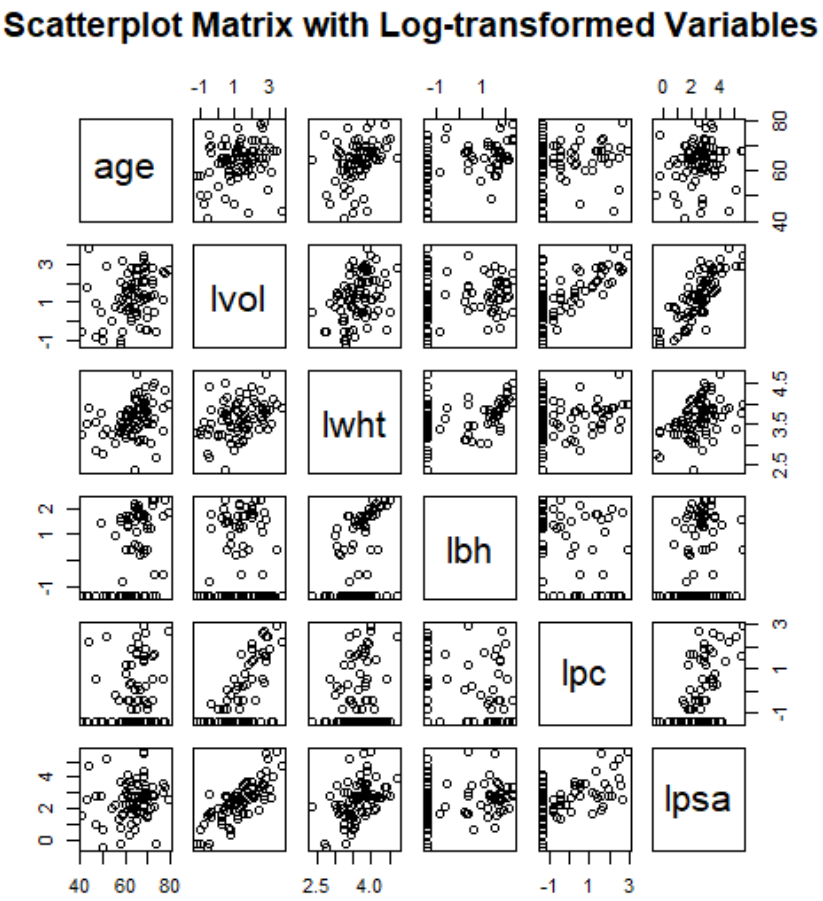
Transformation logarithmique :

```
> # Appliquer la transformation logarithmique aux variables spécifiées
> prostate.dataset$lvol <- log(prostate.dataset$vol)
> prostate.dataset$lwht <- log(prostate.dataset$wht)
> prostate.dataset$lbh <- log(prostate.dataset$bh)
> prostate.dataset$lpc <- log(prostate.dataset$pc)
> prostate.dataset$lpsa <- log(prostate.dataset$psa)
>
> # Supprimer les colonnes non-transformées, à l'exception de 'age'
> prostate.dataset$vol <- NULL
> prostate.dataset$wht <- NULL
> prostate.dataset$bh <- NULL
> prostate.dataset$pc <- NULL
> prostate.dataset$psa <- NULL
```

Visualiser à nouveau le scatterplot matrix :

```
> # Créer un scatterplot matrix avec les variables transformées
> pairs(prostate.dataset, main = "Scatterplot Matrix with Log-transformed Variables")
```

Output :



Interprétation :

Observons la matrice de nuages de points (scatterplot matrix) des variables transformées logarithmiquement :

Comme précédemment, les graphiques sur la diagonale principale montrent la distribution des données pour chaque variable individuellement.

lvol (Volume du cancer logarithmiquement transformé) :

lvol montre une forte corrélation positive avec lpsa. L'augmentation du volume du cancer semble être associée à une augmentation du taux de PSA.

Il y a aussi une relation positive entre lvol et lwht, ce qui est logique car une prostate plus grande peut avoir un plus grand volume de cancer.

lwht (Poids de la prostate logarithmiquement transformé) :

lwht semble avoir une corrélation positive modérée avec lpsa. Cela pourrait suggérer que le poids de la prostate peut influencer le taux de PSA.

La relation entre lwht et lvol est également clairement positive, comme mentionné ci-dessus.

age (Âge du patient) :

L'âge ne semble pas avoir de forte corrélation avec les autres variables logarithmiquement transformées, y compris lpsa. Cela pourrait suggérer que l'âge, par lui-même, n'est pas un prédicteur fort du taux de PSA.

lbh (Hyperplasie bénigne logarithmiquement transformée) :

lbh ne montre pas de corrélation claire avec lpsa. Cela indique que l'hyperplasie bénigne, même après transformation logarithmique, n'a pas une relation linéaire forte avec le taux de PSA dans cet ensemble de données.

lpc (Pénétration capsulaire logarithmiquement transformée) :

lpc montre une corrélation positive modérée forte avec lpsa. Cela suggère que plus la pénétration capsulaire est élevée (le cancer se propage au-delà de la prostate), plus le taux de PSA est élevé.

Une relation positive est également observée entre lpc et lvol.

lpsa (Taux d'antigène prostatique spécifique logarithmiquement transformé) :

Comme discuté, lpsa montre des corrélations positives avec lvol, lwht, et lpc.

En conclusion, après avoir effectué la transformation logarithmique et examiné le scatterplot matrix, il semble que lvol et lpc soient les variables qui ont la plus forte relation avec lpsa. Cela pourrait indiquer que le volume du cancer et la pénétration capsulaire sont des facteurs déterminants du taux de PSA. La transformation logarithmique a permis de clarifier ces relations.

II. Analyse en composantes principales (ACP)

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

Non, si deux variables sont parfaitement corrélées dans le jeu de données, il ne serait pas approprié de les inclure toutes les deux lors de la réalisation de l'ACP.

