ENSTA IP PARIS

École Nationale d'Ingénieurs de Tunis

ÉCOLE NATIONALE D'INGÉNIEURS DE TUNIS

ÉCOLE NATIONALE SUPÉRIEURE DES TECHNIQUES AVANCÉES

STA 201

Projet Statistique

Réalisé par :
Rayen ZARGUI

Classe :
2ème année Technique Avancée

Majeure :
Mathématiques appliquées

Encadré par :
Mme Anissa Rabhi

RStudio®

Année universitaire 2024/2025

Table des matières

Table des figures

Introduction

L'objectif de cette analyse est d'étudier la consommation de carburant des véhicules en utilisant des données techniques telles que la cylindrée, la puissance, le poids ou encore le type de transmission comme variables explicatives. Pour ce faire, j'utiliserai le logiciel R afin de mener à bien l'analyse statistique.

Ma démarche consistera tout d'abord à explorer les statistiques descriptives des différentes variables pour dégager des tendances générales et visualiser les distributions. Ensuite, je réaliserai une analyse en composantes principales (ACP) afin de réduire la dimension du jeu de données et d'explorer les relations entre les variables et les observations.

Enfin, une régression linéaire multiple sera menée dans le but d'identifier les facteurs les plus influents sur la consommation de carburant. Une comparaison entre différents modèles, avec ou sans sélection de variables, ainsi que l'ajout d'une nouvelle observation, permettra d'analyser la robustesse et l'interprétation du modèle. L'utilisation conjointe de l'ACP et de la régression sur les composantes principales permettra également de souligner les avantages et les limites de chaque méthode dans le cadre de cette étude.

Chapitre 1

Analyse descriptive et visuelle

1.1 Aquisition des données

Dans un premier temps, j'ai importé le jeu de données à partir d'un fichier CSV, et je l'ai stocké dans l'objet **mtcars**, qui servira de base à l'ensemble de l'analyse. Ensuite, j'ai extrait les variables pertinentes pour l'étude et les ai organisées au sein de sous-ensembles distincts afin de faciliter les différentes étapes de traitement statistique, telles que l'analyse descriptive, l'ACP et la régression multiple.

```
# Chargement des données
mtcars <- read_csv("mtcars.csv")
data("mtcars")
```

FIGURE 1.1 – Code pour importer les données

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

FIGURE 1.2 – Les données

L'ensemble de données mtcars comprend 32 observations et 11 variables. Parmi celles-ci, deux variables sont binaires : **am** (type de transmission, manuel ou automatique) et **vs** (type de moteur, en V ou droit). Trois variables sont discrètes : **gear** (nombre de vitesses), **carb** (nombre de carburateurs) et **cyl** (nombre de cylindres). Les six autres variables, à savoir **mpg** (consommation de carburant), **disp** (cylindrée), **hp** (chevaux), **drat** (rapport d'essieu arrière), **wt** (poids) et **qsec** (temps pour 1/4 de mile), présentent une distribution continue.

1.2 Analyse des données

1.2.1 Analyse descriptive de l'ensemble de données `mtcars`

J'ai entamé l'analyse des données en calculant plusieurs statistiques descriptives pour chacune des variables. Ces mesures incluent le minimum, le maximum, le premier quartile, la médiane, le troisième quartile ainsi que la moyenne. Ces valeurs ont été obtenues à l'aide de la fonction **summary**.

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

FIGURE 1.3 – Représentation descriptive des données

Interprétation : Les statistiques descriptives de l'ensemble de données `mtcars` révèlent des informations clés sur les 11 variables étudiées, représentant 32 modèles de voitures.

La variable `mpg` (consommation de carburant) varie de 10,4 à 33,9 miles par gallon, avec une médiane de 19,2 et une moyenne de 20,09. Cela indique une distribution légèrement asymétrique à droite, où la majorité des voitures ont une consommation modérée, mais quelques modèles sont très économes.

Les variables continues telles que `disp` (cylindrée) et `hp` (puissance en chevaux) présentent une grande variabilité, avec des valeurs allant respectivement de 71,1 à 472 et de 52 à 335. Leurs médianes (196,3 pour `disp` et 123 pour `hp`) sont inférieures aux moyennes (230,7 et 146,7), suggérant une asymétrie à droite due à quelques véhicules très puissants.

Le poids (`wt`) varie de 1,513 à 5,424, avec une médiane de 3,325. Cela signifie que la moitié des voitures pèsent moins de 3,325 unités, mais quelques modèles très lourds tirent la moyenne vers le haut (3,217). La variable `qsec` (temps pour le quart de mile) présente une distribution plus symétrique, allant de 14,5 à 22,9, avec une médiane de 17,71 proche de la moyenne (17,85).

Concernant les variables discrètes, `cyl` (nombre de cylindres) montre que la majorité des voitures possèdent 4, 6 ou 8 cylindres, avec une médiane de 6, ce qui

reflète une prédominance de moteurs de taille moyenne à grande. La variable `gear` (nombre de vitesses) indique une répartition entre 3, 4 et 5 vitesses, avec une médiane de 4. La variable `carb` (nombre de carburateurs) varie de 1 à 8, mais la médiane est de 2, suggérant que les voitures avec 1 ou 2 carburateurs sont les plus courantes.

Enfin, les variables binaires `am` (type de transmission) et `vs` (type de moteur) présentent une répartition équilibrée. Pour `am`, la médiane est 0 (transmission automatique), mais le troisième quartile est à 1, ce qui indique que 40,6 % des voitures ont une transmission manuelle. Quant à `vs`, la médiane est également 0 (moteur en ligne), avec 43,75 % des voitures équipées d'un moteur en V.

Ces données illustrent une diversité dans les caractéristiques des voitures, avec des tendances marquées : les voitures plus lourdes et plus puissantes tendent à consommer davantage de carburant, tandis que les modèles plus légers et équipés de transmissions manuelles sont généralement plus économes. Ces observations préliminaires orientent les analyses ultérieures, notamment sur l'impact de variables telles que le poids et la puissance sur la consommation de carburant.

1.2.2 Nuage des points

En utilisant la fonction 'pairs', j'ai obtenu le nuage de points de mes données pour explorer les relations entre les différentes variables.

