Projet STA 203

Analyse multivariée de la consommation automobile

Réalisé par :

Mahdi HADJTAIEB

Encadré par :

Mme Anissa Rabhi

2eme année Techniques Avancées

Année universitaire : 2024/2025

Introduction

L'objectif de ce projet est d'étudier la consommation de carburant de différents véhicules à partir de plusieurs caractéristiques techniques, en utilisant des outils statistiques classiques et multivariés. Plus précisément, il s'agit de modéliser la variable `mpg` (miles per gallon) à l'aide de différentes variables explicatives, puis de compléter cette analyse par une réduction de dimension via l'Analyse en Composantes Principales (ACP).

Ce travail s'inscrit dans une démarche d'analyse complète d'un jeu de données multivarié, avec une attention portée à la visualisation, à la sélection de variables, à la modélisation, et à l'interprétation des résultats.

Pour mener à bien cette étude, nous utiliserons le langage R, outil adapté pour le traitement statistique et la visualisation graphique. L'analyse se déroule en plusieurs étapes :

1. Une étude statistique descriptive des variables permettra de dégager les tendances générales du jeu de données, à l'aide de représentations graphiques telles que les boxplots et les histogrammes.

2. Un modèle de régression multiple initial sera construit afin d'expliquer la consommation à partir de toutes les variables disponibles.

3. Une sélection de variables sera effectuée manuellement, en s'appuyant sur les corrélations et les résultats du modèle initial, afin d'obtenir un modèle plus simple et plus pertinent. Une comparaison des modèles sera réalisée.

4. L'impact de l'ajout d'un individu extrême au jeu de données sera analysé pour tester la robustesse du modèle.

5. Une Analyse en Composantes Principales (ACP) sera ensuite menée afin de réduire la dimension du jeu de données et de visualiser les relations entre les variables.

6. L'ACP permettra également d'examiner de potentiels regroupements d'individus ou de variables, afin d'identifier des profils similaires parmi les observations.

7. Une régression sur les composantes principales sera construite et comparée au modèle classique.

8. Enfin, les avantages et les limites des deux approches (régression classique et régression sur ACP) seront discutés, afin de mettre en perspective les choix méthodologiques et leur impact sur l'interprétation et la modélisation.

Toutes les analyses ont été réalisées sous R, et les codes sont intégrés directement dans le corps du document grâce à l'utilisation de RMarkdown, ce qui permet d'assurer à la fois la traçabilité, la reproductibilité et une lecture fluide des résultats.

Acquérir les données

```
rm(list = objects())
graphics.off()
data <- read.table("C:/Users/Utilisateur/Desktop/projet stat
2025/mtcars.csv", header = TRUE, sep = ",", dec = ".")
head(data)
```

```
##                model  mpg cyl disp  hp drat    wt  qsec vs am gear carb
## 1          Mazda RX4 21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## 2      Mazda RX4 Wag 21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## 3         Datsun 710 22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## 4     Hornet 4 Drive 21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## 5  Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## 6            Valiant 18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

Le jeu de données `mtcars` regroupe les caractéristiques techniques de 32 modèles de voitures. Il contient plusieurs variables explicatives, principalement numériques, décrivant la performance, la consommation et la structure des véhicules.

Voici une brève description de chaque variable :

- `mpg` (Miles per gallon) : Consommation de carburant en miles parcourus par gallon d'essence (variable à expliquer dans notre étude).
- `cyl` (Cylinders) : Nombre de cylindres du moteur.
- `disp` (Displacement) : Cylindrée du moteur, en pouces cubes (mesure du volume des cylindres).
- `hp` (Horsepower) : Puissance du moteur en chevaux.
- `drat` (Rear axle ratio) : Rapport de démultiplication du pont arrière.
- `wt` (Weight) : Poids du véhicule (en milliers de livres).
- `qsec` (1/4 mile time) : Temps nécessaire pour parcourir un quart de mile (accélération).
- `vs` (V/S) : Type de moteur (0 = en V, 1 = en ligne).
- `am` (Transmission) : Type de transmission (0 = automatique, 1 = manuelle).
- `gear` (Number of gears) : Nombre de vitesses de la boîte de transmission.
- `carb` (Number of carburetors) : Nombre de carburateurs.

La plupart des variables du jeu de données mtcars sont quantitatives. Les variables comme `mpg`, `disp`, `hp`, `drat`, `wt` et `qsec` sont quantitatives continues, tandis que `cyl`, `gear` et `carb` sont quantitatives discrètes. En revanche, `vs` (type de moteur) et `am` (type de transmission) sont des variables qualitatives binaires, codées sous forme numérique (0 ou 1).

```r
summary(data)
```

```
##     model                mpg             cyl              disp
##  Length:32          Min.   :10.40   Min.   :4.000   Min.   : 71.1
##  Class :character   1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8
##  Mode  :character   Median :19.20   Median :6.000   Median :196.3
##                     Mean   :20.09   Mean   :6.188   Mean   :230.7
##                     3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0
##                     Max.   :33.90   Max.   :8.000   Max.   :472.0
##        hp             drat             wt             qsec
##  Min.   : 52.0   Min.   :2.760   Min.   :1.513   Min.   :14.50
##  1st Qu.: 96.5   1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89
##  Median :123.0   Median :3.695   Median :3.325   Median :17.71
##  Mean   :146.7   Mean   :3.597   Mean   :3.217   Mean   :17.85
##  3rd Qu.:180.0   3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90
##  Max.   :335.0   Max.   :4.930   Max.   :5.424   Max.   :22.90
##        vs               am              gear            carb
##  Min.   :0.0000   Min.   :0.0000   Min.   :3.000   Min.   :1.000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000
##  Median :0.0000   Median :0.0000   Median :4.000   Median :2.000
##  Mean   :0.4375   Mean   :0.4062   Mean   :3.688   Mean   :2.812
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000
##  Max.   :1.0000   Max.   :1.0000   Max.   :5.000   Max.   :8.000
```

1) Statistiques descriptives

```r
# Créer df en excluant la variable 'model'
df <- data[, -1]
# Aperçu des premières lignes
head(df)
```

```
##    mpg cyl disp  hp drat    wt  qsec vs am gear carb
## 1 21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## 2 21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## 3 22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## 4 21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## 5 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## 6 18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

```r
# Structure (types de variables)
str(df)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : int  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : int  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
```

```
##  $ vs  : int  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : int  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: int  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: int  4 4 1 1 2 1 4 2 2 4 ...
```

La fonction `str` est utilisée pour afficher la structure des données pour chaque colonne. Elle affiche le type de données, le nombre d'observations et de variables, et les noms des variables.

```
dim(df)
```

```
## [1] 32 11
```

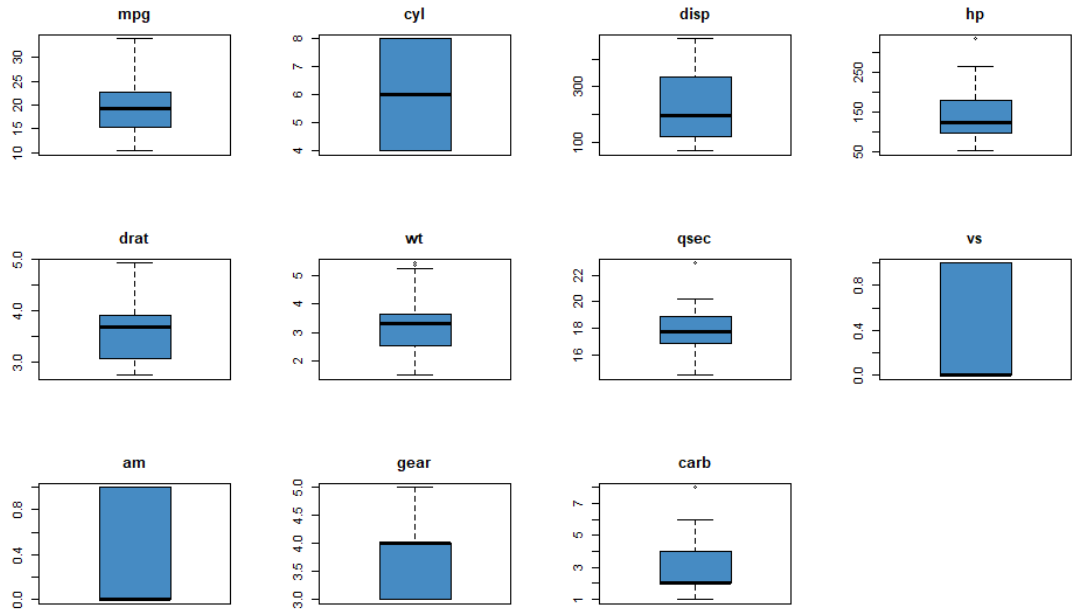
La fonction `dim` est utilisée pour afficher le nombre d'observations et de variables dans le data frame.

