efrei PARIS
Ecole d'ingénieurs du numérique

# MATHEMATICS FOR DATASCIENTIST

RAPPORT

## TRAVAUX PRATIQUES N°1

### M1 APP BDML


MEMBRES DU GROUPE :

AYMEN ZEMMOURI

RÉMY KOUAME

# SOMMAIRE

Introduction ..... 3
I. Analyse des données ..... 4
1. Nombre d'observations ..... 4
2. Statistiques descriptives ..... 4
3. Corrélations « psa » / autres variables ..... 6
4. Nuages de points entre « psa » et les autres variables ..... 7
5. Mise à l'échelle logarithmique ..... 8
II. Analyse en composantes principales (ACP) ..... 8
1. Redondance d'informations / variables fortement corrélées ..... 8
2. Moyennes et variances ..... 9
3. Nécessité de normaliser les variables ..... 9
4. Analyse en Composantes Principales (ACP) ..... 9
5. PVE et PVE Cumulés ..... 12
6. Choix des axes ..... 13
III. Régression linéaire ..... 13
1. Question théorique ..... 13
2. Calcul de la corrélation entre la variable *lpsa* et les autres variables existant dans le jeu de données ..... 13
3. Les estimations des coefficients ..... 14
4. Élaboration du test d'hypothèse ..... 15
5. La valeur du coefficient de détermination R2 ..... 15
Conclusion ..... 16

# Introduction

L'objectif de cette étude est de mieux comprendre les facteurs influençant le taux d'antigène prostatique spécifique. Une étude a été menée sur des hommes atteints d'un cancer de la prostate et ayant subi une prostatectomie radicale, c'est-à-dire une ablation chirurgicale complète de la prostate. Avant la chirurgie, le niveau d'antigène spécifique de la prostate a été déterminé par un test sanguin. Les tissus prélevés lors de l'opération ont été examinés plus précisément afin de caractériser le cancer.

Ce travail pratique (TP) vise à explorer et analyser un jeu de données relatif au cancer de la prostate. Il se décompose en trois étapes essentielles :

- La première étape consiste en une analyse préliminaire des données, comprenant la description du jeu de données, le calcul de statistiques descriptives pour la variable PSA, ainsi que l'analyse de la corrélation entre PSA et les autres variables.
- La deuxième étape s'attache à réaliser une Analyse en Composantes Principales (ACP) afin d'explorer la structure sous-jacente des données.
- Enfin, la troisième étape implique l'utilisation d'une régression linéaire simple pour étudier la relation entre la variable transformée "lpsa" et les autres variables du jeu de données, dans le but de prédire le taux d'antigène spécifique de la prostate.

# I. Analyse des données

1. Nombre d'observations

Il y a 80 observations dans le jeu de données. Le but de cette étude statistique est d'établir si oui ou non il est possible de relier les différentes variables du jeu de données entre elle. Plus précisément, il s'agira de déterminer les liens entre les données afin de pouvoir établir un modèle de régression linéaire qui pourra nous permettre de savoir approximativement si un homme est atteint ou nom du cancer de la prostate.

2. Statistiques descriptives

La sortie de la commande `summary()` fournit un résumé statistique des différentes variables de votre ensemble de données. En fonction de votre description, voici comment interpréter les résultats pour chaque variable :

| 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 |

*Figure 1: Résumé des variables*

- vol (Volume du cancer)
  - La valeur minimale (Min.) est de 0.3.
  - Le premier quartile (1st Qu.) est à 1.65, ce qui signifie que 25% des valeurs sont inférieures ou égales à 1.65.
  - La médiane (Median) est à 3.565, c'est-à-dire que 50% des valeurs sont inférieures ou égales à 3.565.
  - La moyenne (Mean) est d'environ 6.771.
  - Le troisième quartile (3rd Qu.) est à 8.06, indiquant que 75% des valeurs sont inférieures ou égales à 8.06.
  - La valeur maximale (Max.) est de 45.65.

- wht (Poids de la prostate)
  - La valeur minimale (Min.) est de 10.75, ce qui signifie que le poids de la prostate le plus bas observé dans les données est de 10.75.
  - Le premier quartile (1st Qu.) est à 29.20, ce qui indique que 25% des valeurs de poids de la prostate sont inférieures ou égales à 29.20.