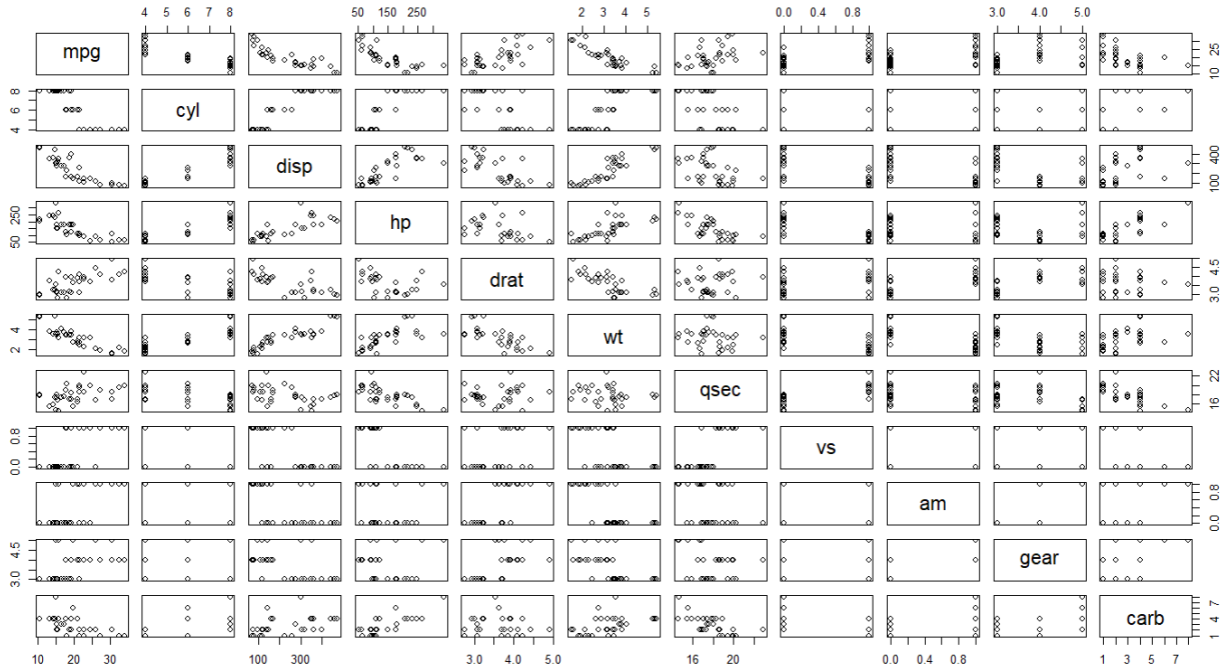

FIGURE 1.4 – Nuage des points

Consommation de carburant (`mpg`) : La variable `mpg`, représentant la consommation en miles par gallon, montre des corrélations notables avec plusieurs autres variables :

— Une **corrélation négative marquée** avec le poids (`wt`) : les voitures plus lourdes consomment davantage de carburant.
— Des **corrélations négatives** également observées avec la cylindrée (`disp`) et la puissance (`hp`), ce qui indique que les voitures plus puissantes ou avec un moteur de plus grande capacité tendent à être moins économes.
— Une **corrélation légèrement positive** avec le temps au quart de mile (`qsec`), suggérant que les voitures moins performantes sur le plan de l'accélération sont souvent plus économes.
— Une association avec la transmission (`am`) : les véhicules à transmission manuelle (`am` = 1) tendent à avoir une consommation plus faible.

Corrélations entre variables techniques :
— Une **forte corrélation positive** est observée entre la cylindrée (`disp`) et la puissance (`hp`).
— Le poids (`wt`) est positivement corrélé à la puissance et à la cylindrée, ce qui est cohérent avec des véhicules plus massifs ayant besoin de moteurs plus puissants.

Variables discrètes : Les variables `cyl` (nombre de cylindres), `gear` (nombre de vitesses), `carb` (nombre de carburateurs), `am` (type de transmission) et `vs` (type de moteur) affichent des regroupements visibles sous forme de bandes, typiques des variables catégorielles ou discrètes. Ces regroupements mettent en évidence certaines configurations mécaniques dominantes :
— La majorité des véhicules ont 4, 6 ou 8 cylindres.
— Les transmissions manuelles sont davantage associées à des voitures à 4 ou 5 vitesses.

1.2.3 Histogrammes

Code

```r
# Histogrammes des variables
plot_list <- list()
for (i in 1:ncol(mtcars)) {
  plot_list[[i]] <- ggplot(mtcars, aes(x = .data[[names(mtcars)[i]]])) +
    geom_histogram(fill = colors[i %% length(colors) + 1], color = "white", bins = 10, alpha = 0.8) +
    labs(title = paste("Distribution de", names(mtcars)[i]),
         x = names(mtcars)[i],
         y = "Fréquence") +
    theme_minimal() +
    theme(plot.title = element_text(size = 10))
}
grid.arrange(grobs = plot_list, ncol = 3)
```

FIGURE 1.5 – Code utilisé

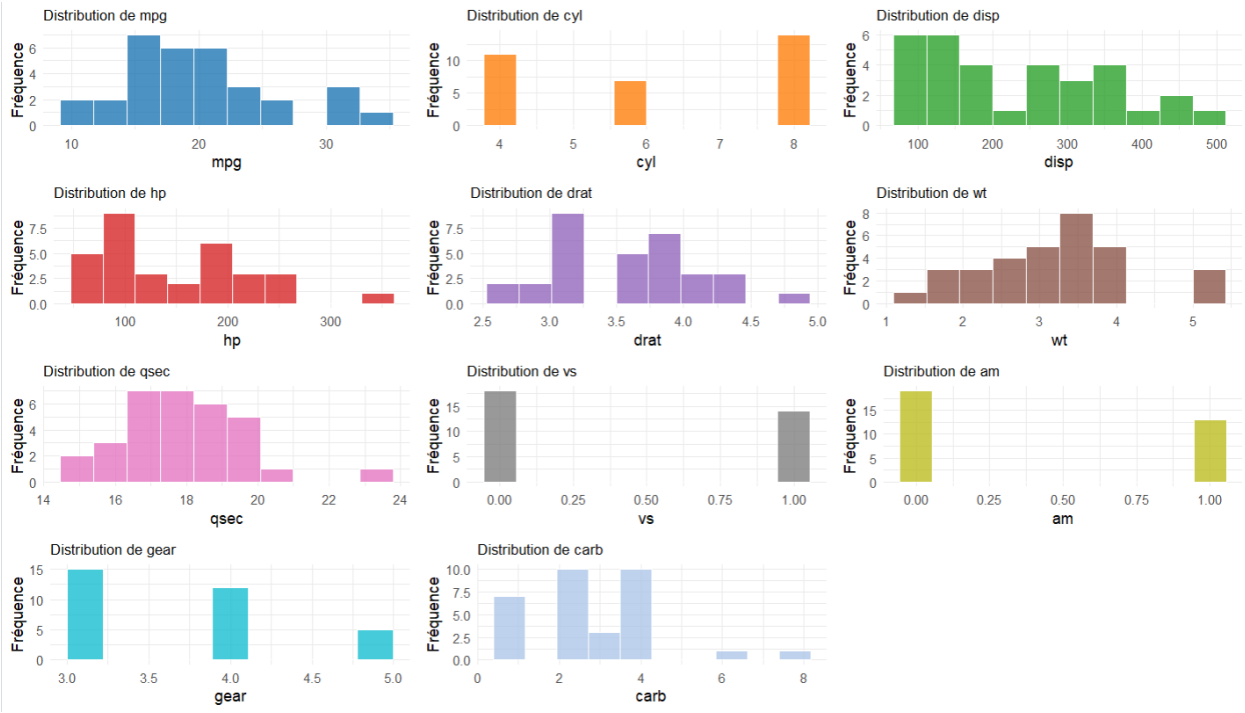

FIGURE 1.6 – Les histogrammes

Interprétation

L'analyse des histogrammes des 11 variables de l'ensemble de données `mtcars` met en évidence les tendances générales et les formes de distribution suivantes :