Voici quatre points expliquant pourquoi :

Redondance d'information : L'ACP est une technique utilisée pour réduire la dimensionnalité des données en conservant autant d'information que possible. Si deux variables sont parfaitement corrélées, elles fournissent essentiellement la même information. Inclure ces deux variables ne ferait qu'ajouter de la redondance sans ajouter d'information nouvelle ou utile.

Multicollinéarité : La multicollinéarité se produit lorsque deux variables ou plus sont fortement corrélées. Elle peut compliquer l'estimation des coefficients dans les modèles de régression et réduire la précision des prévisions.

Interprétation des composantes : La présence de variables parfaitement corrélées peut également compliquer l'interprétation des composantes principales. Les coefficients de ces variables dans les vecteurs propres seraient identiques ou très proches, rendant difficile la distinction de l'importance relative de chaque variable dans les composantes.

Efficacité : En éliminant la redondance, nous pouvons réduire la dimensionnalité de vos données plus efficacement et obtenir des composantes principales qui capturent l'essentiel de la variation dans vos données avec moins de composantes.

Pour toutes ces raisons, si nous identifions des variables qui sont parfaitement (ou même très fortement) corrélées, il est généralement recommandé de ne conserver qu'une seule de ces variables pour l'ACP, ou d'envisager d'autres méthodes pour gérer la multicollinéarité avant d'appliquer l'ACP.

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.

Calcule de la variance de chaque variable dans notre jeu de données : apply()

```
> # Calculer la variance de chaque variable
> variances <- apply(prostate.dataset, 2, var)
>
> # Afficher les variances
> print(variances)
       age       lvol       lwht        lbh        lpc       lpsa
62.3669304  1.4085737  0.1853071  2.1625915  1.8679841  1.4438723
```

Interprétation des résultats :

La variance mesure la dispersion des données autour de la moyenne. Une variance élevée indique que les valeurs sont éloignées de la moyenne, tandis qu'une faible variance indique que les valeurs sont proches de la moyenne.

Variables avec une dispersion élevée : age et lbh montrent les dispersions les plus élevées parmi les variables. Cela signifie que ces caractéristiques varient considérablement parmi les patients de l'ensemble de données.

Variables avec une dispersion modérée : lvol, lpc, et lpsa ont des variances modérées, indiquant une variation significative, mais pas aussi prononcée que pour age et lbh.

Variables avec une dispersion faible : lwht a la dispersion la plus faible parmi les variables, ce qui suggère qu'après transformation logarithmique, le poids de la prostate est relativement constant parmi les patients.

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 généralement nécessaire de normaliser (ou standardiser) les variables avant d'effectuer un ACP, en particulier lorsque les variables sont mesurées à des échelles différentes ou ont des unités de mesure différentes.

Voici pourquoi quelque explication supplémentaire expliquant pourquoi :

Échelle des Variables : L'ACP est sensible à la magnitude des variables. Si une variable a une variance très élevée parce qu'elle est mesurée à une grande échelle, elle dominera les premières composantes principales, ce qui pourrait masquer les structures intéressantes dans les données.

Interprétation des Résultats : La normalisation permet d'interpréter les composantes principales en termes de la variance relative des variables plutôt qu'en termes de leur échelle absolue.

Distances Euclidiennes : L'ACP est basée sur la maximisation de la variance. Si les variables ne sont pas standardisées, les variables avec des échelles plus grandes auront un poids disproportionné dans la détermination de la distance.

Compatibilité avec d'autres Analyses : La standardisation facilite la comparaison des coefficients de chargement des composantes principales et rend les résultats plus compatibles avec d'autres analyses multivariées.

Pour notre jeu de données, étant donné les variances que nous avons obtenus précédemment, il est clair que les variables sont à des échelles différentes (par exemple, var age =62.3 par rapport à lwht=0.18). Par conséquent, il serait approprié de normaliser les variables avant d'effectuer l'ACP pour s'assurer que chaque variable est traitée équitablement dans l'analyse.

Normalisation :

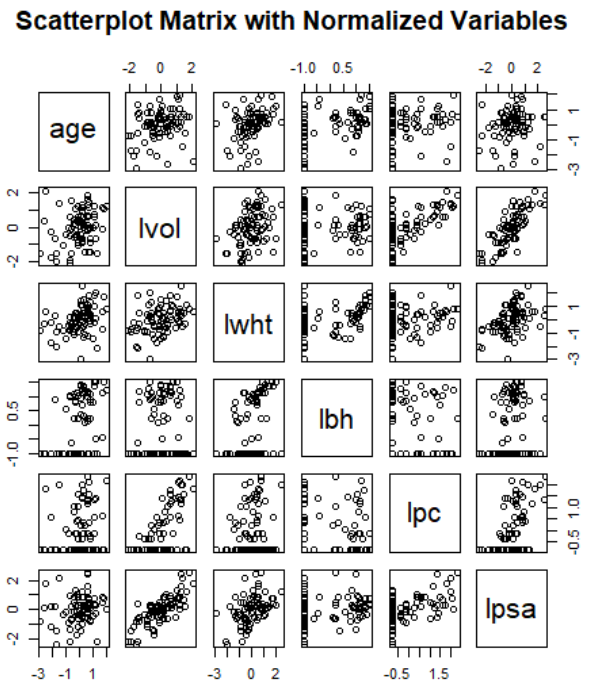
```
> # Fonction de normalisation
> normalize_data <- function(data){
+     return((data - mean(data)) / sd(data))
+ }
>
> # Normalisation des données
> prostate.dataset.normalized <- as.data.frame(lapply(prostate.dataset, normalize_data))
>
> # Afficher le résumé statistique des données normalisées pour vérification
> print(summary(prostate.dataset.normalized))
      age               lvol              lwht               lbh               lpc               lpsa
 Min.   :-2.8633   Min.   :-2.10007   Min.   :-2.93317   Min.   :-1.009   Min.   :-0.8199   Min.   :-2.4449
 1st Qu.:-0.4574   1st Qu.:-0.66387   1st Qu.:-0.61187   1st Qu.:-1.009   1st Qu.:-0.8199   1st Qu.:-0.5781
 Median : 0.1757   Median :-0.01458   Median : 0.01818   Median : 0.112   Median :-0.3898   Median : 0.1333
 Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.00000   Mean   : 0.000   Mean   : 0.0000   Mean   : 0.0000
 3rd Qu.: 0.5556   3rd Qu.: 0.67253   3rd Qu.: 0.56562   3rd Qu.: 1.039   3rd Qu.: 0.6498   3rd Qu.: 0.4610
 Max.   : 1.9485   Max.   : 2.13387   Max.   : 2.51002   Max.   : 1.516   Max.   : 2.3193   Max.   : 2.5598
>
```