1.1) Analyse par boxplots

Les boxplots, ou boîtes à moustaches, sont des représentations graphiques simples et efficaces permettant de résumer la distribution d'une variable quantitative. Ils mettent en évidence la médiane, les quartiles (Q1 et Q3), et les valeurs potentiellement extrêmes. La boîte représente l'intervalle interquartile (50 % des observations), tandis que les "moustaches" s'étendent généralement jusqu'à 1.5 fois cet intervalle. Les observations situées en dehors sont signalées comme des points (lorsqu'elles existent).

Ce type de graphique est particulièrement utile pour repérer des asymétries ou des écarts notables dans les données.

```
par(mfrow = c(3, 4), mar = c(4, 4, 3, 2))
for (i in 1:ncol(df)) {
  var <- df[[i]]
  boxplot(var, main = colnames(df)[i], col = "#468bc3", axes = FALSE,
          ylim = c(min(var), max(var)))
  axis(2, at = pretty(range(var)))
  box()
}
```



L’analyse des boxplots met en évidence plusieurs éléments importants sur la répartition et la variabilité des variables :

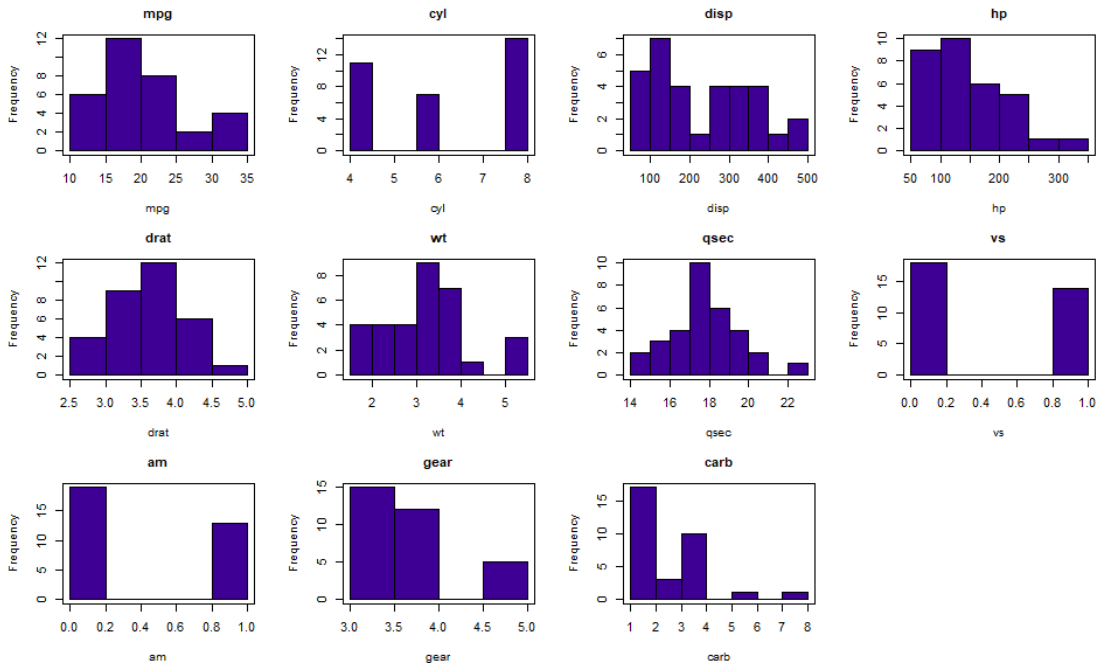
- `mpg` (consommation) est modérément étalée, avec une médiane autour de 19–20 mpg. On ne note pas de valeur fortement atypique.
- `disp` (cylindrée) et `hp` (`puissance`) présentent une forte dispersion. Pour hp, on observe un point isolé, représentant une voitures très puissante, donc potentiellement atypique.
- `wt` (poids) est assez bien répartie autour de la médiane (~3.2), mais affiche deux points supérieurs isolés, signalant deux véhicules sensiblement plus lourds que les autres.
- `qsec` (temps d’accélération) est relativement symétrique, mais présente aussi un point légèrement isolé, indiquant un modèle avec une accélération un peu plus rapide que la norme.
- `carb` (nombre de carburateurs) montre également une ou deux valeurs isolées, reflétant des véhicules dotés de nombreux carburateurs (jusqu’à 8), ce qui est peu courant dans l’échantillon.
- `drat` (rapport de démultiplication du pont arrière) présente une dispersion modérée, avec une répartition relativement homogène autour de sa médiane. Aucune valeur fortement isolée n’est observée.
- Les variables `gear`, `cyl`, et `vs`, `am` sont soit discrètes ou binaires, donc leurs boxplots sont plus compacts.

1.2) Analyse par histogrammes

Les histogrammes permettent de visualiser la répartition d'une variable quantitative en regroupant les valeurs dans des classes (ou intervalles). Chaque barre indique le nombre d'observations appartenant à une classe donnée, ce qui permet de repérer facilement la forme générale de la distribution (symétrie, étalement, concentration, présence de pics ou de creux).

C'est un outil essentiel pour identifier les tendances globales et d'éventuelles déformations (asymétrie, biais) dans les données.

```r
par(mfrow = c(3, 4), mar = c(4, 4, 3, 2))
for (i in 1:ncol(df)) {
  var <- df[[i]]
  hist(var,
       main = colnames(df)[i],
       xlab = colnames(df)[i],
       col = "#3e0094",
       axes = FALSE,
       breaks = "Sturges",
       cex.main = 1, cex.lab = 0.9)
  axis(1, at = pretty(range(var)))
  axis(2)
  box()
}
```



Les histogrammes permettent de visualiser la forme de distribution de chaque variable. Ils confirment certaines observations des boxplots et apportent des compléments utiles :

- `mpg` (consommation) est légèrement asymétrique à gauche, avec un pic de densité entre 15 et 20 mpg. La distribution est globalement étalée, mais sans concentration marquée autour d'une moyenne.
- `disp` (cylindrée) présente une petite asymétrie à gauche.
- `hp` (puissance) présente un asymétrie marquée à gauche, ce qui signifie que la majorité des véhicules ont une puissance faible, mais quelques-uns ont des valeurs très élevées.
- `wt` (poids du véhicule) suit une distribution plutôt symétrique autour d'une valeur centrale (environ 3.2), ce qui se rapproche d'une distribution normale.
- `qsec` (temps d'accélération) présente une distribution bien centrée et approximativement normale, avec un pic autour de 17–18 secondes.
- `drat` est relativement symétrique mais légèrement étalée, sans réel pic marqué.
- Les variables discrètes comme `gear`, `cyl` et `carb` présentent des distributions en barres distinctes, correspondant aux différentes modalités possibles. Certaines d'entre elles (comme carb) montrent une concentration sur quelques valeurs (1, 2 ou 4).
- Enfin, `vs` et `am`, qui sont des variables binaires, affichent logiquement deux barres correspondant aux modalités 0 et 1.

1.3) Corrélations entre les variables

Avant de construire un modèle de régression ou d'effectuer une ACP, il est important d'analyser les relations linéaires entre les variables. Pour cela, nous utilisons la matrice des corrélations de Pearson, qui permet de mesurer l'intensité et le sens du lien entre deux variables quantitatives.

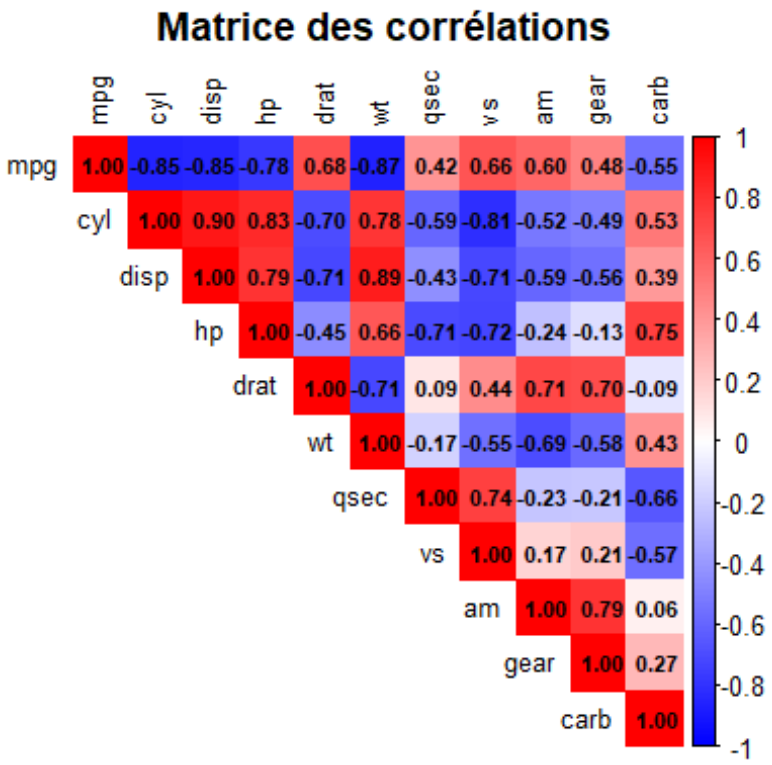
Une valeur de corrélation proche de 1 ou -1 indique une forte relation (positive ou négative), tandis qu'une valeur proche de 0 traduit une indépendance linéaire. Cette étape est essentielle pour détecter :

- d'éventuelles redondances entre variables (fortement corrélées),
- ou au contraire des variables indépendantes, ce qui orientera la construction du modèle et justifiera éventuellement une réduction de dimension par ACP.