- **mpg (consommation de carburant)** : la distribution est légèrement asymétrique à droite, avec une majorité de valeurs comprises entre 15 et 25 miles par gallon. Quelques observations atteignent 30 à 35 mpg, correspondant à des véhicules particulièrement économes.
- **cyl (nombre de cylindres)** : la variable est discrète, avec trois niveaux dominants (4, 6 et 8 cylindres). On observe une forte représentation des moteurs à 4 et 8 cylindres, traduisant une polarisation entre véhicules économiques et puissants.
- **disp (cylindrée)** : la distribution est fortement asymétrique à droite. La plupart des observations se situent entre 100 et 200, mais certaines atteignent jusqu'à 500, indiquant la présence de moteurs très volumineux.
- **hp (puissance)** : cette variable suit une distribution similaire à celle de la cylindrée, avec un regroupement entre 50 et 200 chevaux. Quelques modèles très puissants dépassent les 300 chevaux.
- **drat (rapport de pont)** : la distribution est bimodale, avec deux pics autour de 3,0 et 4,0. Cela pourrait refléter deux types de conception de véhicules, l'un favorisant l'accélération, l'autre l'efficacité énergétique.
- **wt (poids)** : la distribution est légèrement asymétrique à droite, avec un pic autour de 3,0 à 3,5. Certaines voitures très lourdes dépassent 5,0, indiquant une hétérogénéité dans les caractéristiques structurelles.
- **qsec (temps pour parcourir 1/4 de mile)** : cette variable présente une distribution relativement symétrique, centrée autour de 17 à 18 secondes.

Quelques valeurs extrêmes sont observées entre 14 et 23 secondes.

— **vs (type de moteur)** : variable binaire (0 = moteur en V, 1 = moteur en ligne), la répartition est presque équilibrée avec une légère majorité de moteurs en V.

— **am (type de transmission)** : cette variable binaire montre une dominance des transmissions automatiques (0), avec environ 40% de transmissions manuelles (1).

— **gear (nombre de vitesses)** : la distribution discrète est concentrée autour de 3 et 4 vitesses. Les véhicules à 5 vitesses sont minoritaires, souvent associés à des modèles sportifs ou récents.

— **carb (nombre de carburateurs)** : la distribution est asymétrique avec un pic à 2 carburateurs. Certains modèles atteignent jusqu'à 8 carburateurs, généralement liés à des performances élevées.

1.2.4 Corrélation

En utilisant la fonction **cor**, j'ai obtenu la matrice de corrélation qui donne la corrélation entre les différentes variables.

	mpg	cyl	disp	hp	drat	wt	qsec	vs
mpg	1.0000000	-0.8521620	-0.8475514	-0.7761684	0.68117191	-0.8676594	0.41868403	0.6640389
cyl	-0.8521620	1.0000000	0.9020329	0.8324475	-0.69993811	0.7824958	-0.59124207	-0.8108118
disp	-0.8475514	0.9020329	1.0000000	0.7909486	-0.71021393	0.8879799	-0.43369788	-0.7104159
hp	-0.7761684	0.8324475	0.7909486	1.0000000	-0.44875912	0.6587479	-0.70822339	-0.7230967
drat	0.6811719	-0.6999381	-0.7102139	-0.4487591	1.00000000	-0.7124406	0.09120476	0.4402785
wt	-0.8676594	0.7824958	0.8879799	0.6587479	-0.71244065	1.0000000	-0.17471588	-0.5549157
qsec	0.4186840	-0.5912421	-0.4336979	-0.7082234	0.09120476	-0.1747159	1.00000000	0.7445354
vs	0.6640389	-0.8108118	-0.7104159	-0.7230967	0.44027846	-0.5549157	0.74453544	1.0000000
am	0.5998324	-0.5226070	-0.5912270	-0.2432043	0.71271113	-0.6924953	-0.22986086	0.1683451
gear	0.4802848	-0.4926866	-0.5555692	-0.1257043	0.69961013	-0.5832870	-0.21268223	0.2060233
carb	-0.5509251	0.5269883	0.3949769	0.7498125	-0.09078980	0.4276059	-0.65624923	-0.5696071

	am	gear	carb
mpg	0.59983243	0.4802848	-0.55092507
cyl	-0.52260705	-0.4926866	0.52698829
disp	-0.59122704	-0.5555692	0.39497686
hp	-0.24320426	-0.1257043	0.74981247
drat	0.71271113	0.6996101	-0.09078980
wt	-0.69249526	-0.5832870	0.42760594
qsec	-0.22986086	-0.2126822	-0.65624923
vs	0.16834512	0.2060233	-0.56960714
am	1.00000000	0.7940588	0.05753435
gear	0.79405876	1.0000000	0.27407284
carb	0.05753435	0.2740728	1.00000000

FIGURE 1.7 – Matrice de corrélation

Et la fonction **corrplot** pour visualiser la corrélation entre les variables :

