Analyse de données :
Le cancer de la prostate

Mathematics for Data Science

Evan ARNEAU **[20220313]**
Mariam MAKSIMOUS **[20190471]**

Table des matières

1 – Introduction

Pour établir ce rapport nous disposons d'un jeu de données déterminé par 7 variables récupérées sur 80 patients différents tous atteints d'un cancer de la prostate. Ces 7 variables sont : le volume du cancer (vol), le poids de la prostate (wht), l'âge du patient (age), la quantité d'une hyperplasie bénigne (bh, tumeur bénigne), la propagation dans les vésicules séminales (vs), la pénétration capsulaire (pc, pénétration du cancer dans la capsule entourant la prostate), et le taux spécifique d'antigène produit par la prostate (psa). La prostate sécrète initialement un antigène spécifique, une protéine. Lorsqu'on observe un patient atteint d'un cancer de la prostate, nous remarquons que cette protéine est sécrétée jusqu'à dix fois plus que la normale. Cependant, le taux spécifique de l'antigène peut aussi être affecté par d'autres facteurs tels que le volume de la prostate, les inflammations, etc. L'objectif de notre étude est donc de comprendre s'il existe une corrélation entre les diverses données dont nous disposons et le taux d'antigène de la prostate d'un patient. Pour cela nous allons calculer et dresser différents modèles statistiques, et les interpréter au mieux afin de déterminer si oui ou non on peut estimer que l'augmentation du taux d'antigène dans la prostate implique la présence d'un cancer. Une réponse positive permettrait ainsi au corps médical de diagnostiquer plus rapidement un cancer, et ainsi stopper sa propagation dès ses premiers stades.

2 – Première analyse : description du jeu de données

2.1 - Calcul de statistiques descriptives pour la variable PSA

En calculant quelques statistiques descriptives concernant le taux d'antigène spécifique, on remarque rapidement que la plage de données semble large (min : 0.650, max : 265.850), la moyenne, la médiane ainsi que les quartiles restent tout de même très faibles par rapport au maximum. On en déduit donc que malgré qu'une moyenne s'établit autour de 25.473, certains patients possèdent des taux anormalement élevés d'antigène.

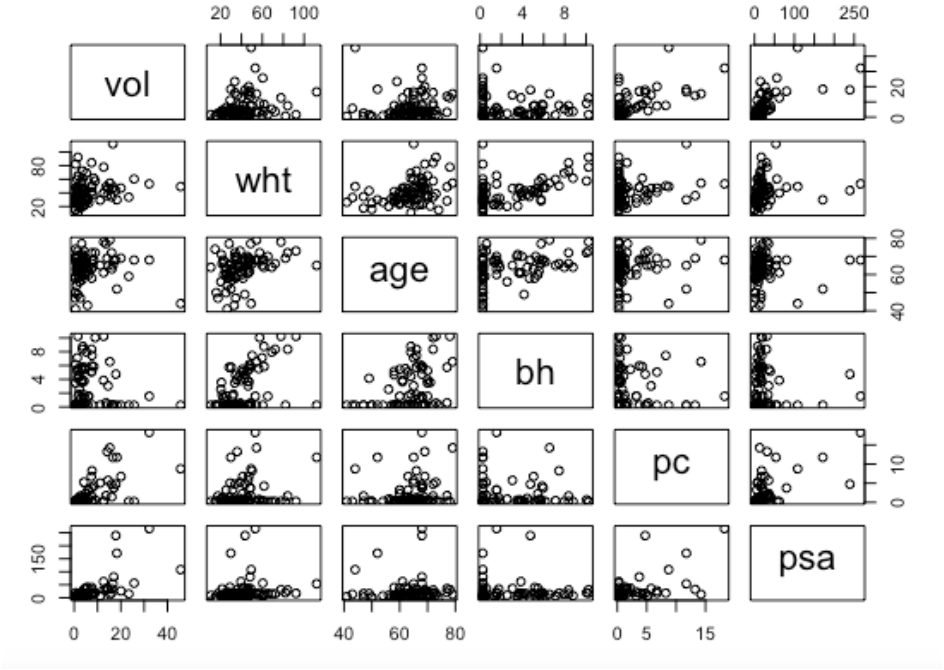
```
      psa
Min.   :  0.650
1st Qu.:  6.125
Median : 14.400
Mean   : 25.473
3rd Qu.: 21.350
Max.   :265.850
```

2.2 - Analyse de la variable PSA : quelle est sa corrélation avec les autres variables ?