Matrice de corrélation simple (numérique)

```
# install.packages("corrplot")
library(corrplot)

## corrplot 0.95 loaded

# Calcul de la matrice de corrélation
cor_mat <- cor(df)
# Affichage de la matrice des corrélations
```

```r
corrplot(cor_mat,
         method = "color",
         type = "upper",
         tl.col = "black",
         addCoef.col = "black",
         col = colorRampPalette(c("blue", "white", "red"))(200),
         number.cex = 0.7,
         tl.cex = 0.8,
         title = "Matrice des corrélations", mar = c(0,0,2,0))
```



L'analyse de la matrice des corrélations met en évidence plusieurs liens linéaires intéressants :

- Une corrélation négative marquée entre mpg (consommation) et des variables comme wt (poids), hp (puissance), disp et cyl, ce qui est intuitif : des véhicules plus lourds ou puissants consomment davantage.
- cyl, disp, hp et wt sont entre elles fortement corrélées positivement, ce qui laisse penser qu'elles mesurent des aspects liés à la taille ou la puissance des véhicules.
- D'autres variables comme drat ou qsec sont moins corrélées avec les autres.

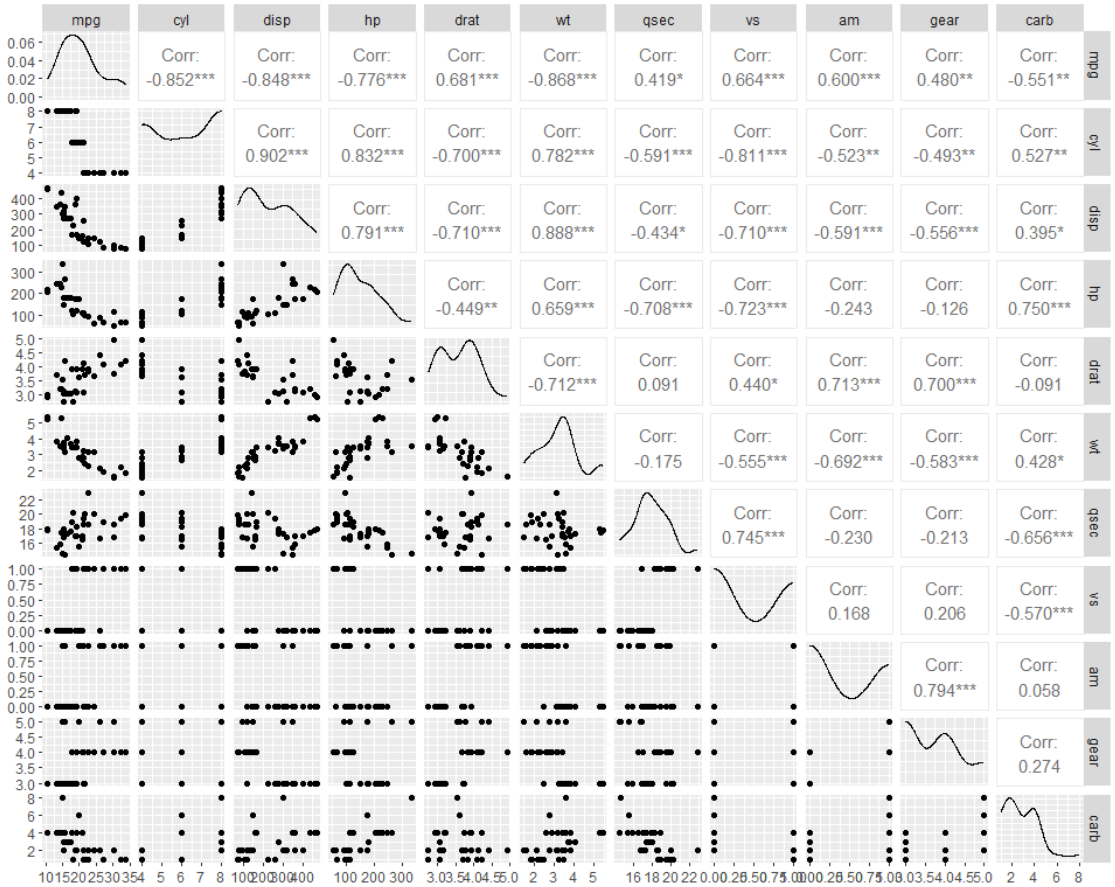
Ces observations renforcent l'intérêt d'une réduction de dimension via une ACP, car certaines variables semblent redondantes, tandis que d'autres apportent de l'information indépendante.

En complément de la matrice de corrélation classique, nous utilisons également la fonction `ggpairs()` du package `GGally`, qui permet une visualisation complète des relations entre les variables :

- Sur la diagonale, les courbes de densité représentant la distribution de chaque variable.
- En dessous, les nuages de points entre chaque paire de variables.
- Au-dessus, les coefficients de corrélation de Pearson.

Cette méthode a l'avantage de combiner analyse graphique et valeurs numériques dans un seul visuel lisible, ce qui facilite l'interprétation visuelle des dépendances, l'identification des structures linéaires, des relations non linéaires, ou des éventuelles ruptures.

```
# install.packages("GGally")
library(GGally)

## Le chargement a nécessité le package : ggplot2

## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2

ggpairs(df)
```



2) Régression linéaire multiple — modèle initial

L'objectif de cette étape est de modéliser la variable mpg (miles per gallon) en fonction des autres caractéristiques techniques des véhicules à l'aide d'un modèle de régression linéaire multiple.

Ce modèle permet d'étudier l'effet simultané de plusieurs variables explicatives sur une variable à expliquer, ici la consommation.

Dans un premier temps, nous incluons toutes les variables disponibles dans un modèle complet, sans sélection préalable.

```r
# Modèle de régression linéaire multiple complet
mod_complet <- lm(mpg ~ ., data = df)

# Résumé du modèle
summary(mod_complet)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4506 -1.6044 -0.1196  1.2193  4.6271
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.30337   18.71788   0.657   0.5181
## cyl         -0.11144    1.04502  -0.107   0.9161
## disp         0.01334    0.01786   0.747   0.4635
## hp          -0.02148    0.02177  -0.987   0.3350
## drat         0.78711    1.63537   0.481   0.6353
## wt          -3.71530    1.89441  -1.961   0.0633 .
## qsec         0.82104    0.73084   1.123   0.2739
## vs           0.31776    2.10451   0.151   0.8814
## am           2.52023    2.05665   1.225   0.2340
## gear         0.65541    1.49326   0.439   0.6652
## carb        -0.19942    0.82875  -0.241   0.8122
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.65 on 21 degrees of freedom
## Multiple R-squared:  0.869,  Adjusted R-squared:  0.8066
## F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

Le modèle initial inclut toutes les variables explicatives disponibles dans le jeu de données pour prédire la variable mpg (consommation en miles par gallon).

L'ajustement global du modèle est satisfaisant, avec un R^2 multiple de 0.869, indiquant que près de 87 % de la variabilité de la consommation est expliquée par les variables du modèle. L'ajustement corrigé (Adjusted R^2 = 0.8066) reste également élevé, ce qui montre que le modèle reste pertinent malgré le nombre important de variables.

Cependant, à l'analyse des p-values (colonnes Pr(>|t|)), on constate que la plupart des coefficients ne sont pas significativement différents de zéro au seuil de 5 %. Seule la variable `wt` (poids du véhicule) approche le seuil de signification avec un p-value ≈ 0.063. Cela indique que certaines variables pourraient ne pas contribuer significativement au modèle, ou que des effets de multicolinéarité réduisent la clarté de leurs contributions.

Par ailleurs, l'erreur standard des résidus est de 2.65, ce qui donne une idée de l'écart typique entre les valeurs observées et les valeurs prédites par le modèle.

Ces observations justifient l'étape suivante de sélection de variables, afin de simplifier le modèle sans perdre en qualité de prédiction.

3) Sélection manuelle de variables et comparaison de modèles

3.1) Choix initial basé sur le modèle complet et les corrélations

À partir des résultats du modèle complet, on observe que plusieurs variables présentent des p-values élevées, indiquant qu'elles ne sont pas statistiquement significatives pour expliquer la variable `mpg`.

Parmi elles, des variables comme `cyl`, `disp`, `hp`, `drat` ou `carb` semblent avoir peu d'impact individuel sur la consommation lorsqu'elles sont incluses avec les autres.

En complément, l'analyse de la matrice des corrélations montre que certaines variables sont fortement corrélées entre elles :

- `wt` (poids), `cyl` (cylindres), `hp` (puissance) et `disp` (cylindrée) sont redondantes : elles capturent globalement la même information liée à la taille et puissance du moteur.
- Il n'est donc pas pertinent de toutes les inclure dans un même modèle, sous peine de colinéarité.

Nous choisissons donc dans un premier temps de construire un modèle réduit avec les variables suivantes :

- `wt` : car elle est significative (p ≈ 0.06) et représente efficacement la masse du véhicule.
- `qsec` : car elle capte l'effet de l'accélération, et est faiblement corrélée aux autres.
- `am` : la transmission, qui est qualitative binaire et potentiellement explicative.

- `cyl` : meme si logiquement parlant, le nombre de cylindres intervient dans la consommation `mpg`, cette variable est incluse à titre de test, car elle est très corrélée à `wt`, mais nous voulons tester son apport dans un modèle partiel .

3.2) Modèle partiel : mpg ~ wt + qsec + am + cyl