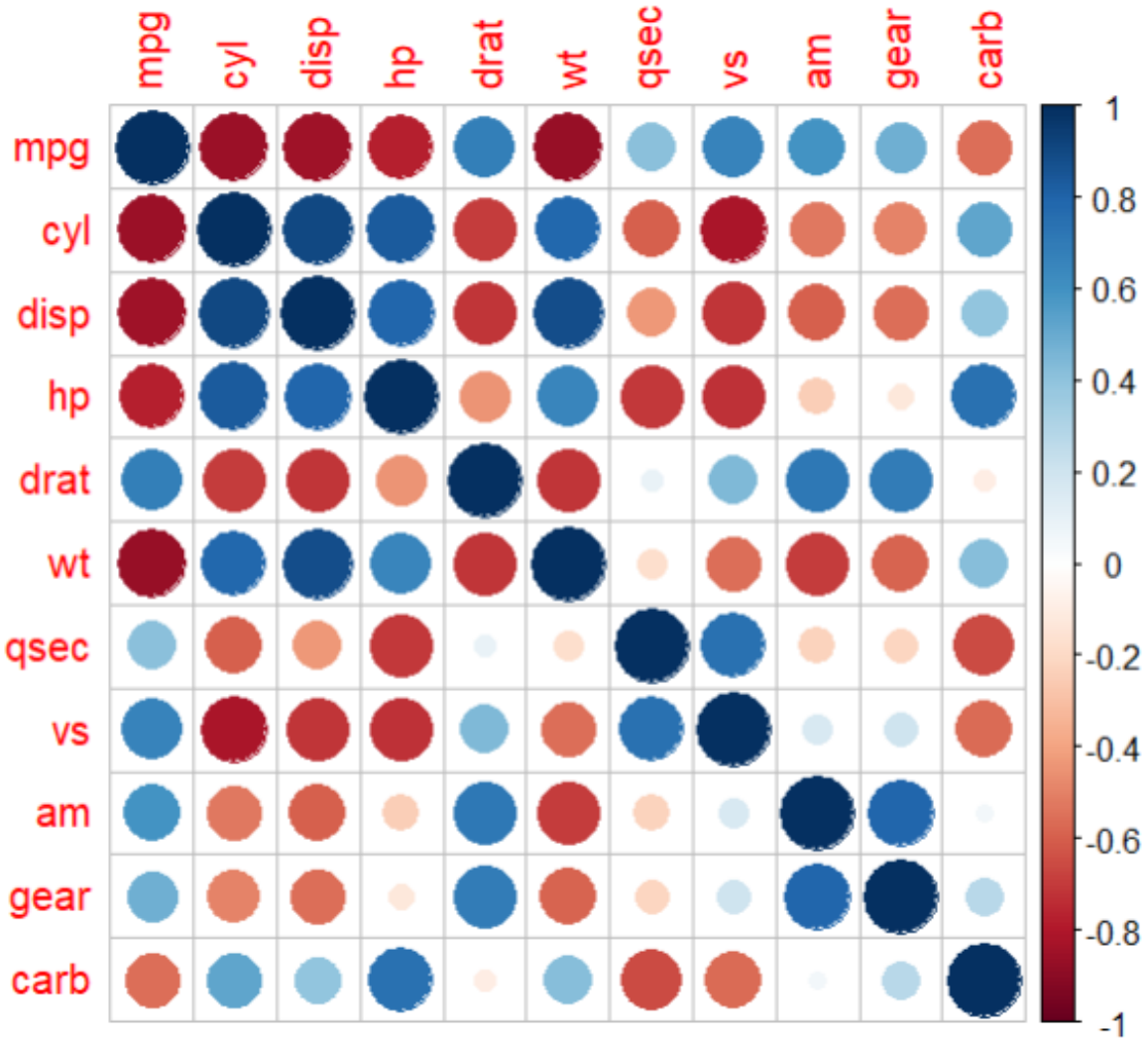

FIGURE 1.8 – Corrplot

Interpétation : La matrice de corrélation des variables de l'ensemble `mtcars` met en évidence plusieurs relations significatives entre les caractéristiques techniques des véhicules. La consommation de carburant (`mpg`) présente des corrélations négatives marquées avec le poids (`wt`, $r = -0{,}87$), la cylindrée (`disp`, $r = -0{,}85$), la puissance (`hp`, $r = -0{,}78$) et le nombre de cylindres (`cyl`, $r = -0{,}85$). Cela confirme que les véhicules plus lourds, puissants ou dotés de moteurs volumineux sont généralement moins économes.

En revanche, `mpg` est positivement corrélée avec le rapport de démultiplication final (`drat`, $r = 0{,}68$) ainsi qu'avec le type de transmission (`am`, $r = 0{,}60$), suggérant que des rapports d'essieux plus élevés et les transmissions manuelles contribuent à une meilleure efficacité énergétique.

Par ailleurs, les variables liées à la puissance et à la taille du moteur – `cyl`, `disp`, `hp` et `wt` – sont fortement interconnectées, avec des coefficients de corrélation allant de 0,79 à 0,90, traduisant leur interdépendance dans la conception des véhicules. Enfin, le temps au quart de mile (`qsec`) est modérément corrélé négativement avec la puissance (`hp`, $r = -0{,}71$), ce qui indique que les voitures plus puissantes tendent

à être plus rapides à l'accélération.

1.2.5 Boxplot

```
1 boxplot(mtcars, main = "Boxplots des variables de mtcars",
    col = "lightblue", las = 2)
```

Listing 1.1 – Boxplot des variables du jeu mtcars

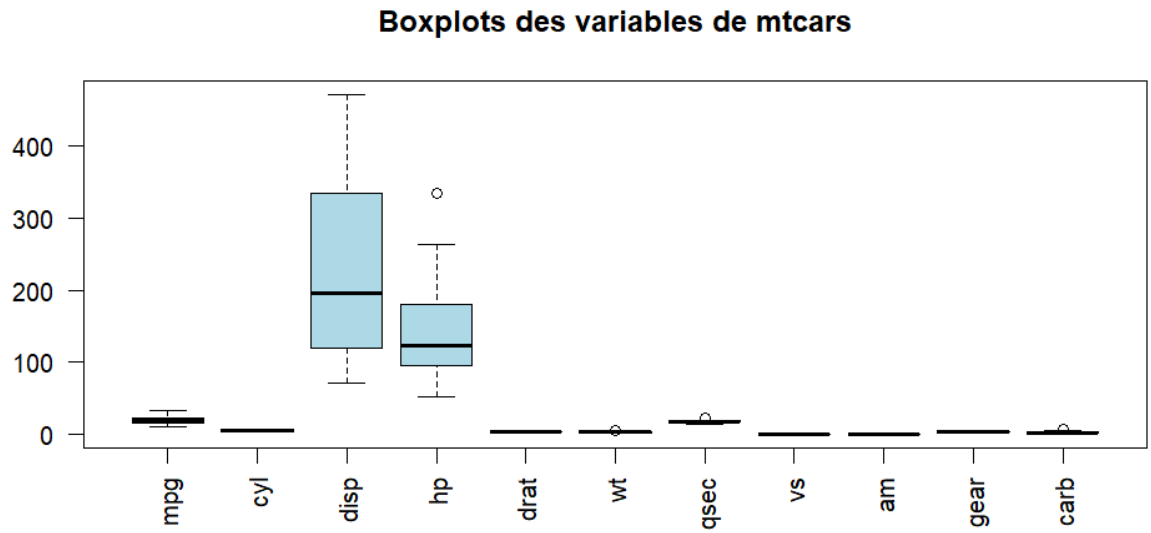

FIGURE 1.9 – Boxplot

Chapitre 2

Régression multiple

2.1 Modéle initial

```r
full_model <- lm(mpg ~ cyl + disp + hp + drat + wt + qsec +
   vs + am + gear + carb, data = mtcars)
```

Listing 2.1 – Modèle de régression linéaire multiple ajusté sur `mtcars`

Le modèle initial est une régression linéaire multiple qui vise à expliquer la variable `mpg` (consommation de carburant) à partir de plusieurs caractéristiques techniques du véhicule. Il inclut des variables telles que le nombre de cylindres (`cyl`), la cylindrée (`disp`), la puissance (`hp`), le poids (`wt`), le type de transmission (`am`), ainsi que d'autres variables.

```
> summary(full_model)

Call:
lm(formula = mpg ~ cyl + disp + hp + drat + wt + qsec + vs +
    am + gear + carb, data = mtcars)

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
```

FIGURE 2.1 – Summary du modéle linéaire

2.2 Modéle par selection

```r
modele_selection <- stepAIC(full_model, direction = "both",
    trace = FALSE)
```

Listing 2.2 – Sélection du modèle par la méthode stepwise (AIC)

Une sélection de variables a été réalisée à l'aide de la méthode pas à pas selon le critère d'Akaike (AIC). Le modèle initial incluait toutes les variables disponibles, mais la sélection a conduit à un modèle plus simple ne retenant que trois variables significatives : le poids du véhicule (wt), le temps sur 1/4 mile (qsec), et le type de transmission (am). Ce modèle réduit améliore la lisibilité tout en conservant une bonne capacité explicative, ce qui en fait un choix pertinent pour prédire la consommation de carburant.

```
> summary(modele_selection)

Call:
lm(formula = mpg ~ wt + qsec + am, data = mtcars)

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
```

FIGURE 2.2 – Summary du modéle par selection

2.3 Comparaison des deux modéles linéaires

Deux modèles de régression multiple ont été construits pour étudier les facteurs influençant la consommation de carburant (`mpg`). Le premier modèle inclut l'ensemble des variables explicatives disponibles dans le jeu de données, tandis que le second a été obtenu par une procédure de sélection afin de ne conserver que les variables les plus pertinentes.