Output :

	age	lvol	lwht	lbh	lpc	lpsa
1	-1.72369614	-1.57416893	-2.01661167	-1.00860192	-0.81990997	-2.44489530
2	-0.71068831	-1.92336179	-0.73855980	-1.00860192	-0.81990997	-2.22164195
3	1.31532735	-1.51603704	-2.19830886	-1.00860192	-0.81990997	-2.22164195
4	-0.71068831	-2.10006774	-0.82413212	-1.00860192	-0.81990997	-2.22164195
5	-0.20418439	-0.45249970	-0.47664499	-1.00860192	-0.81990997	-1.77717042
6	-1.72369614	-1.97018374	-0.94948955	-1.00860192	-0.81990997	-1.44935731
7	0.04906757	-0.46450814	-0.38106399	0.35241653	-0.81990997	-1.44935731
8	-0.71068831	-0.50159581	-0.22776531	0.97916555	-0.81990997	-1.37533387
9	-2.10357407	-1.73991273	-0.22776531	-1.00860192	-0.81990997	-1.21479651
10	-0.07755841	-0.89761062	-0.91297759	-1.00860192	-0.81990997	-1.21479651

```
> # Créer un scatterplot matrix avec les variables normalisées
> pairs(prostate.dataset.normalized, main = "Scatterplot Matrix with Normalized Variables")
```

Output :



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.

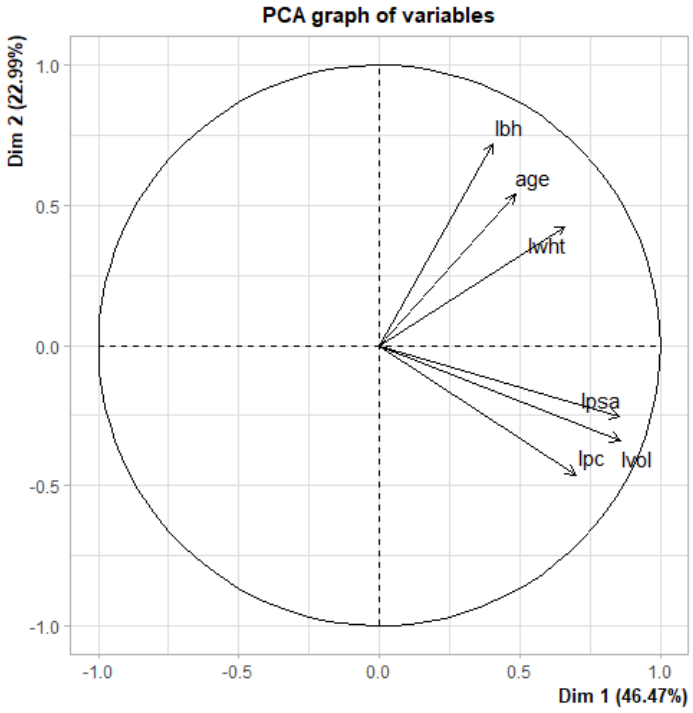
Installation et chargement du package FactoMiner :

```
> install.packages("FactoMineR")
WARNING: Rtools is required to build R packages but is not currently installed. Pl
ease download and install the appropriate version of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/
Installation du package dans 'C:/Users/Morgan/AppData/Local/R/win-library/4.3'
(car 'lib' n'est pas spécifié)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.3/FactoMineR_2.9.zip'
Content type 'application/zip' length 3799984 bytes (3.6 MB)
downloaded 3.6 MB

le package 'FactoMineR' a été décompressé et les sommes MD5 ont été vérifiées avec
succés

Les packages binaires téléchargés sont dans
        C:\Users\Morgan\AppData\Local\Temp\RtmpITwUzh\downloaded_packages
> library(FactoMineR)
```

Effectuer l'ACP :

```
res.pca <- PCA(prostate.dataset, scale.unit=TRUE)
```

Output :



Interprétation :

Ce graphique est un "biplot" des variables, obtenu à partir de l'ACP que nous avons effectuée sur notre jeu de données prostate.dataset. Le biplot représente la contribution des variables à chacune des deux premières composantes principales.

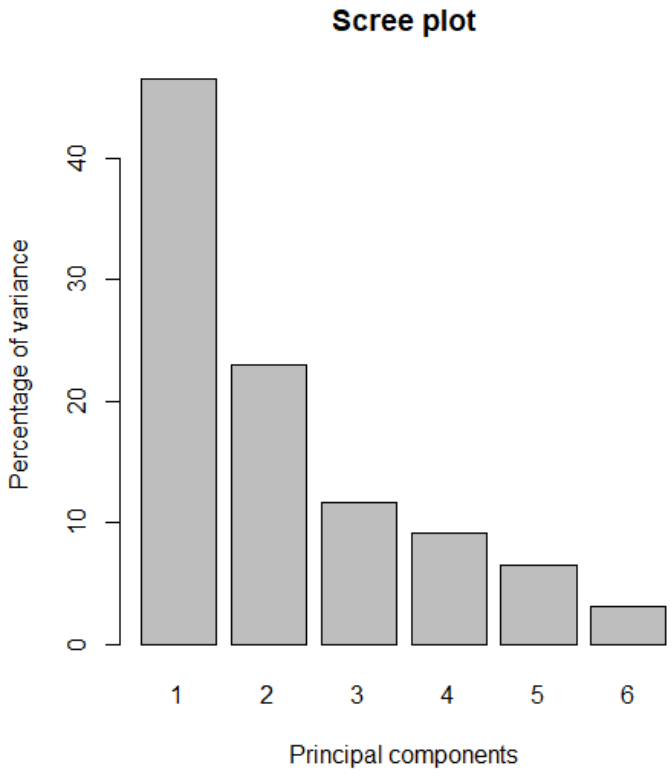
Analyse de la sortie | Variance expliquée :

```
> print(res.pca$eig)
       eigenvalue percentage of variance cumulative percentage of variance
comp 1  2.7883690              46.472817                           46.47282
comp 2  1.3793518              22.989197                           69.46201
comp 3  0.7031853              11.719755                           81.18177
comp 4  0.5494491               9.157486                           90.33925
comp 5  0.3933084               6.555140                           96.89439
comp 6  0.1863364               3.105606                          100.00000
```