```
mod_partiel <- lm(mpg ~ wt + qsec + am + cyl, data = df)
summary(mod_partiel)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am + cyl, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5711 -1.4461 -0.7698  1.5246  4.7207
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  14.9289    12.9059   1.157 0.257507
## wt           -3.6439     0.9097  -4.006 0.000436 ***
## qsec          1.0126     0.5234   1.935 0.063559 .
## am            2.4679     1.7183   1.436 0.162404
## cyl          -0.3542     0.7207  -0.491 0.627060
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.493 on 27 degrees of freedom
## Multiple R-squared:  0.851,  Adjusted R-squared:  0.8289
## F-statistic: 38.55 on 4 and 27 DF,  p-value: 8.603e-11
```

Nous avons construit un modèle partiel incluant les variables `wt`, `qsec`, `am` et `cyl`, sélectionnées manuellement en se basant sur les résultats du modèle complet et la matrice des corrélations.

Le résumé du modèle mod_partiel montre un ajustement global très satisfaisant, avec un R^2 ajusté de 0.8289, légèrement supérieur à celui du modèle complet (0.8066). Cela signifie que ce modèle parvient à expliquer une part légèrement plus importante de la variabilité de la consommation (**mpg**) avec moins de variables, ce qui va dans le sens d'un modèle plus efficace et plus parcimonieux.

- La variable `wt` (poids du véhicule) ressort comme fortement significative avec un p-value < 0.001, confirmant son rôle central dans l'explication de la consommation.

- La variable qsec (temps d'accélération) est modérément significative ($p \approx 0.06$), ce qui peut être acceptable dans un contexte exploratoire ou au seuil de 10 %.

- La variable `am` (type de transmission) n'est pas significative au seuil de 5 %, mais reste relativement proche ($p \approx 0.16$) et possède un effet interprétable, ce qui justifie sa présence pour le moment.

- En revanche, la variable `cyl` (nombre de cylindres) présente une p-value très élevée (≈ 0.63), ce qui indique qu'elle n'apporte aucune information significative dans ce modèle, probablement en raison de sa forte corrélation avec `wt`. Autrement dit, l'effet de `cyl` est déjà capté par `wt`, et son inclusion n'améliore pas le modèle.

Pour vérifier si ce modèle est statistiquement équivalent ou meilleur que le modèle complet, on peut comparer les deux via un test ANOVA.

```
anova(mod_partiel, mod_complet)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ wt + qsec + am + cyl
## Model 2: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     27 167.78
## 2     21 147.49  6     20.29 0.4815 0.8146
```

Le test ANOVA permet de savoir si l'ajout des 6 variables supplémentaires dans le modèle complet améliore significativement la qualité du modèle.

Ici, le p-value = 0.8146 est largement supérieur à 0.05, ce qui signifie qu'il n'y a pas de gain significatif en termes de réduction de l'erreur (RSS).

➤ On ne rejette pas l'hypothèse nulle : les variables ajoutées dans le modèle complet n'apportent pas d'information significative par rapport au modèle partiel.

3.3) Nouveau modèle sans cyl : mpg ~ wt + qsec + am

À la suite de l'analyse du modèle partiel, nous avons décidé de supprimer la variable `cyl`, qui s'était révélée non significative dans le modèle précédent. Ce choix est justifié par :

- sa p-value élevée (≈ 0.63),
- sa corrélation forte avec `wt`, qui capture déjà l'essentiel de son effet,
- et la volonté de simplifier le modèle sans perte d'information.

Nous avons donc estimé un nouveau modèle avec uniquement les variables `wt`, `qsec` et `am`.

```
mod_final <- lm(mpg ~ wt + qsec + am, data = df)
summary(mod_final)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## am            2.9358     1.4109   2.081 0.046716 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.459 on 28 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

Le modèle final présente un excellent ajustement, avec un :

- R^2 ajusté de 0.8336, très proche de celui du modèle partiel (0.8289),
- une erreur standard résiduelle de 2.459, légèrement meilleure que celle du modèle partiel (2.493),
- et surtout, toutes les variables sont significatives au seuil de 5 %, ce qui renforce la robustesse du modèle.

Nous avons comparé mod_final au modèle précédent mod_partiel (qui incluait `cyl`) via un test ANOVA :

```
anova(mod_final, mod_partiel)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ wt + qsec + am
## Model 2: mpg ~ wt + qsec + am + cyl
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     28 169.29
## 2     27 167.78  1    1.5011 0.2416 0.6271
```

Le p-value = 0.6271 est largement supérieur à 0.05, ce qui indique que l'ajout de `cyl` n'apporte pas d'amélioration significative au modèle.

De plus, les indicateurs statistiques sont similaires, voire légèrement meilleurs dans le modèle final : R^2 ajusté plus élevé et erreur plus faible.

3.4) Conclusion sur les modèles

Au terme de cette sélection manuelle et progressive, nous retenons le modèle :

mpg ~ wt + qsec + am

comme modèle final, pour les raisons suivantes :

- Il est simple (3 variables),
- Il est statistiquement performant (R^2 ajusté élevé, résidus faibles),
- Toutes ses variables sont significatives et interprétables,
- Il évite la redondance causée par la multicolinéarité.

Ce modèle constitue donc une base solide et cohérente pour modéliser la consommation (**mpg**) et sera utilisé pour la suite de l'analyse, notamment pour les comparaisons avec l'ACP.

4) Impact de l'ajout d'un individu

Afin d'évaluer la robustesse du modèle de régression face à des valeurs atypiques, nous avons ajouté un individu artificiel extrême au jeu de données.

Cet individu représente une voiture aux caractéristiques extrêmes, très éloignées de celles des autres véhicules de l'échantillon :

```
# Nouvelle observation
new_row <- data.frame(
mpg = 5, cyl = 12, disp = 500, hp = 400, drat = 2,
wt = 6, qsec = 12, vs = 0, am = 0, gear = 3, carb = 8)

# créer un nouveau dataframe contenant la base de données initiale + nouvelle
observation
df2 <- rbind(df, new_row)
```

Ce véhicule est extrêmement lourd, puissant, avec une forte cylindrée et une consommation très élevée (`mpg = 5`), une accélération très rapide (`qsec = 12`), et une transmission automatique. Il s'agit d'un véhicule très éloigné des données habituelles, inséré volontairement pour tester la stabilité du modèle.

Nous avons ensuite ré-estimé le modèle de régression avec les variables `wt`, `qsec` et `am`, et comparé les résultats avant et après l'ajout de cet individu :

```
mod_avec_new <- lm(mpg ~ wt + qsec + am, data = df2)

# Comparaison des coefficients
summary(mod_final)$coefficients

##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  9.617781  6.9595930  1.381946 1.779152e-01
## wt          -3.916504  0.7112016 -5.506882 6.952711e-06
## qsec         1.225886  0.2886696  4.246676 2.161737e-04
## am           2.935837  1.4109045  2.080819 4.671551e-02

summary(mod_avec_new)$coefficients
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 11.285348  6.9514058  1.623463 1.153125e-01
## wt          -3.683002  0.7003424 -5.258860 1.235372e-05
## qsec         1.093564  0.2758152  3.964842 4.401044e-04
## am           3.002411  1.4307062  2.098552 4.467916e-02
```

```
# Comparaison des R² et erreurs
summary(mod_final)$adj.r.squared
```

```
## [1] 0.8335561
```

```
summary(mod_avec_new)$adj.r.squared
```

```
## [1] 0.8521187
```

Après l'ajout de cette observation atypique, on constate que :

- Les coefficients du modèle sont légèrement modifiés, mais conservent le même signe et restent statistiquement significatifs.
- Le R^2 ajusté augmente légèrement, passant de 0.834 à 0.852, ce qui montre que l'individu, bien que très différent, renforce la relation linéaire globale au lieu de la perturber.
- L'interprétation des coefficients reste stable, et aucun comportement aberrant n'apparaît dans la régression.