Modèle	**R^2 ajusté**	**Variables retenues**
Modèle complet	0,869	`cyl, disp, hp, drat, wt, qsec, vs, am, gear, carb`
Modèle sélectionné	0,8497	`wt, qsec, am`

TABLE 2.1 – Comparaison des performances des deux modèles de régression

Le modèle réduit, bien que plus simple, présente un R^2 ajusté supérieur à celui du modèle complet, ce qui témoigne d'une meilleure capacité explicative avec un

nombre restreint de variables. Les variables conservées (`wt`, `qsec`, `am`) sont donc les plus influentes sur la consommation de carburant. Ce modèle est également plus facile à interpréter et plus adapté à une analyse pratique.

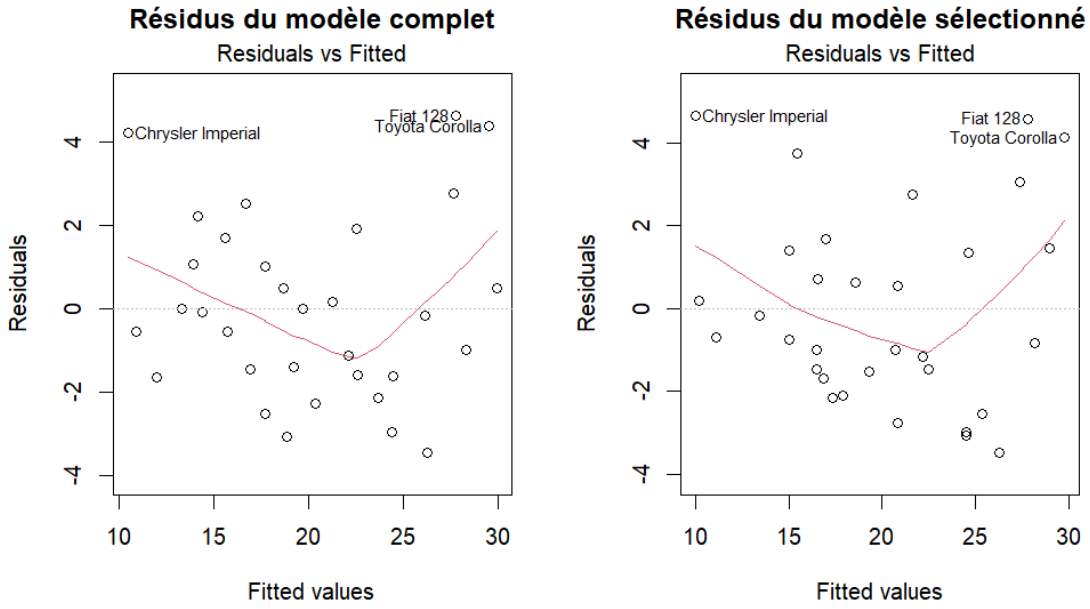

FIGURE 2.3 – Résidus en fonction des valeurs

2.4 Ajout au jeu de données

On fait l'ajout de cette voiture :

```
nouveau_vehicule <- data.frame( mpg = 35, cyl = 4, disp = 95,
    hp = 80, drat = 4.1, wt = 2.0, qsec = 18.5, vs = 1, am =
   1, gear = 5, carb = 1)
```

Listing 2.3 – Création d'un nouveau véhicule avec des caractéristiques spécifiques

2.4.1 Modèle complet

```
> summary(modele_complet_ajout)

Call:
lm(formula = mpg ~ ., data = mtcars_ajout)

Residuals:
    Min      1Q  Median      3Q     Max
-4.5308 -1.5788 -0.0716  1.3848  4.8971

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.38249   20.06726   0.368    0.716
cyl          0.22034    1.11649   0.197    0.845
disp         0.01017    0.01922   0.529    0.602
hp          -0.02080    0.02351  -0.885    0.386
drat         0.68854    1.76571   0.390    0.700
wt          -3.10237    2.02311  -1.533    0.139
qsec         0.74738    0.78854   0.948    0.354
vs           0.67097    2.26625   0.296    0.770
am           2.45247    2.22117   1.104    0.281
gear         1.94186    1.47908   1.313    0.203
carb        -0.74046    0.85330  -0.868    0.395

Residual standard error: 2.863 on 22 degrees of freedom
Multiple R-squared:  0.8656,    Adjusted R-squared:  0.8046
F-statistic: 14.17 on 10 and 22 DF,  p-value: 2.082e-07
```

FIGURE 2.4 – Summary de modèle complet

2.4.2 Modèle réduit

```
> summary(modele_reduit_ajout)

Call:
lm(formula = mpg ~ wt + qsec + am, data = mtcars_ajout)

Residuals:
    Min      1Q  Median      3Q     Max
-4.1045 -1.7014 -0.5203  1.4748  6.9485

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   8.8895     7.8280   1.136 0.265423
wt           -4.0275     0.7993  -5.039 2.28e-05 ***
qsec          1.2889     0.3240   3.978 0.000425 ***
am            3.3715     1.5793   2.135 0.041350 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.767 on 29 degrees of freedom
Multiple R-squared:  0.8345,    Adjusted R-squared:  0.8173
F-statistic: 48.73 on 3 and 29 DF,  p-value: 1.914e-11
```

FIGURE 2.5 – Summary de modèle réduit

2.4.3 Impact sur la regression

Afin d'évaluer la robustesse du modèle, une nouvelle observation représentant un véhicule léger, peu puissant mais très économe a été ajoutée artificiellement au jeu de données. Après réestimation, on observe une légère modification des coefficients du modèle, notamment une réduction de l'effet du poids (`wt`) sur la consommation.

Ce nouveau point a une forte consommation spécifique et un faible poids, ce qui tend à renforcer le lien inverse entre `wt` et `mpg`. Toutefois, les résidus indiquent que cette observation reste dans des limites raisonnables et ne déstabilise pas le modèle de manière excessive.

Chapitre 3

Analyse en composantes principales

L'analyse en composantes principales a été réalisée sur l'ensemble des variables quantitatives du jeu de données mtcars :

```r
# Variables quantitatives completes
vars_quant <- c("mpg", "cyl", "disp", "hp", "drat", "wt", "
   qsec", "vs", "am", "gear", "carb")
data_acp <- mtcars[, vars_quant]

# ACP avec centrage-reduction
res_pca <- PCA(data_acp, scale.unit = TRUE, graph = FALSE)
```

Listing 3.1 – Code R pour l'ACP complète

3.0.1 Valeurs propres et variance expliquée

	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	6.60840025	60.0763659	60.07637
comp 2	2.65046789	24.0951627	84.17153
comp 3	0.62719727	5.7017934	89.87332
comp 4	0.26959744	2.4508858	92.32421
comp 5	0.22345110	2.0313737	94.35558
comp 6	0.21159612	1.9236011	96.27918
comp 7	0.13526199	1.2296544	97.50884
comp 8	0.12290143	1.1172858	98.62612
comp 9	0.07704665	0.7004241	99.32655
comp 10	0.05203544	0.4730495	99.79960
comp 11	0.02204441	0.2004037	100.00000