Graphique des valeurs propres :

```
> barplot(res.pca$eig[, 2], names.arg=1:nrow(res.pca$eig), main="Scree plot", ylab="Percentage of variance", xlab="Principal components")
```

Output :



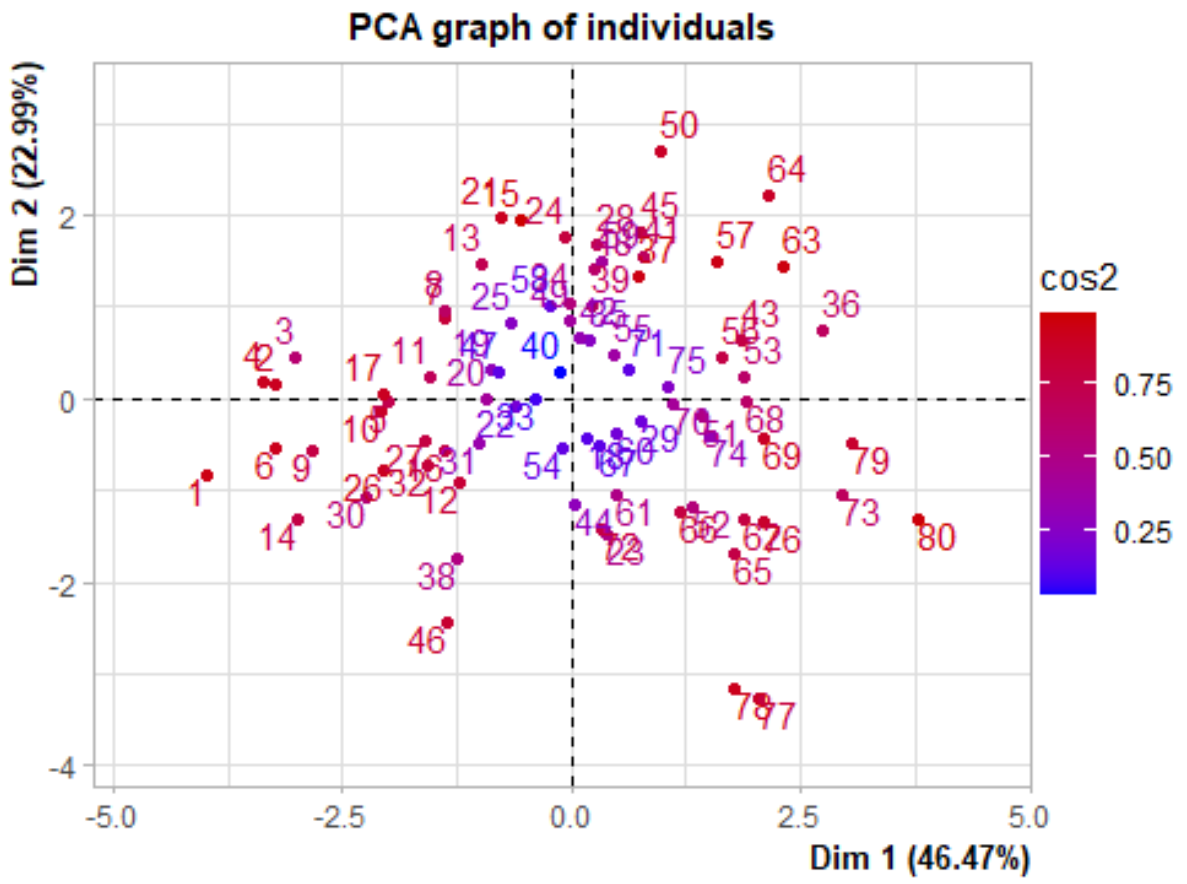
Interprétation :

CP1 et CP2 combinées expliquent près de 70% de la variance totale des données, ce qui est assez significatif. Les composantes suivantes (de CP3 à CP6) ajoutent de la variance supplémentaire, mais chacune d'elles explique progressivement moins de variance que les précédentes. Avec toutes les six composantes, 100% de la variance totale des données est expliquée. Cependant, en pratique, on pourrait considérer la réduction de dimensionnalité en ne conservant que les premières composantes (CP1 et CP2) pour une visualisation ou une analyse simplifiée, car elles capturent une grande partie de l'information.

Graphique des Individus :

```
plot(res.pca, choix="ind", habillage="cos2")
```

Output :



Interprétation :

Principal regroupement : La majorité des individus sont centrés autour de l'origine, ce qui suggère que la plupart des observations ont des valeurs moyennes pour les caractéristiques associées aux deux premières composantes.

Individus distincts : Quelques individus (par exemple, ceux situés à l'extrême gauche ou à l'extrême droite) sont assez éloignés du groupe central, ce qui suggère qu'ils ont des valeurs atypiques pour certaines caractéristiques.