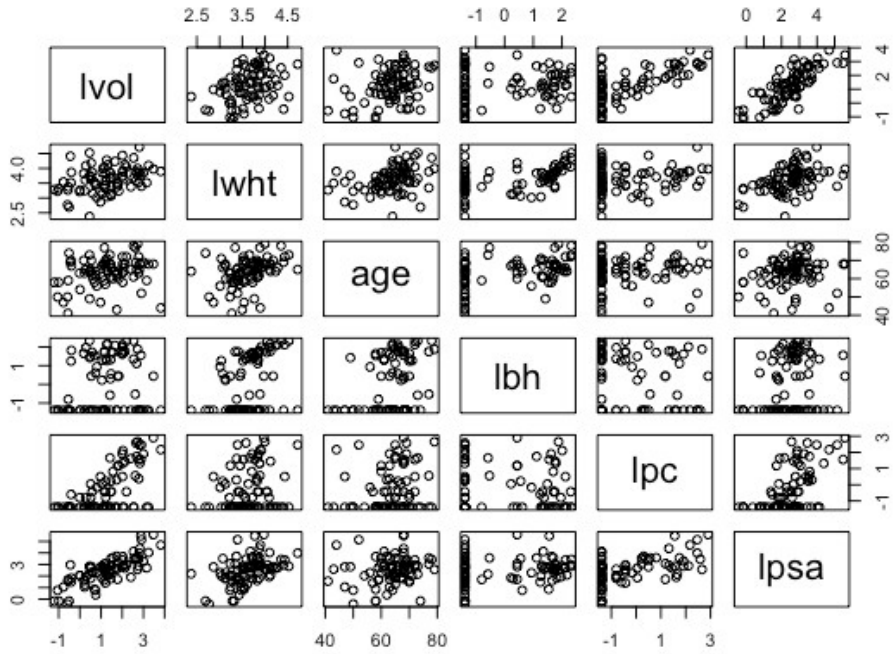
```
> cor(prostate$vol,prostate$psa,use="complete.obs")
[1] 0.6664723
> cor(prostate$wht,prostate$psa,use="complete.obs")
[1] 0.1662757
> cor(prostate$age,prostate$psa,use="complete.obs")
[1] 0.01304884
> cor(prostate$bh,prostate$psa,use="complete.obs")
[1] -0.02203714
> cor(prostate$pc,prostate$psa,use="complete.obs")
[1] 0.5957111
```

On sait que plus le coefficient de corrélation s'approche de 1, plus les variables étudiées sont corrélées positivement. On peut donc en déduire ici que le volume de la prostate et la pénétration capsulaire sont les variables les plus corrélées (positivement) au taux d'antigène dans la prostate.

2.3 - Clarifier le modèle visuel : transformation logarithmique



plot(prostate) avant transformation logarithmique



plot(prostate) après transformation logarithmique

Comme on peut le voir sur notre graphe après transformation logarithmique, il semble il y avoir une corrélation positive entre le taux d'antigène spécifique sécrété par la prostate et le volume de la prostate ainsi que la pénétration capsulaire. En effet, plus le taux d'antigène spécifique sécrété est important, plus ces deux variables augmentent. En revanche, il ne semble y avoir aucune corrélation entre le taux d'antigène spécifique et l'âge, le poids de la prostate, et la présence d'une tumeur bénigne.

3 - Analyse en composantes principales (ACP)

3.1 - La corrélation parfaite entre deux variables : présente-t-elle un intérêt dans le cas d'une ACP ?

Si on retrouve deux variables parfaitement corrélées dans notre jeu de données, il nous semble inapproprié d'inclure les deux dans notre analyse lors de la réalisation de l'ACP, vu qu'elles représentent exactement la même chose. L'inclusion de l'une retournera le même résultat que si on inclut l'autre ou les deux en même temps.