FIGURE 3.1 – Valeurs propres et variance cumulée

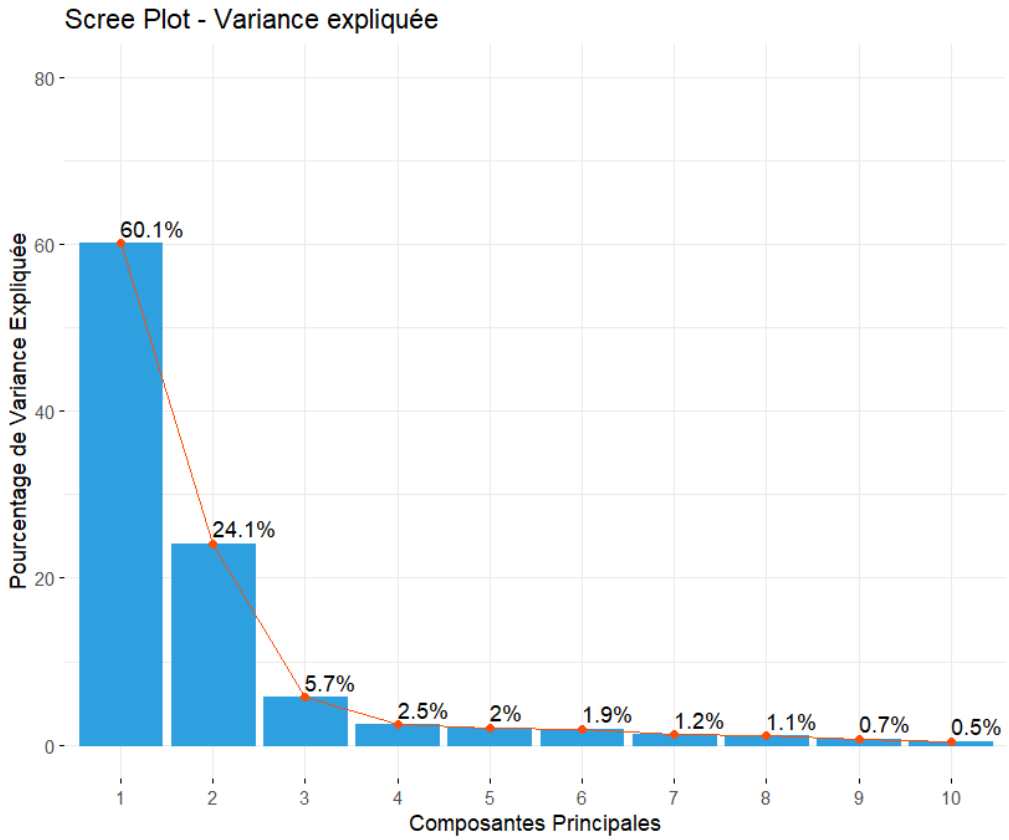

FIGURE 3.2 – Éboulis des valeurs propres (Scree plot)

Le scree plot (Fig. 3.2) montre que :

— Les deux premières composantes capturent **84,2%** de la variance totale
  — Dim1 : 60,1%
  — Dim2 : 24,1%
— La troisième composante explique 5.7% de variance supplémentaire

3.0.2 Cercle des corrélations

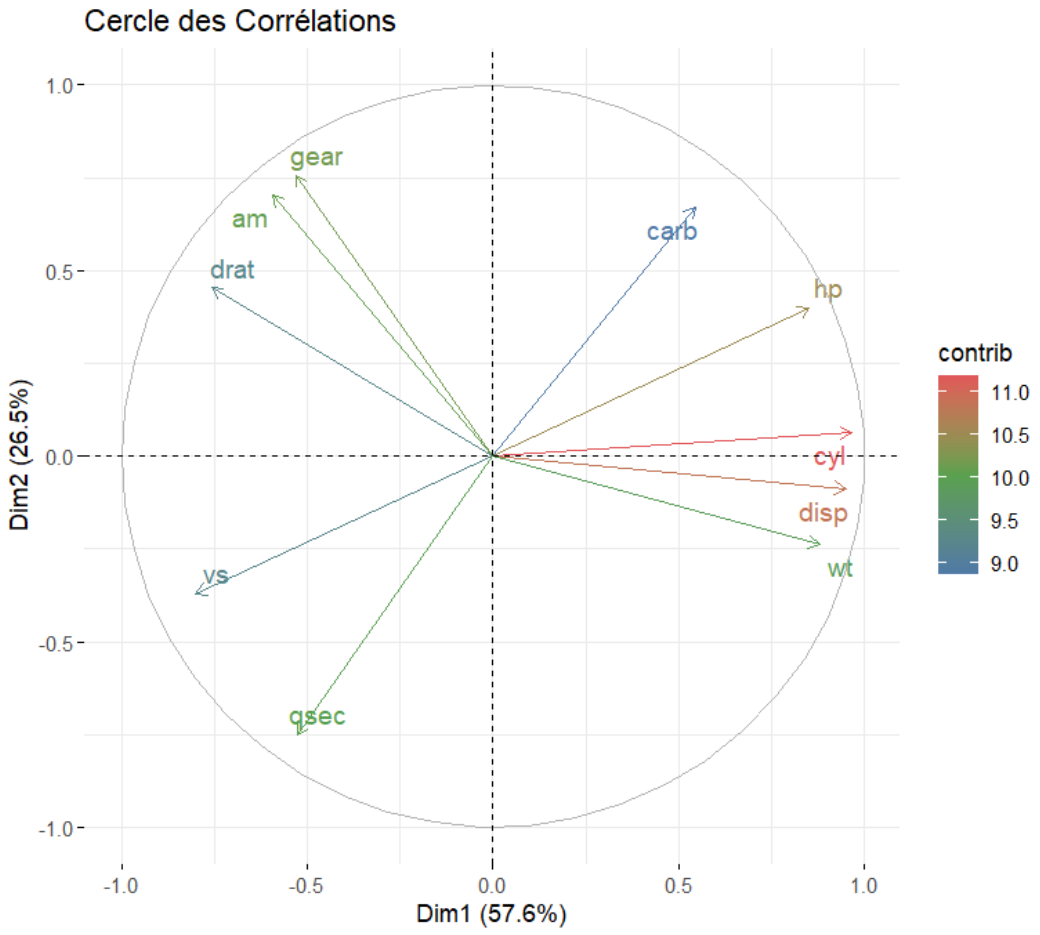

FIGURE 3.3 – Cercle des corrélations des variables

L'analyse du cercle de corrélation (Fig. 3.3) révèle :

— **Axe 1** oppose :
  — Variables de puissance (`hp`, `disp`, `cyl`) à gauche
  — Variables d'efficacité (`mpg`, `drat`) à droite
— **Axe 2** sépare :
  — Variables d'accélération (`qsec`) en haut
  — Variables de transmission (`am`, `gear`) en bas