- La médiane (Median) est à 38.30, ce qui signifie que 50% des valeurs sont inférieures ou égales à 38.30.
- La moyenne (Mean) est d'environ 41.58, représentant la valeur moyenne du poids de la prostate dans l'échantillon.
- Le troisième quartile (3rd Qu.) est à 48.48, indiquant que 75% des valeurs de poids de la prostate sont inférieures ou égales à 48.48.
- La valeur maximale (Max.) est de 111.95, ce qui représente le poids de la prostate le plus élevé observé.

- age (Âge du patient)
  - La valeur minimale (Min.) est de 41.00, ce qui signifie que l'âge du patient le plus bas observé dans les données est de 41.00 ans.
  - Le premier quartile (1st Qu.) est à 60.00, ce qui indique que 25% des valeurs d'âge du patient sont inférieures ou égales à 60.00 ans.
  - La médiane (Median) est à 65.00, ce qui signifie que 50% des valeurs sont inférieures ou égales à 65.00 ans.
  - La moyenne (Mean) est d'environ 63.61, représentant l'âge moyen des patients dans l'échantillon.
  - Le troisième quartile (3rd Qu.) est à 68.00, indiquant que 75% des valeurs d'âge du patient sont inférieures ou égales à 68.00 ans.
  - La valeur maximale (Max.) est de 79.00, ce qui représente l'âge du patient le plus élevé observé.

- bh (Hyperplasie bénigne)
  - La valeur minimale (Min.) est de 0.250.
  - Le premier quartile (1st Qu.) est à 0.250.
  - La médiane (Median) est à 1.300.
  - La moyenne (Mean) est d'environ 2.692.
  - Le troisième quartile (3rd Qu.) est à 5.075.
  - La valeur maximale (Max.) est de 10.240.

- pc (Propagation dans les vésicules séminales)
    - La valeur minimale (Min.) est de 0.250.
    - Le premier quartile (1st Qu.) est à 0.250.
    - La médiane (Median) est à 1.300.
    - La moyenne (Mean) est d'environ 2.692.
    - Le troisième quartile (3rd Qu.) est à 5.075.
    - La valeur maximale (Max.) est de 10.240.

- psa (Taux spécifique d'antigène de la prostate)
    - La valeur minimale (Min.) est de 0.250.
    - Le premier quartile (1st Qu.) est à 0.250.
    - La médiane (Median) est à 1.300.
    - La moyenne (Mean) est d'environ 2.692.
    - Le troisième quartile (3rd Qu.) est à 5.075.
    - La valeur maximale (Max.) est de 10.240.

La sortie `summary()` fournit des informations sur la distribution des données, notamment les valeurs minimales, les premiers et troisièmes quartiles, la médiane et les valeurs maximales. Elle vous donne une idée de la tendance centrale, de la dispersion et de la forme de la distribution de chaque variable dans votre ensemble de données. Ces statistiques peuvent vous aider à mieux comprendre les caractéristiques de vos données et à prendre des décisions sur la manière de les analyser ou de les traiter.

3. Corrélations « psa » / autres variables

La sortie de la commande permettant de voir la correlation en « psa » et les autres variables est :

```
       vol         wht         age          bh          pc         psa
0.66647230  0.16627567  0.01304884 -0.02203714  0.59571112  1.00000000
```

*Figure 2: Corrélation entre "psa" et les autres variables*

De ce qui précède, on constate que la variable la plus corrélée à « psa » est « vol ».

4. Nuages de points entre « psa » et les autres variables

Les nuages de points obtenus entre « psa » et les autres variables sont présentés ci-après :