Cette expérience montre que le modèle sélectionné est robuste et stable, même en présence d'un individu atypique. Il conserve des coefficients significatifs, une structure interprétable, et ne se laisse pas fortement influencer par des valeurs extrêmes isolées. Cela renforce la confiance que l'on peut avoir dans son usage pour l'analyse et la prédiction.

5) Analyse en composangtes principales : ACP

L'Analyse en Composantes Principales (ACP) a été réalisée sur l'ensemble des variables quantitatives du jeu de données `mtcars`, à l'exception de la variable `mpg` qui constitue notre variable cible.

Toutes les autres variables ont été centrées et réduites afin de neutraliser les effets d'échelle et de garantir que chaque variable contribue de manière équitable au calcul des composantes principales.

L'objectif de cette ACP est double : d'une part, étudier les corrélations et redondances entre les variables techniques du véhicule, et d'autre part, explorer la possibilité de réduire la dimension du jeu de données sans perte d'information significative.

```
# Suppression de mpg
acp_data <- df[, -1]
```

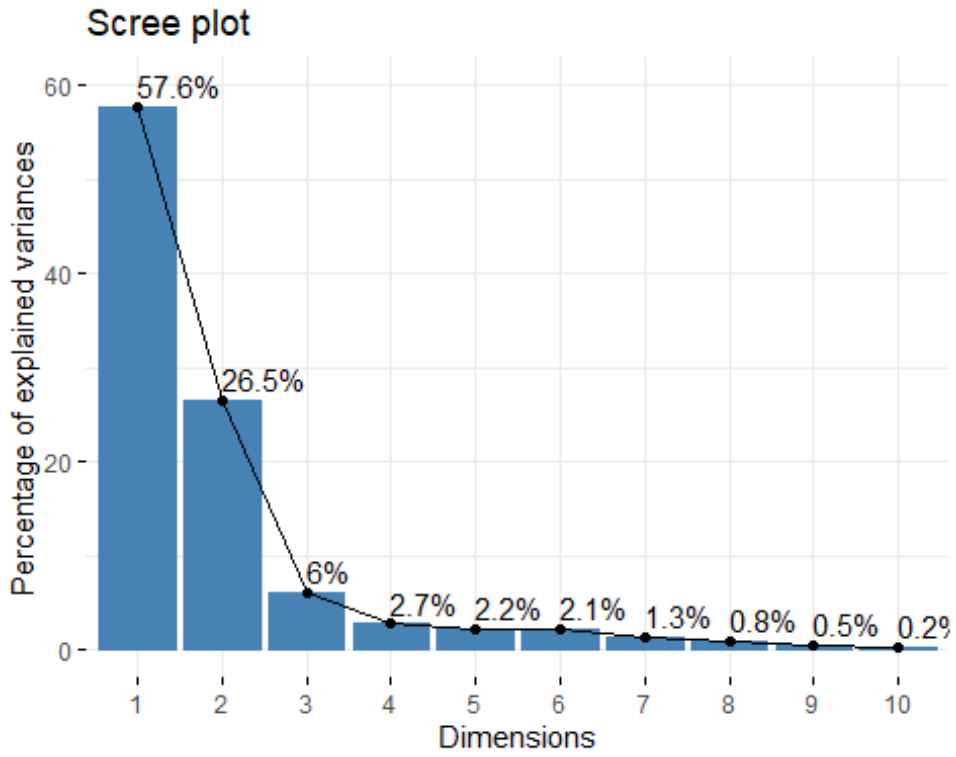
```r
# ACP avec centrage-réduction
res_acp <- prcomp(acp_data, scale. = TRUE)

# Résumé de l'inertie
summary(res_acp)
```

```
## Importance of components:
##                           PC1   PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     2.400 1.628 0.77280 0.51914 0.47143 0.45839 0.36458
## Proportion of Variance 0.576 0.265 0.05972 0.02695 0.02223 0.02101 0.01329
## Cumulative Proportion  0.576 0.841 0.90071 0.92766 0.94988 0.97089 0.98419
##                           PC8     PC9    PC10
## Standard deviation     0.28405 0.23163 0.15426
## Proportion of Variance 0.00807 0.00537 0.00238
## Cumulative Proportion  0.99226 0.99762 1.00000
```

```r
library(factoextra)
```

```
## Welcome! Want to learn more? See two factoextra-related books at
https://goo.gl/ve3WBa
```

```r
fviz_eig(res_acp, addlabels = TRUE, ylim = c(0, 60))
```



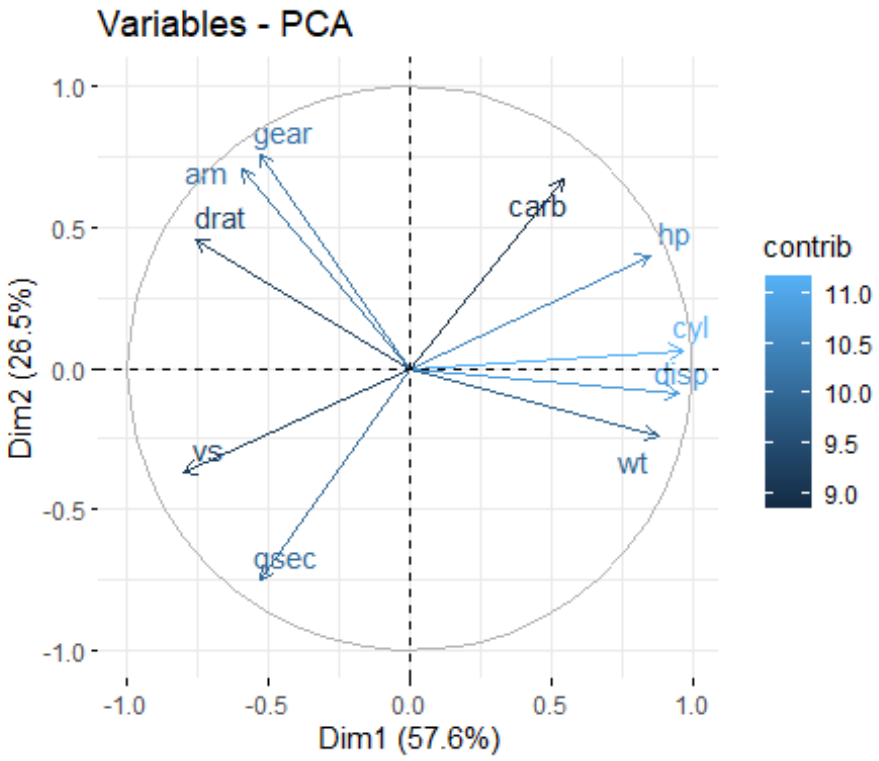
Après réalisation de l'ACP avec la fonction prcomp(), l'analyse de l'inertie cumulée révèle que la première composante principale (PC1) explique à elle seule 57,6 % de la variabilité totale du jeu de données.

La seconde composante (PC2), quant à elle, en explique 26,5 %.

Ainsi, les deux premiers axes résument à eux seuls environ 84,1 % de l'information initiale. Cette forte part de variance cumulée montre qu'il est possible de représenter les données dans un plan bidimensionnel tout en conservant l'essentiel de leur structure.

À partir de la troisième composante, l'apport en variance devient marginal (moins de 6 % par axe), ce qui confirme que l'interprétation peut raisonnablement se concentrer sur le plan formé par PC1 et PC2.

```
# Cercle de corrélation
fviz_pca_var(res_acp, col.var = "contrib", repel = TRUE)
```