3.0.3 Contributions des variables

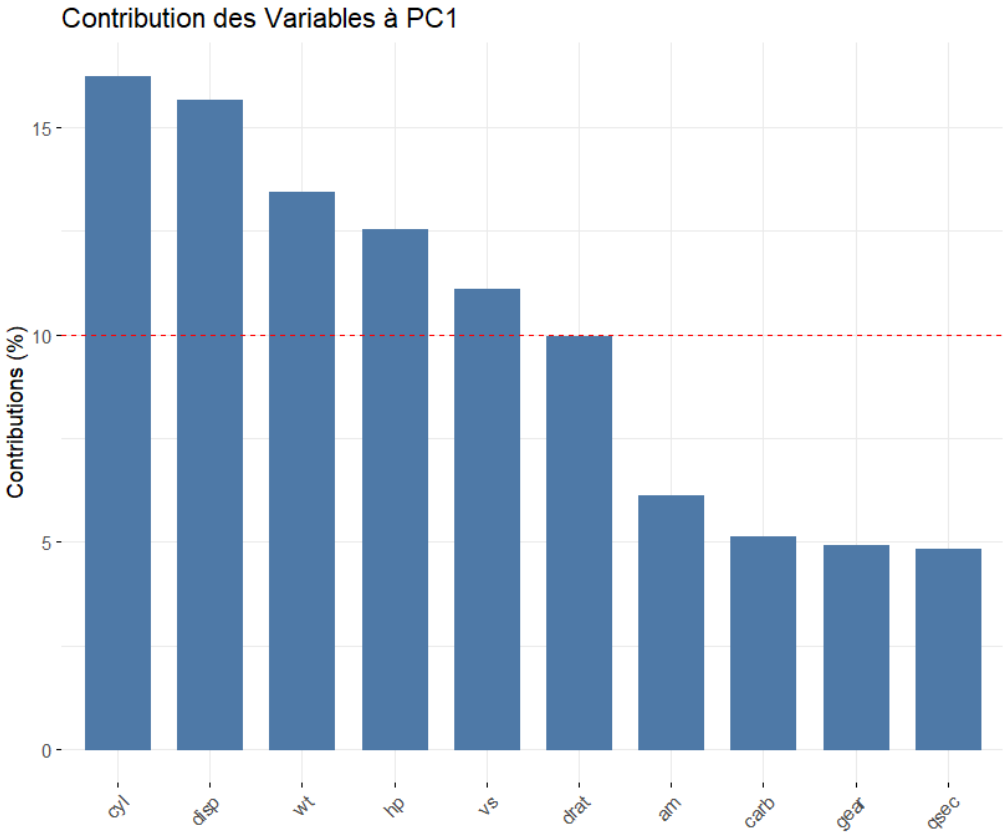

FIGURE 3.4 – Contributions à la première composante

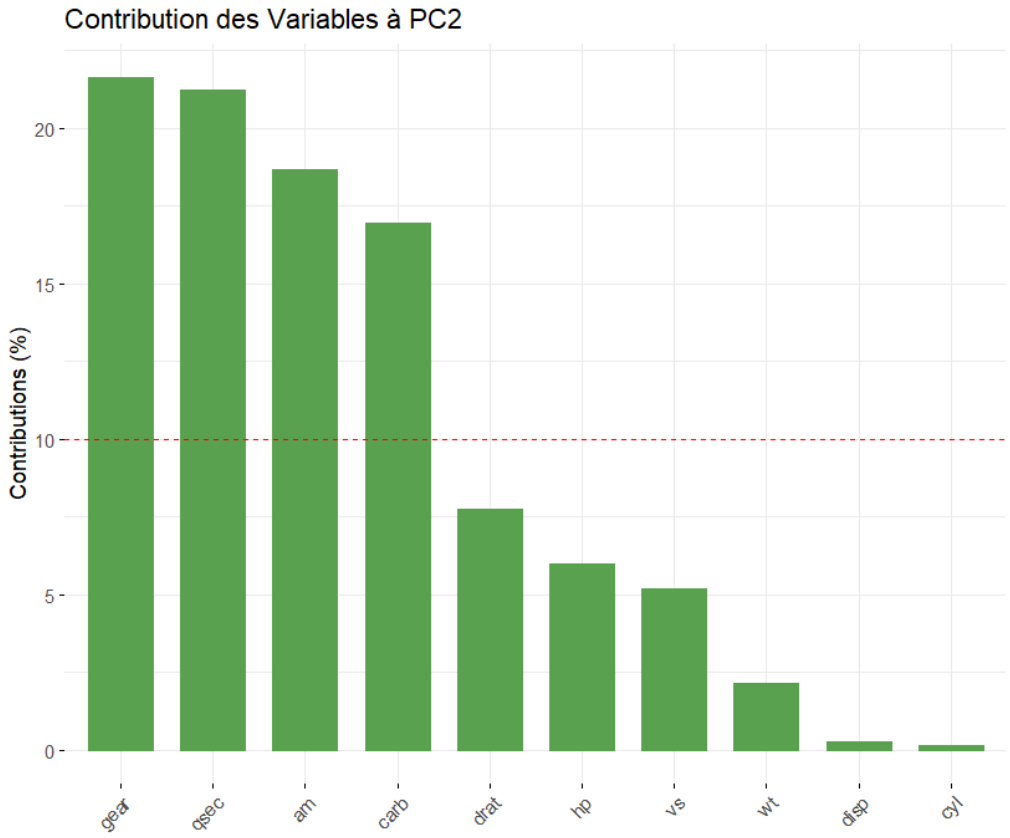

FIGURE 3.5 – Contributions à la deuxième composante

Les figures 3.4 et 3.5 montrent que :

TABLE 3.1 – Principales contributions

Composante	Variables majeures	Contribution
Dim1	`cyl`, `disp`	$> 15\%$ chacune
Dim2	`qsec`, `am`, `gear`	$> 10\%$ chacune