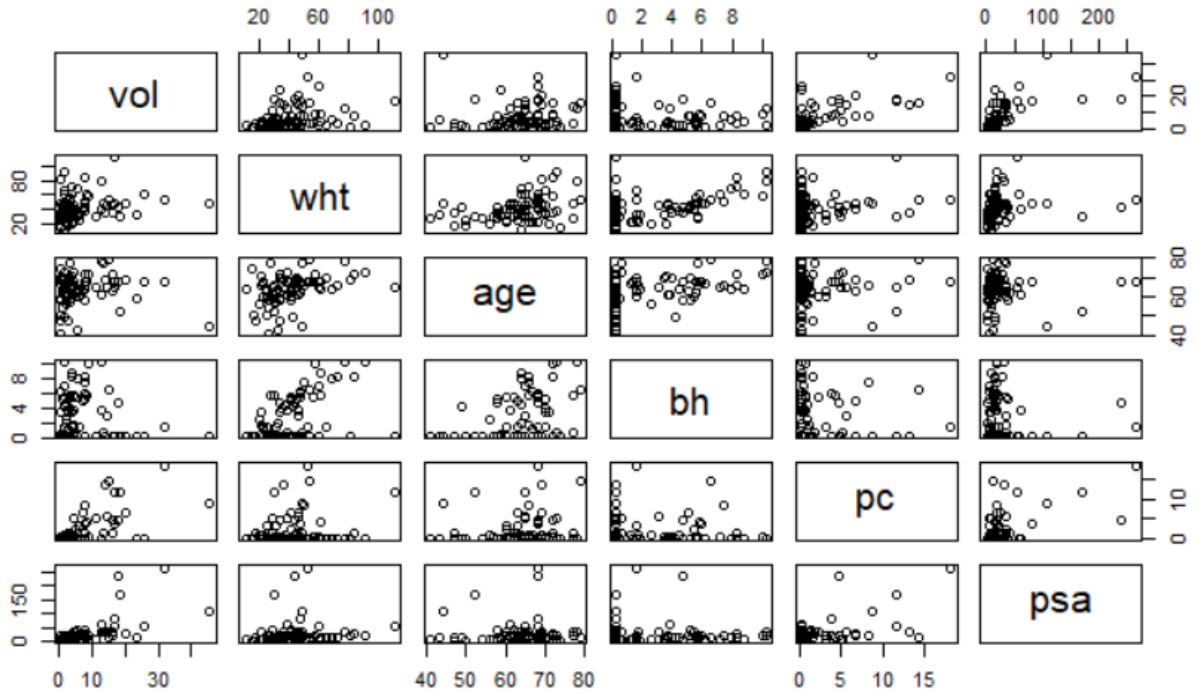


*Figure 3: Corrélation entre les variables d'origines*

On remarque une accumulation de points dans une même zone et cela, sur chaque graphe. Afin d'y remédier, nous pouvons passer à l'échelle logarithmique.

5. Mise à l'échelle logarithmique

Après être passé à l'échelle logarithmique, on observe les nouveaux nuages de points suivants :