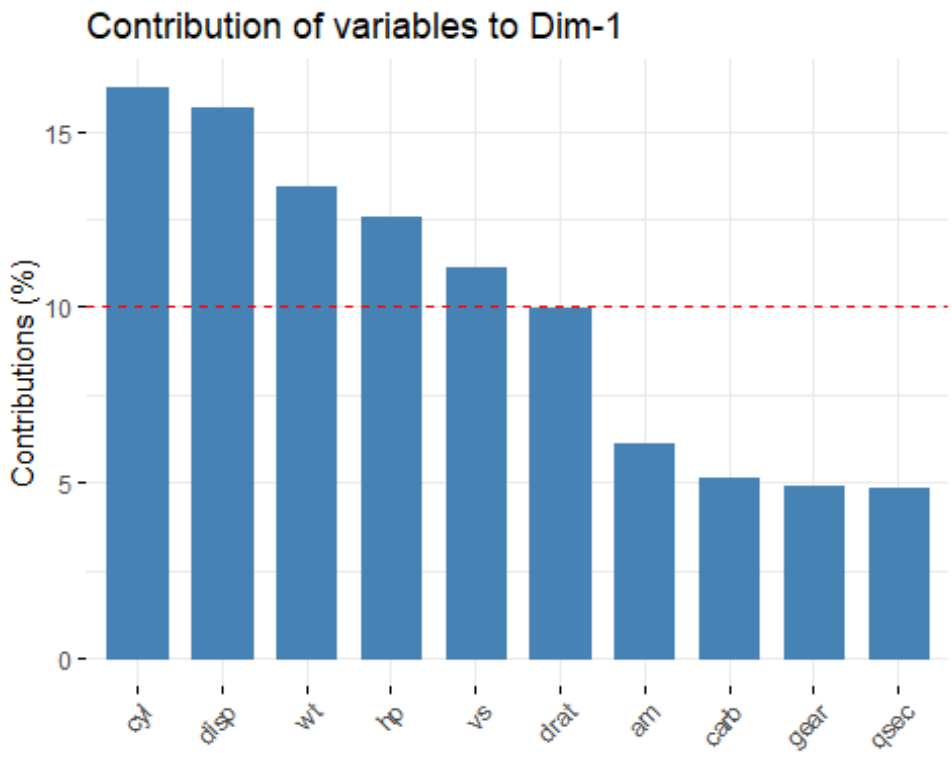
Le cercle des corrélations nous permet d'étudier les relations entre les variables et leur contribution aux axes principaux.

Il met en évidence une forte contribution de certaines variables comme `wt`, `hp`, `disp`, `cyl` et `vs` sur la première composante. Ces variables sont orientées dans la même direction et sont fortement corrélées entre elles, ce qui laisse penser que PC1 synthétise une information liée à la taille et la puissance du véhicule.

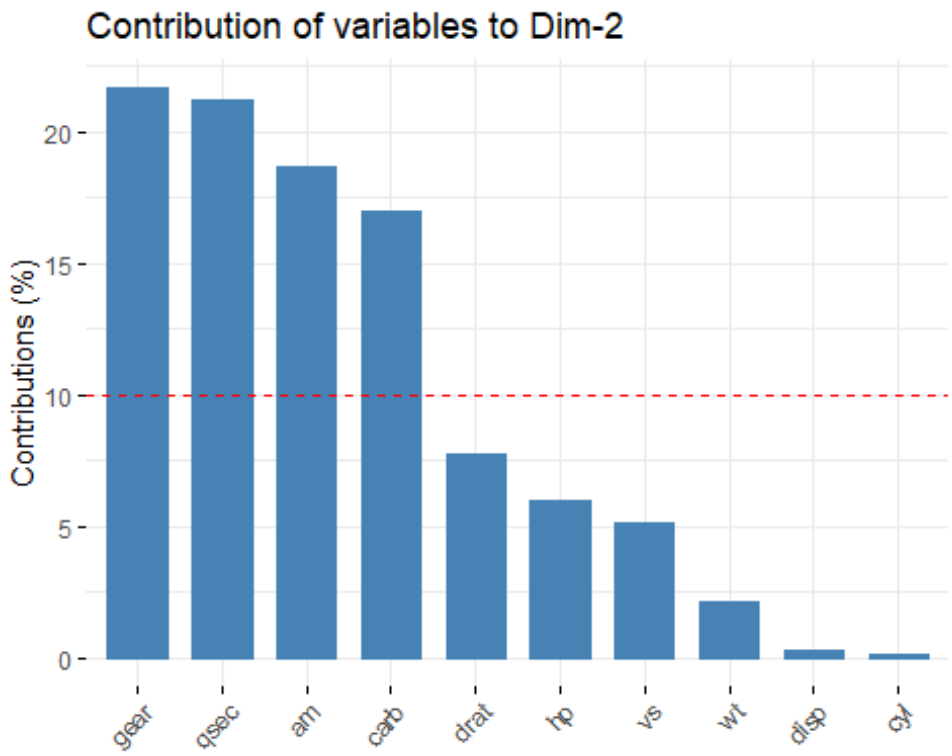
À l'opposé, les variables `qsec` et `drat` apparaissent dans une direction opposée, traduisant une certaine opposition entre les véhicules puissants et les véhicules plus légers ou avec un rapport de pont plus élevé.

Enfin, la variable `am` est positionnée perpendiculairement aux autres, suggérant qu'elle est peu corrélée avec PC1 mais davantage portée par l'axe PC2.

```
# Contribution des variables à PC1
fviz_contrib(res_acp, choice = "var", axes = 1)
```




```
# Contribution des variables à PC2
fviz_contrib(res_acp, choice = "var", axes = 2)
```



L'analyse des contributions des variables confirme cette interprétation.

Pour PC1, les variables les plus contributrices sont `wt`, `hp`, `disp` et `cyl`, ce qui montre que cet axe résume la dimension mécanique du véhicule (puissance, cylindrée, poids). Pour PC2, ce sont `gear`, `qsec` et `am` qui dominent, ce qui donne à cet axe une signification plus comportementale, orientée vers la dynamique du véhicule et le type de transmission.

Ainsi, l'ACP nous permet de résumer l'information contenue dans les 10 variables initiales en deux axes principaux, l'un reflétant une dimension "mécanique / puissance" (PC1) et l'autre une dimension "performance / comportement de conduite" (PC2).

Cette réduction de dimension, en plus de simplifier la représentation graphique, ouvre la voie à une éventuelle régression sur les composantes principales.

6) Regroupement d'individus et de variables

L'objectif ici est de visualiser les éventuelles structures dans les données, à travers deux représentations :

- Le cercle des corrélations permet d'analyser les regroupements de variables.
- La projection des individus (voitures) dans le plan principal (formé par les axes PC1 et PC2) permet de détecter d'éventuels groupes d'individus similaires.

6.1) Regroupement de variables

Le cercle des corrélations ci-dessus présente les variables projetées sur le plan principal formé par les deux premières composantes principales, PC1 (57.6 % de l'inertie) et PC2 (26.5 %), ce qui permet d'interpréter environ 84 % de la variance totale.

Un premier groupe de variables très corrélées entre elles se distingue nettement : `cyl` (nombre de cylindres), `disp` (cylindrée), hp (puissance) et `wt` (poids). Ces variables pointent dans la même direction (à droite) et forment un cluster très resserré. Cela traduit une forte redondance d'information entre ces variables et leur lien avec la première dimension PC1.

À l'opposé (en haut à gauche), un autre petit groupe de variables corrélées positivement entre elles mais négativement corrélées aux précédentes apparaît : `gear` (nombre de vitesses), `am` (type de transmission) et `drat` (rapport de transmission arrière). Ces variables sont donc opposées en termes de profil à celles du premier groupe.

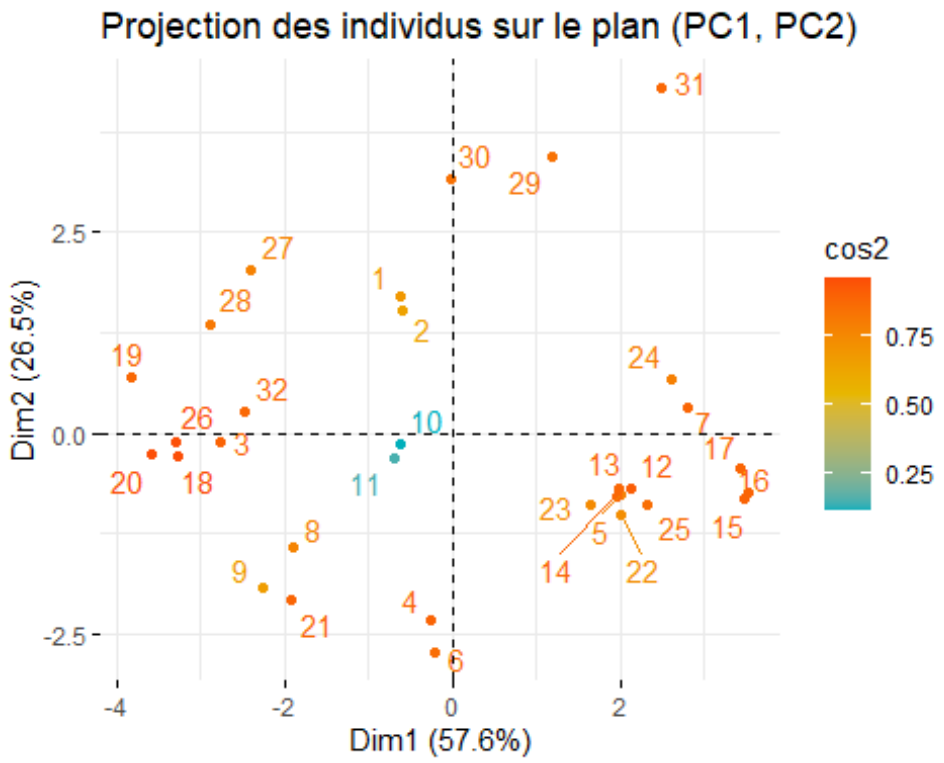
La variable `qsec` (temps sur 1/4 mile) est relativement isolée et projette essentiellement sur PC2 (elle est perpendiculaire aux autres), ce qui indique qu'elle apporte une information indépendante des autres variables principales.