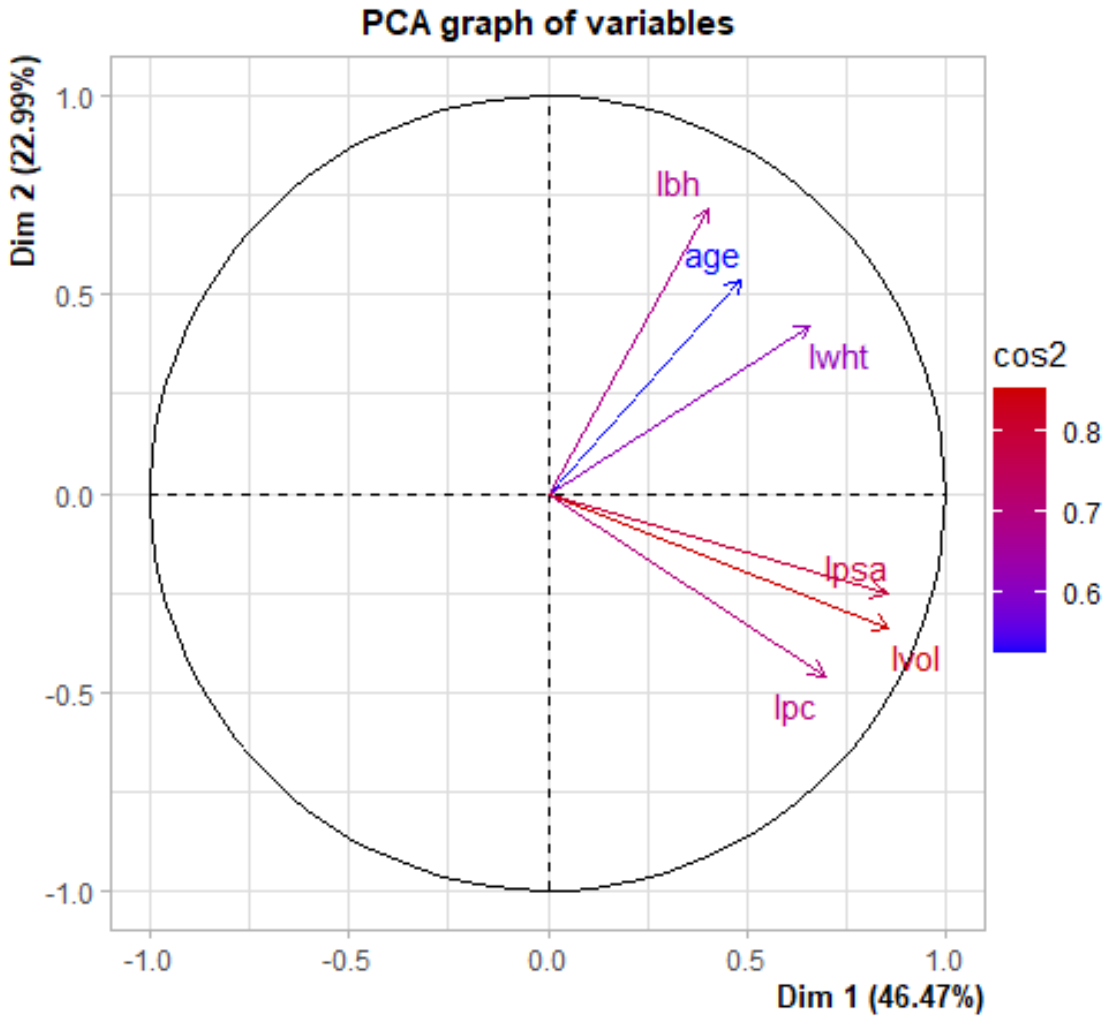
Qualité de la représentation : La plupart des individus sont en rouge ou en rose, ce qui indique qu'ils sont bien représentés par les deux premières composantes. Seuls quelques-uns, en bleu, sont moins bien représentés.

Cette visualisation nous offre un aperçu de la structure des données et de la manière dont les observations se comparent les unes aux autres sur la base des deux premières composantes principales.

Graphique des variables :

```
plot(res.pca, choix="var", habillage="cos2")
```

Output :



Interprétation :

Corrélation entre les variables : La proximité et l'orientation des vecteurs fournissent des informations sur la relation entre les variables. Par exemple, lvol, lpsa et lpc semblent avoir des comportements similaires dans l'ensemble de données.

Importance des variables : Les variables dont les vecteurs sont longs (comme lpsa, lvol et lpc) sont plus corrélées avec les composantes principales, ce qui signifie qu'elles contribuent plus à la structure des données que les variables avec des vecteurs plus courts.

Ce graphique nous offre un aperçu de la relation entre les différentes variables du jeu de données et comment elles se comparent et interagissent les unes avec les autres dans le contexte des deux premières composantes principales.

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.

Voici comment nous pouvons interpréter ce graphique :

- **Axes :**

Dim 1 (46.47%) : La première composante principale (Dim 1) explique 46,47% de la variance totale des données.

Dim 2 (22.99%) : La deuxième composante principale (Dim 2) explique 22,99% de la variance totale des données.

Ensemble, ces deux composantes expliquent environ 70% de la variance totale dans votre jeu de données.

- **Vecteurs des variables :**

Les vecteurs (flèches) représentent les variables de notre jeu de données. La direction et la longueur de chaque vecteur indiquent comment chaque variable contribue aux deux premières composantes principales.

Direction : La direction du vecteur indique la corrélation entre la variable et la composante principale. Par exemple, lpsa et lpc pointent dans des directions similaires par rapport à Dim 1, ce qui signifie qu'elles sont positivement corrélées avec cette composante.

Longueur : La longueur du vecteur est un indicateur de la qualité de la représentation de la variable sur le plan factoriel. Plus un vecteur est long, plus la variable est bien représentée par les deux composantes principales.

- **Corrélations entre variables :**

Les variables dont les vecteurs pointent dans la même direction ou dans des directions proches sont positivement corrélées entre elles. Par exemple, lpsa, lpc et lvol semblent être positivement corrélées car leurs vecteurs pointent dans des directions similaires.

Les variables dont les vecteurs pointent dans des directions opposées sont négativement corrélées.

Les variables perpendiculaires (ou orthogonales) sur le biplot sont non corrélées.

En observant le graphique, on peut dire que **lbh, age**, et **lwht** ont une forte contribution à **Dim 2**, tandis que **lpsa, lpc** et **lvol** ont une forte contribution à **Dim 1**.