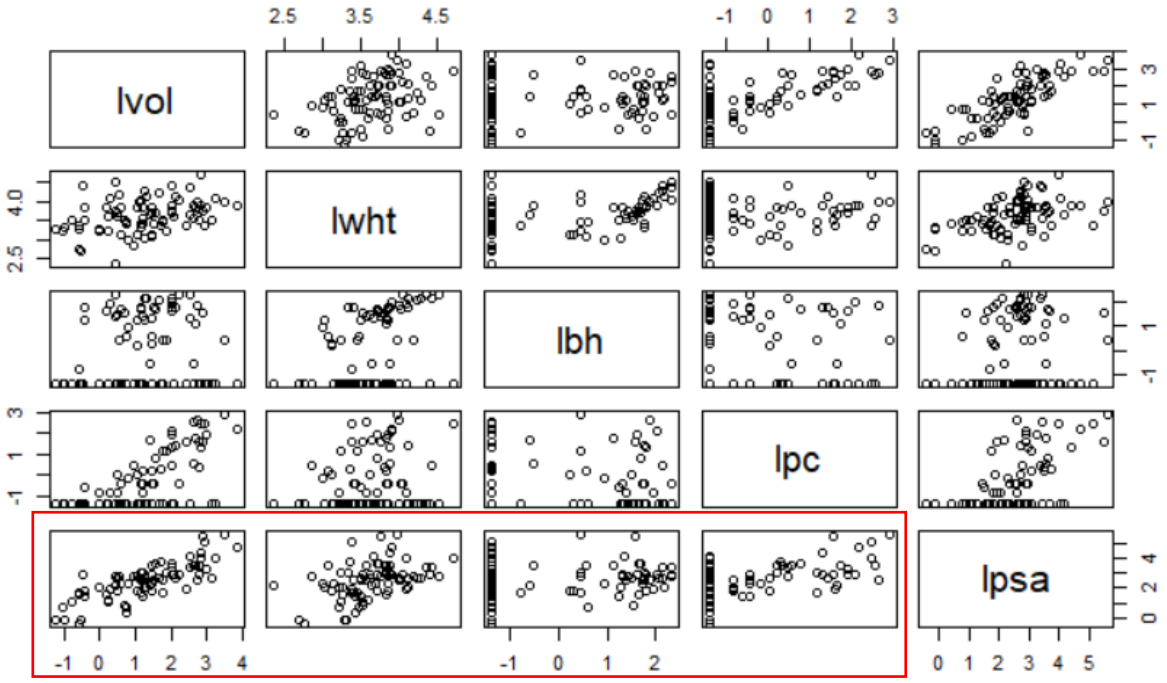


*Figure 4: Corrélations entres variables à l'échelle logarithmique*

# II. Analyse en composantes principales (ACP)

1. Redondance d'informations / variables fortement corrélées

Lorsque deux variables sont parfaitement corrélées (corrélation de 1 ou -1), cela signifie qu'elles contiennent exactement la même information et qu'elles sont redondantes l'une par rapport à l'autre. Dans une ACP, il n'est donc pas nécessaire d'inclure ces deux variables parfaitement corrélées car cela n'apporterait aucune information supplémentaire à l'analyse, bien au contraire, cela pourrait introduire de la **multicollinéarité**, ce qui pourrait être problématique.

2. Moyennes et variances

- Moyennes

```
      lvol       lwht        lbh         lpc        lpsa
1.28845788 3.63755583 0.09693063 -0.26568885 2.50703350
```

*Figure 5: Moyennes*

- Variances

```
    lvol      lwht      lbh       lpc      lpsa
1.4085737 0.1853071 2.1625915 1.8679841 1.4438723
```

*Figure 6: Variances*

- Interprétation

Les valeurs relativement élevées des variances de « lpc » et « lbh » signifient que leurs répartitions autour de leurs moyennes respectives (-0.2656885 et 0.09693063) sont assez irrégulières. C'est-à-dire que tantôt il y a des valeurs très proches de la moyenne, tantôt il y en a qui sont très éloignées de celle-ci. Tandis que la valeur presque nulle de la variance de « lwht » par exemple montre une répartition autour de sa moyenne (3.63755583) assez homogène et uniforme.

3. Nécessité de normaliser les variables

Oui il est nécessaire de normaliser les variables dans notre cas car les résultats précédents (les variances et les moyennes) nous montrent que nos variables ne sont pas centrées et réduites. Cela se déduit par des distributions et une variation anormale entre les valeurs de certaines variables comme « lpc » et « lbh » par exemple. Par conséquent, une normalisation s'impose afin de mener une ACP de qualité et avoir des résultats fiables.

4. Analyse en Composantes Principales (ACP)

Compte tenu du fait que nous avons décidé de normaliser nos données, il faut préciser l'argument `scale = True` lors de l'appel de la fonction `PCA()`. Après l'appel de la fonction `PCA()`, on obtient :