`vs` (configuration moteur) est également partiellement corrélée à PC2 et inversement corrélée à wt, hp, cyl et disp.

En résumé, les axes principaux permettent une bonne séparation des variables en groupes significatifs, reflétant probablement deux grands types de motorisation ou de style de conduite.

6.2) Regroupement d'individus

```r
fviz_pca_ind(res_acp,
             col.ind = "cos2",  # couleur selon la qualité de représentation
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE,
             title = "Projection des individus sur le plan (PC1, PC2)")
```



La projection des individus sur le plan (PC1, PC2) permet d'observer plusieurs regroupements naturels de points :

Les individus situés à droite (ex : 16, 17, 7, 13…) sont ceux qui ont des valeurs élevées sur les variables wt, cyl, disp, hp. Il s'agit donc probablement de véhicules puissants, lourds, à forte cylindrée.

À l'opposé, les individus à gauche (ex : 20, 18, 26, 3…) sont plus légers, avec des moteurs plus petits, et plus efficients. Ils sont probablement associés à une conduite plus économique et/ou des modèles à transmission automatique.

En haut à gauche (ex : 27, 28), on retrouve des véhicules associés à des valeurs élevées de gear, am et drat. Ils pourraient correspondre à des voitures sportives légères à boîte manuelle.

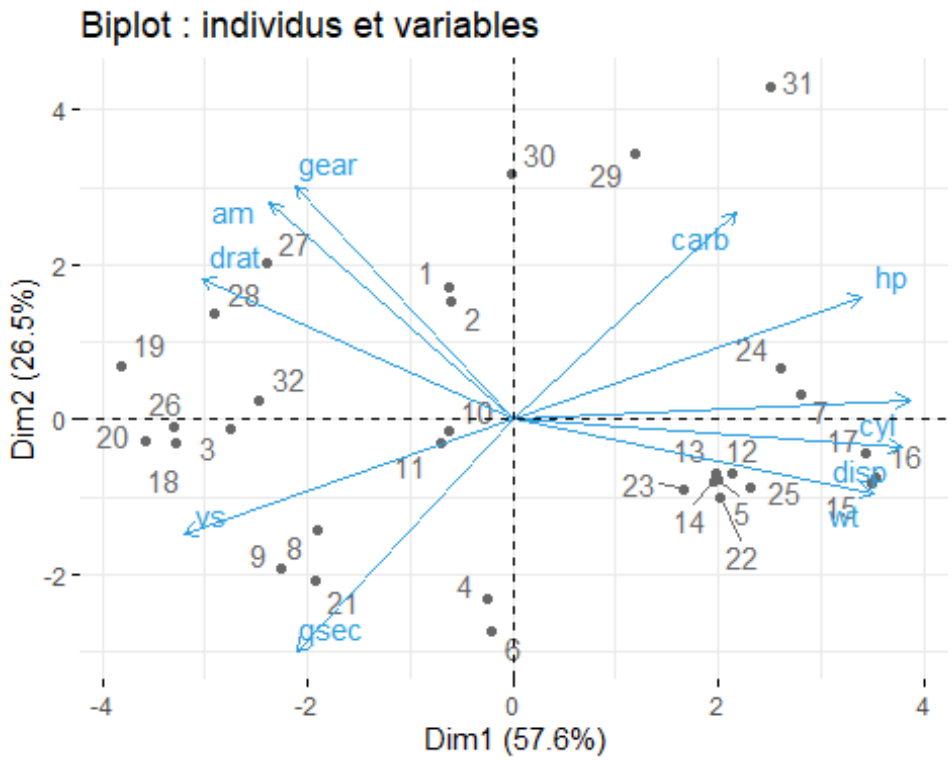
Le point 31, très excentré en haut à droite, semble atypique, avec probablement des valeurs extrêmes sur plusieurs variables (puissance, poids, cylindrée). Il pourrait s'agir d'un modèle unique ou très haut de gamme.

L'intensité de la coloration cos2 dans le graphique des individus permet également d'identifier les points bien représentés par les deux premières composantes (forte couleur orange) par rapport à ceux qui le sont moins (ex. point 10, faible cos2).

6.3) Lien entre les deux

```
fviz_pca_biplot(res_acp,
                repel = TRUE,
```

```
                col.var = "#2E9FDF",
                col.ind = "#696969",
                title = "Biplot : individus et variables")
```



Le biplot combinant variables et individus illustre visuellement les liens entre les variables et les clusters d'individus. Par exemple :

Les individus situés vers la droite sont alignés avec les variables `wt`, `disp`, `cyl`, ce qui confirme leur forte motorisation.

Inversement, ceux placés vers la gauche (près de `am`, `gear`, `drat`) sont associés à des transmissions manuelles et de meilleures performances routières.

Cette double représentation permet d'interpréter le positionnement des véhicules en fonction de leurs caractéristiques techniques.

L'ACP révèle deux grandes familles de variables (puissance/poids vs. maniabilité/transmission) et met en évidence des regroupements clairs parmi les voitures étudiées. Ces clusters peuvent refléter des typologies de véhicules (sportives, familiales, citadines, etc.), confirmant la pertinence de la réduction de dimension pour la visualisation et l'interprétation.

7)Régression sur les deux premières composantes principales (PC1 et PC2)

Après avoir effectué une analyse en composantes principales (ACP), nous avons décidé d'utiliser les deux premières composantes principales (PC1 et PC2) comme variables explicatives pour construire un modèle de régression linéaire visant à prédire la variable cible mpg (miles per gallon).

Ce choix repose sur le fait que PC1 et PC2 capturent ensemble une part très significative de la variance totale des données d'origine (souvent plus de 80 %), tout en limitant le risque de multicolinéarité entre les variables. Utiliser uniquement les premières composantes permet donc de simplifier le modèle tout en conservant une grande partie de l'information pertinente.

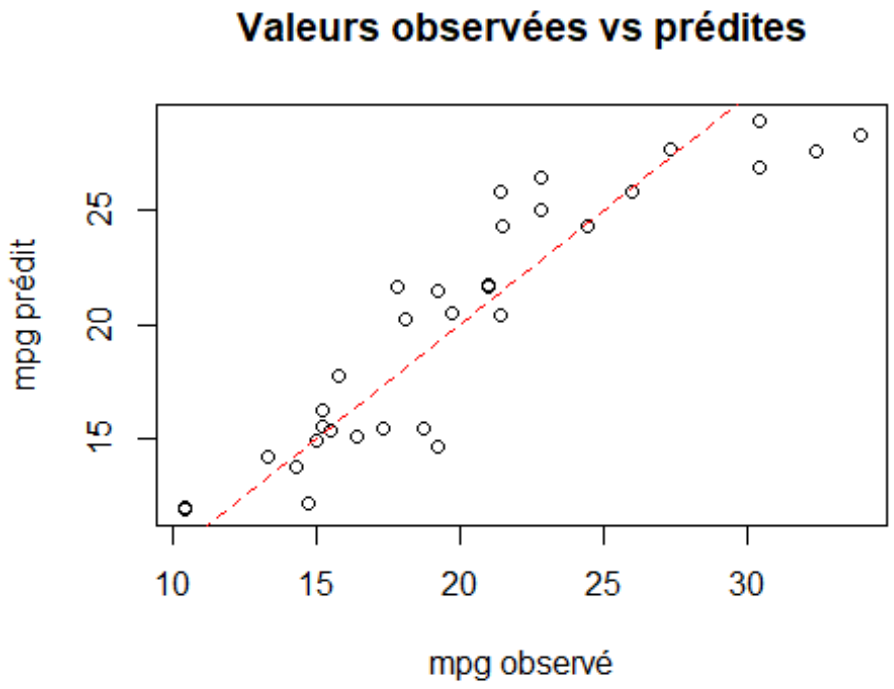
```r
# Récupération de la variable cible
target <- df$mpg

# Récupération des deux premières composantes principales
scores_2PC <- as.data.frame(res_acp$x[, 1:2])  # PC1 et PC2
colnames(scores_2PC) <- c("PC1", "PC2")

# Régression linéaire avec PC1 et PC2 comme variables explicatives
modele_pcr_2 <- lm(target ~ PC1 + PC2, data = scores_2PC)

# Résumé du modèle
summary(modele_pcr_2)
```

```
##
## Call:
## lm(formula = target ~ PC1 + PC2, data = scores_2PC)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3611 -1.7263 -0.3322  1.3208  5.6763
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  20.0906     0.4591  43.760  < 2e-16 ***
## PC1          -2.2813     0.1944 -11.738 1.55e-12 ***
## PC2           0.1163     0.2866   0.406    0.688
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.597 on 29 degrees of freedom
## Multiple R-squared:  0.8263, Adjusted R-squared:  0.8143
## F-statistic: 68.97 on 2 and 29 DF,  p-value: 9.493e-12
```