Summary pour rest.pca (ACP):

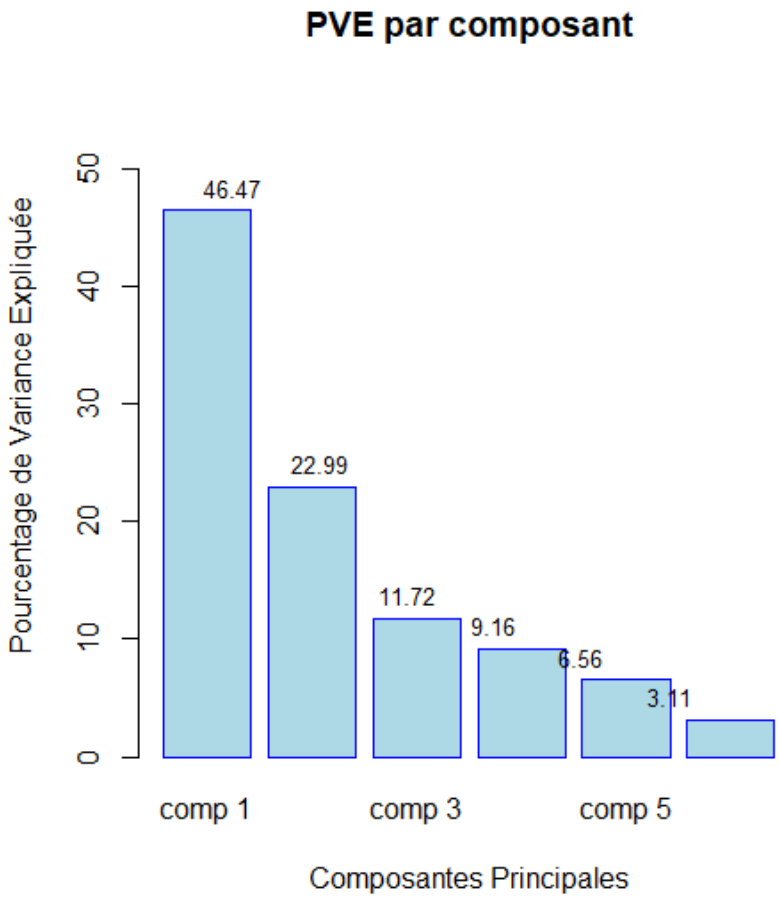
```
> summary(res.pca)

Call:
PCA(X = prostate.dataset, scale.unit = TRUE)


Eigenvalues
                       Dim.1   Dim.2   Dim.3   Dim.4   Dim.5   Dim.6
Variance               2.788   1.379   0.703   0.549   0.393   0.186
% of var.             46.473  22.989  11.720   9.157   6.555   3.106
Cumulative % of var.  46.473  69.462  81.182  90.339  96.894 100.000

Individuals (the 10 first)
         Dist    Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr   cos2
1     |  4.171 | -3.970  7.065  0.906 | -0.831  0.626  0.040 |  0.084  0.012  0.000 |
2     |  3.394 | -3.230  4.677  0.906 |  0.153  0.021  0.002 |  0.306  0.166  0.008 |
3     |  3.960 | -3.008  4.057  0.577 |  0.440  0.175  0.012 |  2.527 11.351  0.407 |
4     |  3.518 | -3.355  5.047  0.910 |  0.174  0.027  0.002 |  0.333  0.197  0.009 |
5     |  2.321 | -1.988  1.772  0.733 | -0.044  0.002  0.000 |  0.543  0.524  0.055 |
6     |  3.419 | -3.237  4.698  0.896 | -0.546  0.270  0.026 | -0.625  0.695  0.033 |
7     |  1.817 | -1.382  0.857  0.579 |  0.877  0.698  0.233 |  0.251  0.112  0.019 |
8     |  2.094 | -1.372  0.844  0.429 |  0.963  0.840  0.211 | -0.608  0.658  0.084 |
9     |  3.287 | -2.821  3.568  0.737 | -0.578  0.303  0.031 | -1.273  2.879  0.150 |
10    |  2.207 | -2.066  1.913  0.876 | -0.137  0.017  0.004 |  0.654  0.759  0.088 |

Variables
        Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr   cos2
age   |  0.485  8.446  0.236 |  0.539 21.036  0.290 |  0.666 62.985  0.443 |
lvol  |  0.858 26.416  0.737 | -0.340  8.405  0.116 |  0.032  0.145  0.001 |
lwht  |  0.661 15.681  0.437 |  0.423 12.996  0.179 | -0.335 16.006  0.113 |
lbh   |  0.404  5.845  0.163 |  0.719 37.467  0.517 | -0.226  7.243  0.051 |
lpc   |  0.698 17.492  0.488 | -0.461 15.376  0.212 |  0.217  6.668  0.047 |
lpsa  |  0.853 26.119  0.728 | -0.255  4.720  0.065 | -0.221  6.952  0.049 |
```

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

Tracez le PVE expliqué par chaque composant

```
> # Calcul du PVE
> pve <- res.pca$eig[, 2]
>
> # Créer un graphique en barres pour le PVE
> barplot(pve,
+         main="PVE par composant",
+         ylab="Pourcentage de Variance Expliquée",
+         xlab="Composantes Principales",
+         ylim=c(0, max(pve) + 10),
+         col="lightblue",
+         border="blue")
>
> # Ajouter le texte au-dessus des barres pour indiquer le pourcentage exact
> text(x=seq_along(pve), y=pve + 2, labels=round(pve, 2), cex=0.8)
```