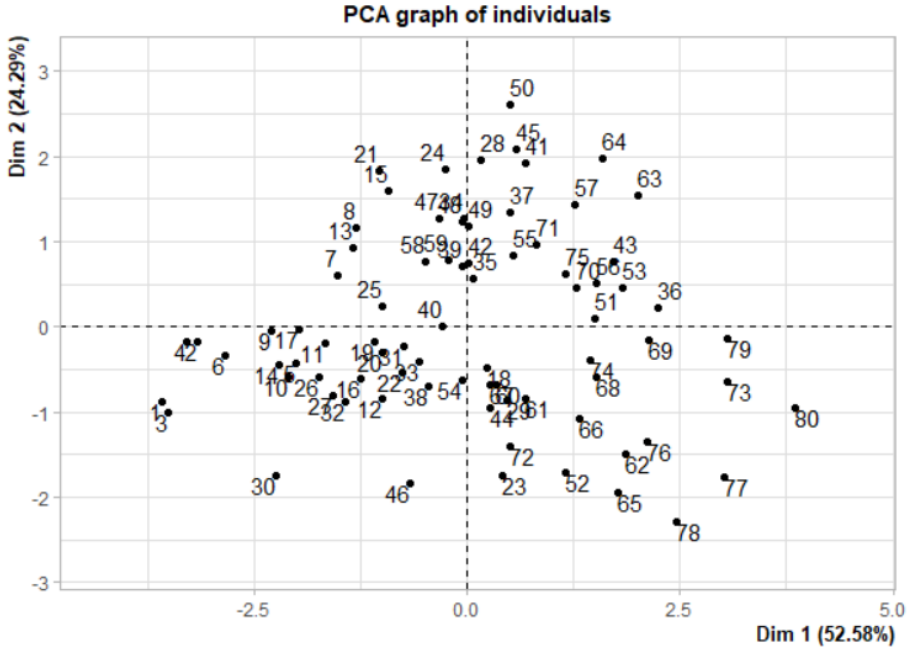


*Figure 7: Nuage d'individus*

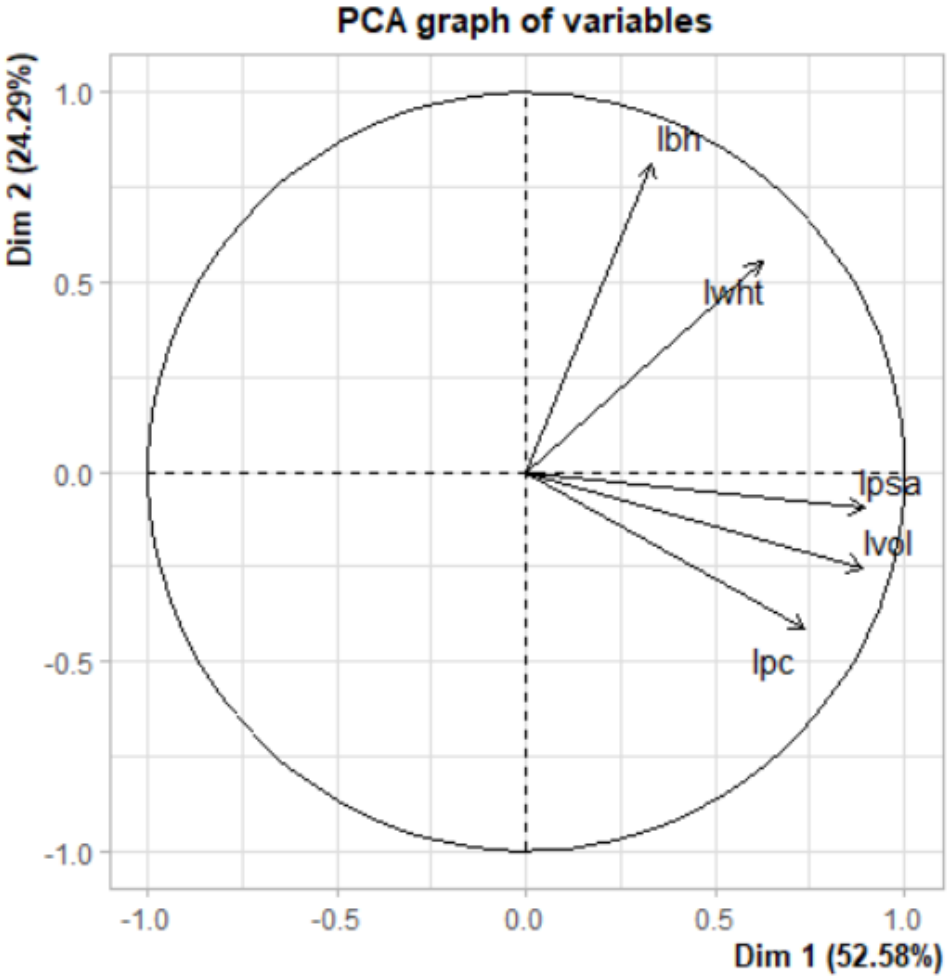


*Figure 8: Cercle des corrélations*

➢ Interprétation

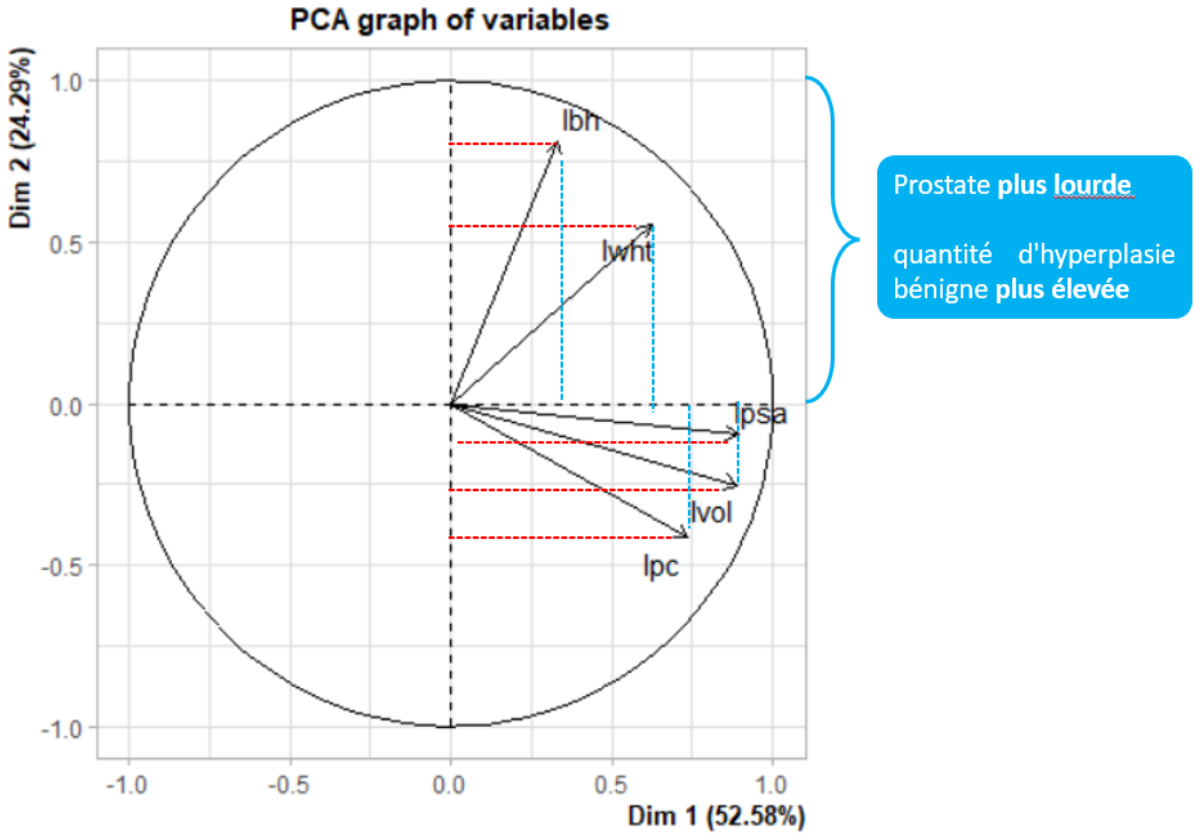


*Figure 9: Projection orthogonale des vecteurs sur les axes*

Le graphe ci-dessus met en évidence le fait que la partie supérieure du cercle regroupe les personnes ayant une prostate assez lourde et ayant une quantité d'hyperplasie bénigne assez élevée. De plus, l'angle de quasi 90° entre « lbh » et « lvol » montre que le volume de la prostate et la quantité d'hyperplasie bénigne ne sont quasiment aucun rapport. Plus explicitement, une personne peut avoir une prostate volumineuse et une quantité d'hyperplasie bénigne faible quand une autre peut avoir le même volume et une quantité d'hyperplasie bénigne élevée.