Les résultats indiquent que le modèle est globalement significatif avec un R^2 ajusté de 0.8143, ce qui signifie que plus de 81 % de la variance de mpg est expliquée par les deux premières

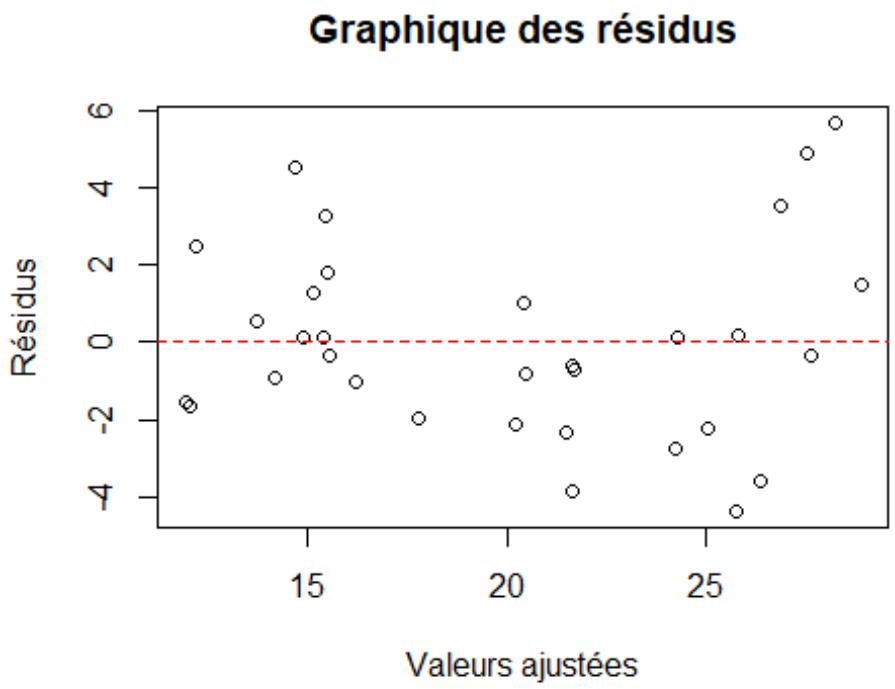
composantes principales. Le coefficient de PC1 est significatif ($p < 0.001$), tandis que celui de PC2 ne l'est pas ($p \approx 0.688$), ce qui suggère que la première composante contient l'essentiel de l'information utile à la prédiction de mpg.

```
# Observé vs. Prédit
plot(target, modele_pcr_2$fitted.values,
     main = "Valeurs observées vs prédites",
     xlab = "mpg observé", ylab = "mpg prédit")
abline(0, 1, col = "red", lty = 2)
```



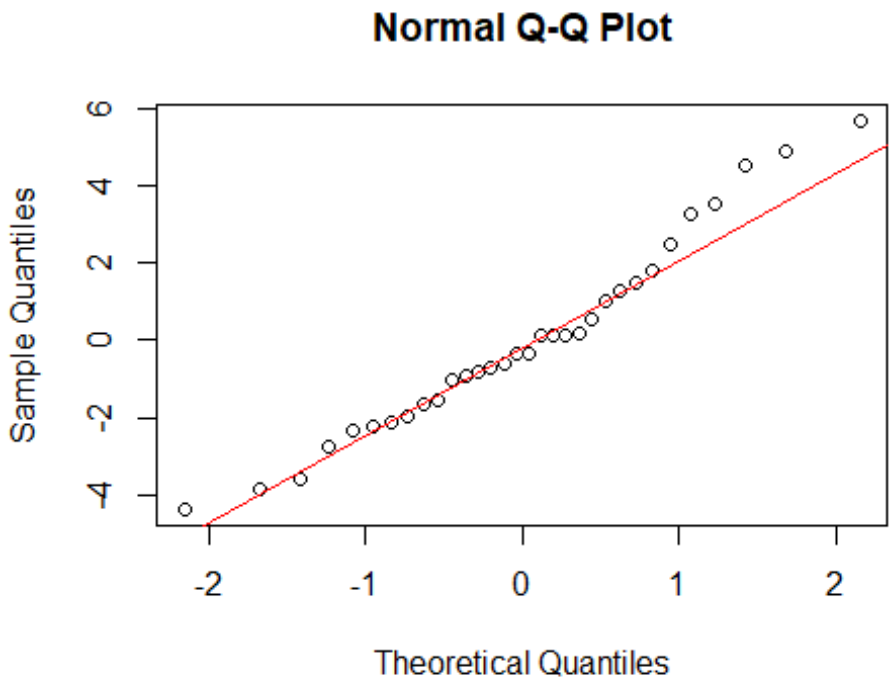
Ce graphique montre une bonne corrélation entre les valeurs prédites par le modèle et les valeurs observées de mpg. Les points sont globalement alignés le long de la droite rouge (régression parfaite), ce qui confirme la qualité du modèle. Quelques écarts sont observables, mais sans tendance particulière, ce qui est attendu dans un modèle imparfait mais robuste.

```
plot(modele_pcr_2$fitted.values, resid(modele_pcr_2),
     xlab = "Valeurs ajustées", ylab = "Résidus",
     main = "Graphique des résidus")
abline(h = 0, col = "red", lty = 2)
```



L'analyse des résidus montre une dispersion relativement homogène autour de zéro, sans motif clair ni structure apparente. Cela suggère que l'hypothèse d'homoscédasticité (variance constante des erreurs) est globalement respectée. Toutefois, quelques résidus plus extrêmes apparaissent, indiquant de potentielles valeurs atypiques (outliers).

```r
# QQ-plot
qqnorm(resid(modele_pcr_2))
qqline(resid(modele_pcr_2), col = "red")
```



Ce graphique permet de vérifier l'hypothèse de normalité des résidus. La majorité des points suivent la ligne droite théorique, bien qu'on observe quelques écarts aux extrémités (queues). Cela peut indiquer une légère déviation de la normalité, mais rien de trop alarmant pour un modèle linéaire.

En résumé, le modèle de régression sur les deux premières composantes principales offre une bonne capacité explicative tout en réduisant la dimension. Cela dit, il convient maintenant de comparer cette approche à celle vue dans les sections précédentes, en soulignant les avantages et les limites de chacune.

8) Avantages et limites des approches classiques et par ACP

Dans cette étude, deux approches de modélisation ont été utilisées pour expliquer la variable `mpg` (consommation) :

- la régression multiple classique, construite à partir des variables originales du dataset,
- et une régression sur les composantes principales, à partir des axes obtenus par l'ACP.

Ces deux méthodes ont chacune leurs forces et leurs faiblesses, que nous comparons ci-dessous.

8.1) Régression avec les variables d'origine

Avantages

- Les coefficients ont une interprétation directe : chaque variable conservant son unité et sa signification réelle, il est facile de comprendre l'effet de chaque prédicteur.
- Permet une analyse détaillée et ciblée de l'effet de chaque variable sur la consommation.
- La mise en œuvre est simple et intuitive, ce qui facilite la lecture des résultats.

Limites

- La présence de multicolinéarité (variables fortement corrélées comme `cyl`, `hp`, `disp`, `wt`) peut rendre les estimations instables ou biaisées.
- L'utilisation d'un grand nombre de variables peut conduire à un modèle complexe et peu interprétable.
- Il existe un risque de surapprentissage si le nombre de variables est élevé par rapport à la taille de l'échantillon (ici, seulement 32 observations).

8.2) Régression sur les composantes principales (ACP)

Avantages

- Permet une réduction de dimension en ne conservant que les composantes expliquant le plus de variance.
- Élimine automatiquement la multicolinéarité, car les composantes principales sont orthogonales (non corrélées).
- Construit un modèle plus stable et plus robuste statistiquement, surtout dans un contexte de données fortement corrélées.

Limites

- Perte d'interprétabilité : les composantes principales sont des combinaisons linéaires de variables, ce qui rend leur interprétation difficile.
- Nécessite une étape supplémentaire de transformation (centrage, réduction, calcul des axes).
- Certaines composantes peuvent être statistiquement significatives, sans qu'on puisse les relier clairement à des variables précises.

8.3) Conclusion

Le choix entre les deux approches dépend des objectifs de l'analyse :

- Si l’objectif est de comprendre et interpréter l’impact de chaque variable technique sur la consommation, la régression classique est à privilégier.
- Si l’objectif est de construire un modèle prédictif plus compact et plus stable, en présence de variables fortement corrélées, alors la régression sur les composantes principales offre une alternative puissante et pertinente.