3.2 - Calcul de variance des variables

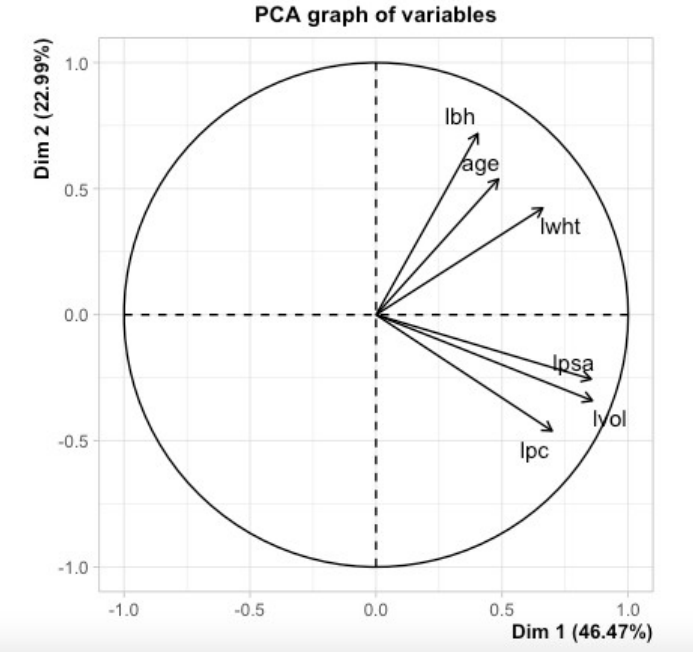
```
> apply(prostate,2,var)
      lvol       lwht        age        lbh        lpc       lpsa
1.4085737  0.1853071 62.3669304  2.1625915  1.8679841  1.4438723
```

On remarque que sur les données logarithmiques, la plupart de nos variables ont une variance relativement proche de 0. La variance étant la valeur représentant le degré de dispersion de nos valeurs autour de la moyenne de la variable, on comprend donc que nos variables comportent des valeurs très concentrées autour de leur moyenne. En revanche, la variable âge présente une variance de 62.366, signifiant qu'elle comporte des données plus dispersées.

3.3 - Normalisation de variables dans le cadre d'une ACP

Dans notre cas, il est nécessaire de normaliser les variables de notre jeu de données puisqu'elles sont toutes d'unités différentes, cela perturberait notre analyse. En faisant cela, on place nos variables à la même échelle.

3.4 - Réalisation de l’ACP

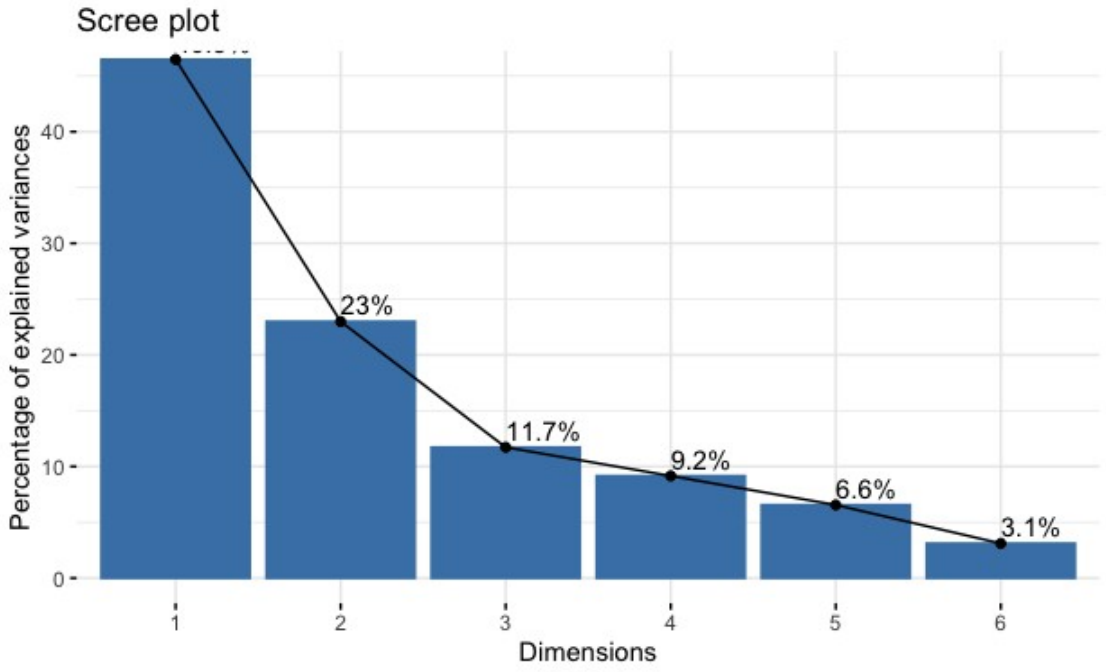


Ci-contre se trouve le schéma tracé de l’ACP. On voit que les données qui semblent être corrélées entre elles sont (par lot) :

- la quantité d’hyperplasie bénigne, l’âge et le poids de la prostate
- le volume de la prostate, la pénétration capsulaire et le taux spécifique d’antigène de la prostate

On constate que ces deux lots ne sont que très peu corrélés entre eux puisque les flèches tracées forment entre elles des angles droits (signe d’un rapport non existant ou très faible entre deux variables lors de la réalisation d’un ACP). Mis à part possiblement le poids de la prostate avec le taux d’antigène de la prostate.