5. PVE et PVE Cumulés

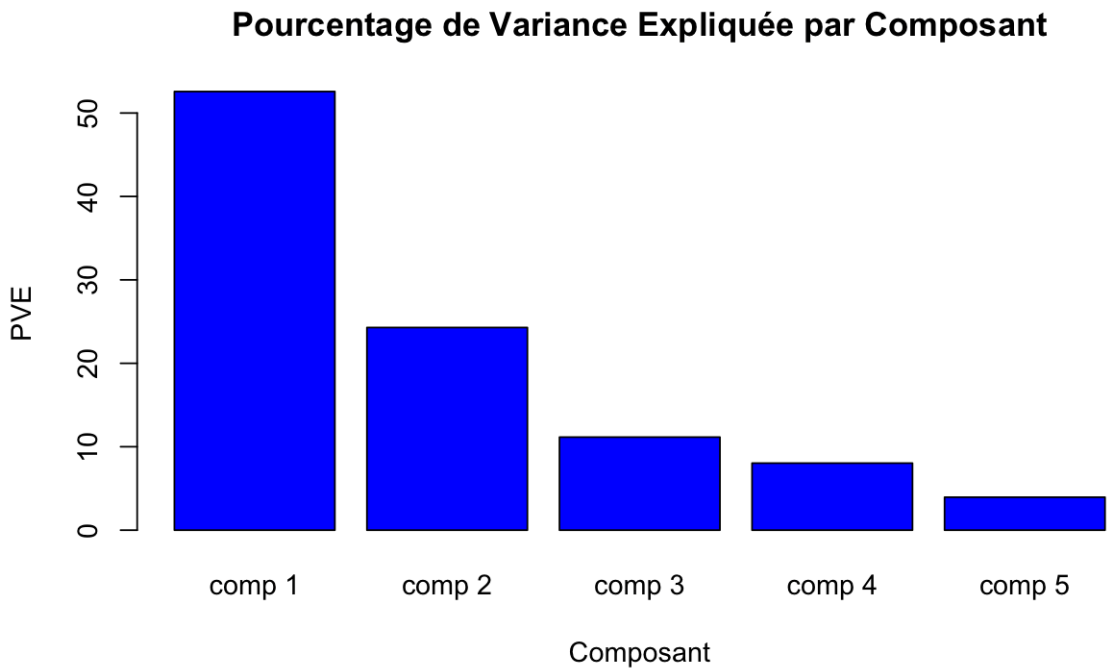


*Figure 10: PVE*

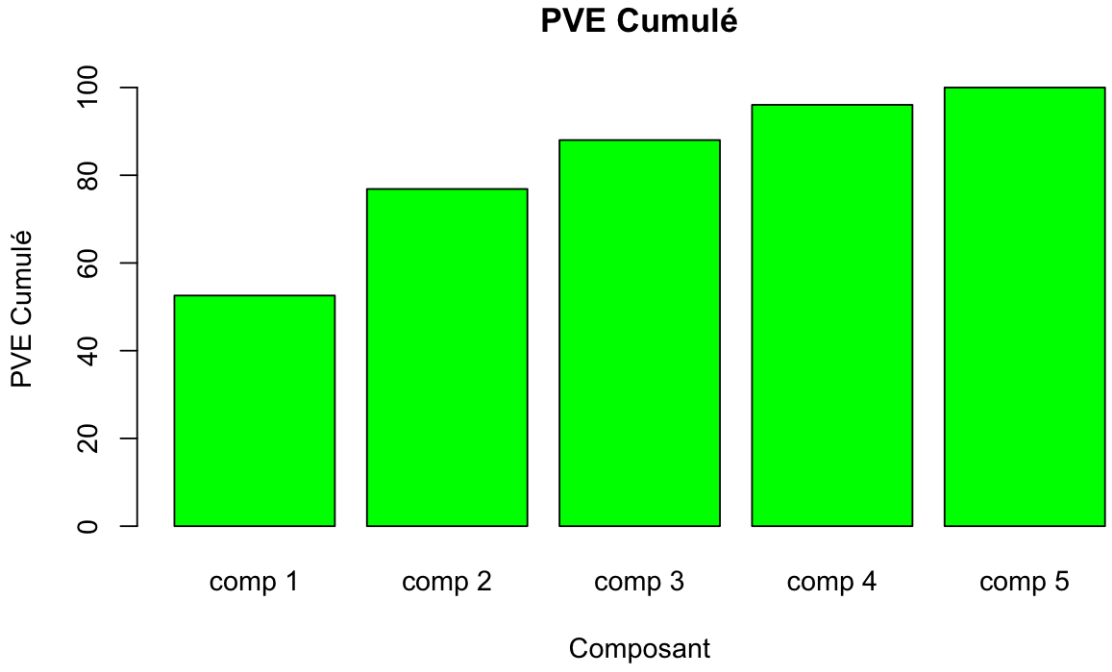


*Figure 11: PVE Cumulée*

6. Choix des axes

De ce qui pécède nous retenons les axes 1 et 2 car à eux seuls, ils représentent 77,87% de l'information.

# III. Régression linéaire

1. Question théorique

La relation entre r et R est la suivante : $R^2 = r^2$.

La corrélation coefficient (r) et le coefficient de détermination ($R^2$) sont deux mesures couramment utilisées pour évaluer la relation entre deux variables, comme dans un modèle de régression linéaire simple.

- La corrélation coefficient (r) mesure la force et la direction de la relation linéaire entre deux variables X et Y.
- Le coefficient de détermination ($R^2$) mesure la proportion de la variance dans la variable dépendante (Y) qui peut être expliquée par la variable indépendante (X) dans le modèle de régression linéaire.