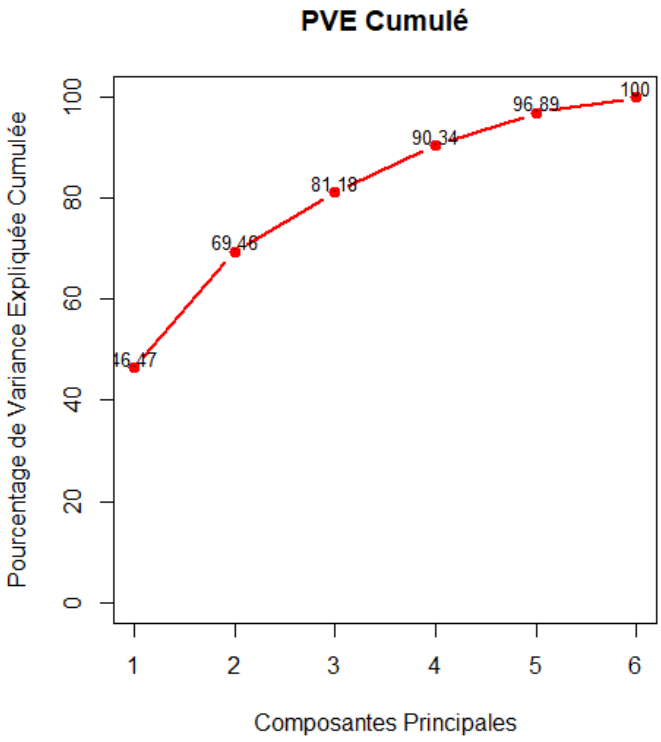
Output :



PVE cumulé :

```
> # Calcul du PVE cumulé
> cumulative_pve <- cumsum(pve)
>
> # Créer un graphique pour le PVE cumulé
> plot(cumulative_pve,
+      type="b",
+      pch=19,
+      col="red",
+      lwd=2,
+      xlab="Composantes Principales",
+      ylab="Pourcentage de Variance Expliquée Cumulée",
+      main="PVE Cumulé",
+      ylim=c(0, 100))
>
> # Ajouter le texte au-dessus des points pour indiquer le pourcentage exact cumulé
> text(x=seq_along(cumulative_pve), y=cumulative_pve + 2, labels=round(cumulative_pve, 2), cex=0.8)
```

Output :



Calcule du pourcentage de variance expliquée (PVE) par chaque composant :

```
> # Calcul du PVE
> pve <- (res.pca$eig[, 1] / sum(res.pca$eig[, 1])) * 100
>
> # Afficher le PVE pour chaque composante
> print(pve)
   comp 1    comp 2    comp 3    comp 4    comp 5    comp 6
46.472817 22.989197 11.719755  9.157486  6.555140  3.105606
```

7. Combien de composants garderiez-vous ? Pourquoi ?

D'après les résultats du PVE de chaque composante et du PVE cumulé, les deux premières composantes expliquent une grande partie de la variance **(69,46%)**. Cela signifie que ces deux composantes capturent une grande partie de l'information contenue dans le jeu de données.

La troisième composante explique 11,72% de la variance, portant le pourcentage cumulatif à 81,18%. Cela signifie qu'avec les trois premières composantes, on peut représenter plus de 80% de la variance totale des données.

Les composantes suivantes (4 à 6) ajoutent de plus en plus peu d'information. En effet, après la quatrième composante, nous avons déjà expliqué 90,34% de la variance. Les cinquième et sixièmes composantes n'ajoutent que 6,56% et 3,11% respectivement, ce qui est relativement faible.

Ainsi, nous pouvons dire que se concentrer sur **les deux** premières composantes pourrait être suffisant car elles capturent la majorité de l'information dans le jeu de données.

III. Régression linéaire

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

La relation entre le coefficient de corrélation r et le coefficient de détermination R2 dans le contexte d'une régression linéaire simple est la suivante :

- R2 = r2
- r est le coefficient de corrélation entre X et Y. Il mesure la force et la direction de la relation linéaire entre X et Y.
  • R2 est le coefficient de détermination. Il représente la proportion de la variance de la variable dépendante Y qui est expliquée par la variable indépendante X dans le modèle.

Concernant la plage de valeurs que r peut prendre :

- r varie entre -1 et 1, inclusivement.
  - r = 1 indique une corrélation linéaire positive parfaite entre X et Y.
  - r = −1 indique une corrélation linéaire négative parfaite entre X et Y.
  - r = 0 indique l'absence de corrélation linéaire entre X et Y.

Ainsi, le R2 variera entre 0 et 1. Si R2 est proche de 1, cela signifie que le modèle de régression linéaire explique une grande proportion de la variance de Y. Si R2 est proche de O, cela signifie que le modèle n'explique pas bien la variance de Y.

2. 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 : $\text{lpsa} = \beta_0 + \beta_1 X + \epsilon$

Calcule de la corrélation entre lpsa et les autres variables :

```
> #Correlation lpsa avec tout les autre variables
> correlations <- cor(prostate.dataset.normalized)[, "lpsa"]
> correlations
      age      lvol      lwht       lbh       lpc      lpsa
0.1755921 0.7858116 0.4558655 0.1927745 0.5545791 1.0000000
```

lvol est la variable la plus corrélé à lpsa donc nous allons la choisir pour notre modelé de régression linéaire simple.

$\hat{\beta}_1$.

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

```
> #estimation des coeficients avec lpsa et lvol
> model <- lm(lpsa ~ lvol, data=prostate.dataset.normalized)
> model_summary <- summary(model)
> model_summary$coefficients
               Estimate Std. Error      t value     Pr(>|t|)
(Intercept) 3.067775e-16 0.06958844 4.408456e-15 1.000000e+00
lvol        7.858116e-01 0.07002749 1.122147e+01 6.005035e-18
```

Les estimations des coefficients sont les suivantes :

Intercept (ou Bo): 3.067775e-16

B1 pour lvol: 7.858116e-01

L'interprétation du coefficient ß1 (associé à `lvol`) est la suivante :

Pour une augmentation d'une unité du logarithme du volume (`lvol`), le logarithme de `lpsa` (taux spécifique d'antigène de la prostate) augmentera en moyenne de 0.7858116 unités, en supposant que toutes les autres variables restent constantes.

La signification de la valeur p (Pr(>lt)) pour le coefficient est extrêmement faible, ce qui suggère que la relation est statistiquement significative.

En conclusion, le modèle est en accord avec les analyses précédentes et montre une relation significative et positive entre `lvol`et `lpsa`.