3.5 - Tracer le PVE, et l’interpréter



PVE tracé sur R

Suite à cette analyse nous allons garder seulement les trois premiers composants car ils retiennent plus de 80 % de l’information importante.

4 – Régression linéaire

4.1 - La corrélation entre lpsa et les autres variables

```
> new_prostate <- data.frame(prostate$lpsa,prostate$lvol,prostate$lwht,prostate$lbh,prostate$lpc,pro
state$age)
> cor_lpsa <- cor(as.matrix(new_prostate[,1]),as.matrix(new_prostate[,-1]))
> cor_lpsa
     prostate.lvol prostate.lwht prostate.lbh prostate.lpc prostate.age
[1,]     0.7858116     0.4558655    0.1927745    0.5545791    0.1755921
```

Après avoir calculé les corrélations de nos variables avec la variable prostate, on remarque que la corrélation la plus forte est celle entre lvol et lpsa. Pour la suite de notre analyse nous utiliserons donc X = lvol, la variable la plus corrélée avec lpsa.

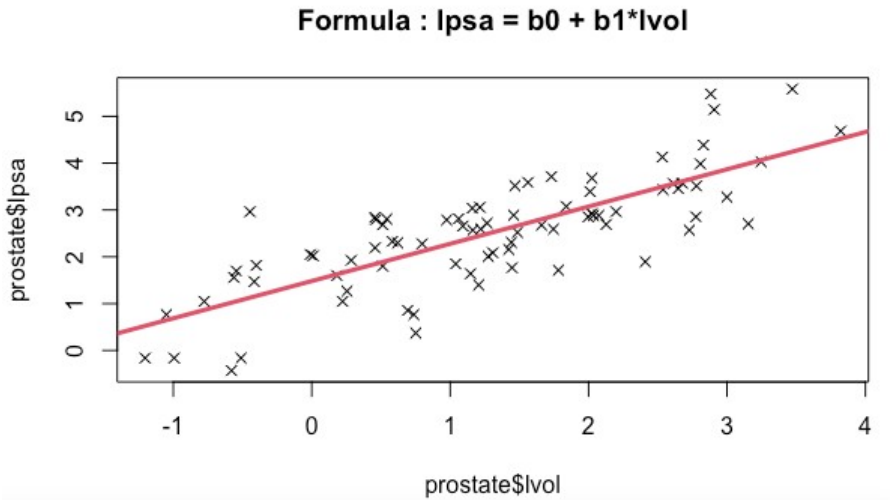
4.2 - Interpréter l'estimation des coefficients

```
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 ***
prostate$lvol   0.7956     0.0709   11.22   <2e-16 ***
```

En exécutant la commande summary(model) on peut retourner nos coefficients B0 = 1.481 et B1 = 0.795. Le coefficient B1 représente l'augmentation moyenne du taux spécifique d'antigène sécrété par la prostate par unité de volume de la prostate (lpsa/lvol). Sa valeur étant inférieure à 1, on estime le coefficient trop faible : ce n'est pas un coefficient de détermination suffisamment élevé pour dresser un modèle fiable quant à la prédiction du taux d'antigène spécifique de la prostate.

4.3 - Test d’hypothèse de pente nulle



Après avoir réalisé notre graphe, on remarque que les points sont peu dispersés de la ligne tracée. On en déduit donc qu’il semble il y avoir une relation positive entre lpsa et lvol, plus lpsa augmente, plus lvol augmente.

4.4 - Le coefficient de détermination

```
> coeff_determination <- summary(model)$r.squared
> coeff_determination
[1] 0.6174998
```

Après calcul, on détermine que notre coefficient de détermination vaut 0.617. Nous estimons que cette valeur n’est pas assez élevée, le coefficient de détermination n’est donc pas fiable.

5 – Conclusion

Pour conclure, après avoir déterminé que notre coefficient de détermination n’est pas fiable, nous estimons que notre modèle n’est donc pas adapté pour prédire le taux d’antigène spécifique de la prostate. Les données de notre data frame ne présentant une corrélation pas assez importante avec notre variable cible, nous ne pouvons donner suite à notre étude.