Plus r est proche de -1 ou 1, plus la relation est forte, et plus $R^2$ est proche de 1, plus X explique une grande proportion de la variance de Y.

Quel est la plage de valeurs que peut prendre r ?

Le coefficient de corrélation r peut varier de -1 à 1, où :

- r = 1 : Une corrélation positive parfaite, ce qui signifie que les variables X et Y sont parfaitement corrélées de manière positive.
- r = -1 : Une corrélation négative parfaite, ce qui signifie que les variables X et Y sont parfaitement corrélées de manière négative.
- r = 0 : Aucune corrélation linéaire entre les variables X et Y.

2. Calcul de la corrélation entre la variable *lpsa* et les autres variables existant dans le jeu de données

```
          lvol      lwht       lbh       lpc
[1,] 0.7858116 0.4558655 0.1927745 0.5545791
```

*Figure 12: Corrélation entre les variables*

La variable la plus corrélée avec *lpsa* est : *lvol*
Donc : X = *lvol*

$$lpsa = \beta_0 + \beta_1.\mathrm{X} + \epsilon$$

- Ajuster le modèle de régression

```
Call:
lm(formula = Y ~ X)

Residuals:
     Min       1Q   Median       3Q      Max
-1.70820 -0.45773  0.06161  0.53102  1.83582

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.4819     0.1238   11.97   <2e-16 ***
X             0.7956     0.0709   11.22   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7479 on 78 degrees of freedom
Multiple R-squared:  0.6175,    Adjusted R-squared:  0.6126
F-statistic: 125.9 on 1 and 78 DF,  p-value: < 2.2e-16
```

*Figure 13: résultat de la régression linéaire*

3. Les estimations des coefficients

- $\widehat{\beta_0}$= 1.482
- $\widehat{\beta_1}$= 0.795

Interprétation de l’estimation de $\beta_1$: L'estimation du coefficient pour la variable X est d'environ 0.795. Cela signifie que lorsque la variable "X" augmente d'une unité, la variable dépendante "Y" est estimée pour augmenter en moyenne d'environ 0.795 unités, toutes les autres variables restant constantes. Cette valeur représente la pente de l'équation de régression linéaire.

4. Élaboration du test d'hypothèse

Nous avons les informations suivantes :

- $\widehat{\beta_0}$ = 0.481
- $\widehat{\beta_1}$= 0.795
- $SE(\widehat{\beta_1})$ = 0.071
- *t-value* = 11.22
- *p-value* = 2 x $10^{-16}$
- Degrés de liberté : 78

- Formulation des hypothèses :
  - Hypothèse nulle (H0) : β1=0 (Pas de relation entre X et Y).
  - Hypothèse alternative (H1) : β1≠0 (Il y a une relation entre

- Interprétation des résultats :

Comme la valeur p est bien inférieure à 0,05, nous rejetons l'hypothèse nulle selon laquelle β = 0. Il existe donc une relation significative entre les variables dans le modèle de régression linéaire de l'ensemble de données fidèle

5. La valeur du coefficient de détermination R2

- R2=0.618
- R2 est un indicateur de la proportion de la variance de Y expliquée par X dans le modèle. En d'autres termes, il mesure à quel point les variations dans *lpsa* sont liées aux variations de la variable indépendante *lvol*.
- Dans notre cas, il signifie que le modèle explique environ 61,8 % de la variance dans le taux d'antigène spécifique de la prostate *lpsa*. Cette constatation indique que la variable *lvol* contribue de manière significative à expliquer les variations de *lpsa*.
- Ces résultats suggèrent que la variable *lvol* a une influence significative sur *lpsa* et que le modèle a une capacité de prédire le taux d'antigènes spécifiques.

## Conclusion

En conclusion, ce travail pratique nous a permis d'explorer un jeu de données lié au cancer de la prostate, en mettant en œuvre des analyses statistiques essentielles. Nous avons examiné les corrélations entre les variables, exploré la structure sous-jacente des données par l'ACP, et évalué la relation entre la variable transformée *lpsa* et les autres variables par le biais d'une régression linéaire simple.

Ces analyses nous ont aidés à mieux comprendre les facteurs influençant le taux d'antigène prostatique spécifique.

En fin de compte, ce TP a illustré comment les concepts et les techniques statistiques peuvent être appliqués de manière concrète pour explorer et interpréter des données réelles, renforçant ainsi notre compréhension de cette maladie potentiellement grave.