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

```
> #Test hypothèse de pente nulle pour le coefficient B1
> # Ajuster le modèle de régression linéaire
> model <- lm(lpsa ~ lvol, data=prostate.dataset.normalized)
>
> # Obtenir un résumé du modèle
> model_summary <- summary(model)
>
> # Afficher le récapitulatif des coefficients
> model_summary$coefficients
               Estimate Std. Error      t value     Pr(>|t|)
(Intercept) 3.067775e-16 0.06958844 4.408456e-15 1.000000e+00
lvol        7.858116e-01 0.07002749 1.122147e+01 6.005035e-18
> # Extraire la valeur p pour lvol
> p_value <- coef(model_summary)["lvol", "Pr(>|t|)"]
>
> # Tester l'hypothèse
> if (p_value < 0.05) {
+     cat("Nous rejetons l'hypothèse nulle. Le coefficient pour lvol est significativement différent de 0.\n")
+ } else {
+     cat("Nous ne parvenons pas à rejeter l'hypothèse nulle. Il n'y a pas suffisamment de preuves pour suggérer que le coefficient de lvol est différent de 0.\n")
+ }
Nous rejetons l'hypothèse nulle. Le coefficient pour lvol est significativement différent de 0.
```

La valeur p est inférieure au seuil de significativité de 0.05, ce qui nous conduit à **rejeter l'hypothèse** nulle Ho. Cela signifie que nous avons des preuves statistiquement significatives suggérant que ß1 ≠ 0. En d'autres termes, il y a une relation significative entre lvol et lpsa, et la variable lvol a un effet significatif sur la variable lpsa dans le modèle de régression.

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

Sur la base des résultats obtenus précédemment, le coefficient ß1 pour `lvol` est estimé à environ 0.7858116 avec une valeur p extrêmement faible (6.005035e-18).

La valeur p est bien en dessous du seuil couramment utilisé de 0,05, ce qui signifie qu'il y a une preuve très forte contre l'hypothèse nulle que ß1 est égal à zéro.

Ainsi, la réponse est oui, ß1 est significativement non nul, indiquant une relation significative entre **`lpsa`** et **`lvol`** dans notre modèle de régression.

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 ?

Nous devons dans un premier temps regarder le coefficient de détermination R2 de notre modèle.

```
> model_summary$r.squared
[1] 0.6174998
```

Le coefficient de détermination R2 est de 0.6175, ce qui signifie que 61,75% de la variance de la variable dépendante `lpsa` est expliquée par le modèle de régression basé sur la variable indépendante `lvol`.

Interprétation :

Avec un R2 de 0.6175, le modèle capture une bonne partie mais pas la totalité de la variabilité de la variable dépendante `lpsa` avec la variable `lvol`. Il existe donc une corrélation significative entre ces deux variables.

• Cependant, un R2 de 0.6175 signifie également que près de 38,25% de la variabilité de `lpsa` n'est pas expliquée par `lvol` dans ce modèle. Il pourrait y avoir d'autres variables ou facteurs non considérés dans le modèle qui pourraient expliquer cette variabilité restante (comme lpc).

• Concernant l'aptitude du modèle à prédire le taux d'antigène spécifique de la prostate : bien que le modèle soit statistiquement significatif et capture une partie considérable de la variabilité, il reste encore une proportion importante de variabilité qui n'est pas expliquée. Cela suggère que, bien que le modèle puisse être utilisé pour des prédictions, il est recommandé d'être prudent et de considérer l'inclusion d'autres variables explicatives ou de regarder d'autres modèles pour obtenir une meilleure prédiction

Maintenant regardons le coefficient R2 ajusté :

```
> model_summary$adj.r.squared
[1] 0.612596
```

Le coefficient R2 ajusté est de 0.6126, ce qui indique que, après avoir ajusté pour le nombre de prédicteurs dans le modèle, environ 61,26% de la variance de la variable dépendante lpsa est expliquée par le modèle de régression basé sur la variable indépendante lvol.

L'interprétation :

R2 ajusté est une version modifiée du R2 qui prend en compte le nombre de prédicteurs dans le modèle. Il est principalement utilisé pour éviter le piège où R2 pourrait augmenter artificiellement avec l'ajout de prédicteurs non pertinents. Dans ce cas, étant donné que nous traitons d'une régression simple, la différence entre R2 et R2 ajusté est minime.

Avec un R2 ajusté de 0.6126, cela confirme que le modèle capture une proportion significative de la variabilité de lpsa, mais pas la totalité. Cela suggère que lvol est un prédicteur pertinent pour lpsa, mais il peut toujours y avoir d'autres facteurs ou variables qui peuvent également jouer un rôle dans la variabilité de lpsa.

Le R2 ajusté est légèrement inférieur au R2 non ajusté, la différence est négligeable dans ce contexte, ce qui renforce la pertinence de lvol comme prédicteur de lpsa.

Conclusion

Au terme de ce travail pratique, nous avons pu explorer en profondeur le jeu de données relatif au taux d'antigène spécifique de la prostate. Grâce à une série d'analyses méthodiques, nous avons identifié des relations clés entre les variables, notamment la variable la plus influente (lvol) sur le taux de PSA. L'analyse en composantes principales a été essentielle pour saisir la structure sous-jacente des données, tandis que la régression linéaire nous a permis de quantifier l'impact d'une variable spécifique sur le taux de PSA.

Le coefficient de détermination R2 est de 0.6175 et le R2 ajusté de 0.6126 indiquent que notre modèle de régression linéaire capte une part substantielle, mais pas totale, de la variabilité du taux de PSA. Environ 61,75% de cette variabilité est expliquée par le modèle, ce pourcentage étant légèrement réduit à 61,26% après ajustement pour le nombre de prédicteurs. Cependant, il est toujours essentiel de considérer d'autres variables (comme lpc) ou méthodes d'analyse pour une compréhension plus approfondie.

En somme, ce TP a été une occasion d'appliquer des techniques statistiques avancées vue en cours de Mathematics For Data Sciences sur des données réelles et d'extraire des informations pertinentes qui pourraient guider des recherches futures ou des décisions cliniques liées au PSA. De plus, il nous aura permis de réviser certaines notions importantes en vue du DE qui s’approche à grand pas.