3.0.4 Projection des individus

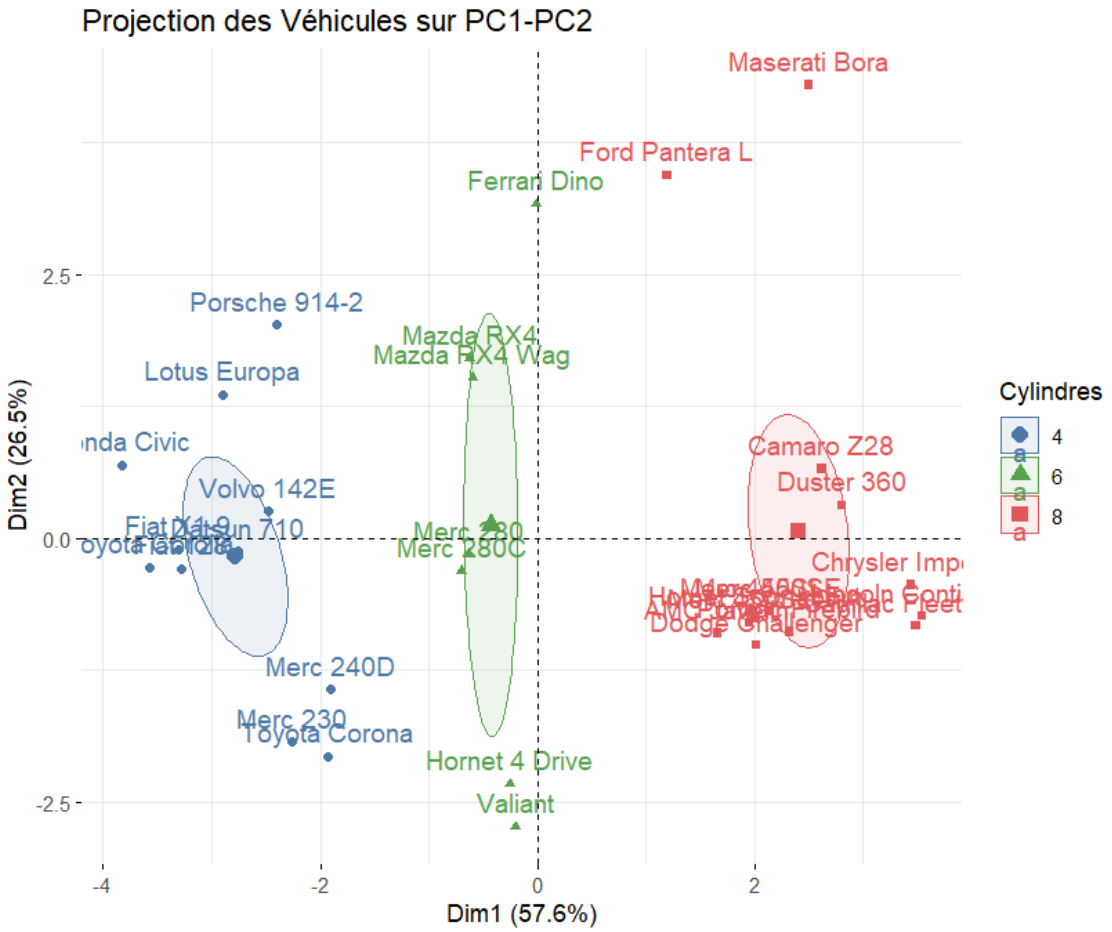

FIGURE 3.6 – Projection des véhicules colorés par cylindres

La projection (Fig. 3.9) montre une séparation nette selon le nombre de cylindres :
— **4 cylindres** : regroupés à droite (haute efficacité)
— **8 cylindres** : à gauche (haute puissance)
— **6 cylindres** : position intermédiaire

3.0.5 Régression sur les Composantes Principales

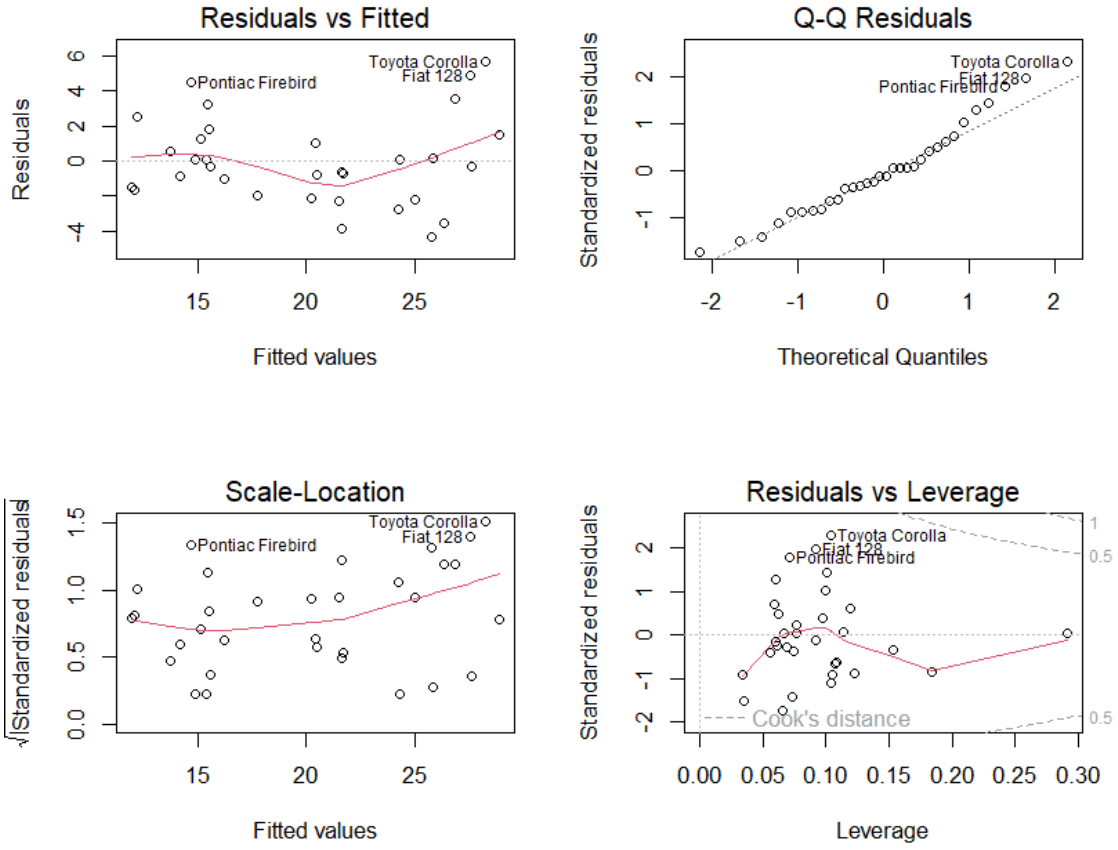

FIGURE 3.7 – Projection des véhicules colorés par cylindres

Les graphiques de diagnostic du modèle de régression multiple ajusté sur l'ensemble de données `mtcars` permettent d'évaluer la validité des hypothèses du modèle.

Résidus vs Valeurs ajustées

Ce graphique montre les résidus standardisés en fonction des valeurs prédites. Les points sont dispersés de manière relativement aléatoire autour de la ligne horizontale à zéro, sans motif clair, ce qui suggère que l'hypothèse de linéarité et d'homoscédasticité (variance constante des résidus) est raisonnablement respectée. Cependant, des points comme Pontiac Firebird, Toyota Corolla et Fiat 128 se distinguent avec des résidus élevés (positifs ou négatifs), indiquant que le modèle prédit moins bien pour ces observations.

Q-Q Résidus

Le graphique quantile-quantile compare la distribution des résidus standardisés à une distribution normale théorique. La plupart des points suivent la ligne diagonale,

ce qui indique que les résidus sont approximativement normaux, une hypothèse clé de la régression linéaire. Cependant, des écarts sont visibles aux extrémités (Toyota Corolla, Pontiac Firebird, Fiat 128), suggérant de légers écarts par rapport à la normalité pour ces observations, bien que cela reste acceptable pour un échantillon de petite taille (32 observations).

Échelle-Localisation

Ce graphique montre la racine carrée des résidus standardisés en fonction des valeurs ajustées, pour évaluer l'homoscédasticité. La dispersion des points reste relativement constante, bien que la ligne de tendance montre une légère augmentation pour les valeurs ajustées plus élevées. Les points éloignés comme Pontiac Firebird, Toyota Corolla et Fiat 128 confirment leur statut de résidus inhabituels, mais l'homoscédasticité est globalement satisfaite.

Résidus vs Effet de levier

Ce graphique identifie les observations influentes en combinant les résidus standardisés et l'effet de levier. La distance de Cook est indiquée par des lignes pointillées. Toyota Corolla et Fiat 128 ont un effet de levier modéré et des résidus élevés, mais leur distance de Cook reste faible, suggérant qu'elles ne sont pas excessivement influentes. Pontiac Firebird a également un résidu notable, mais son effet de levier est faible. Aucune observation ne semble avoir un impact disproportionné sur le modèle.

```
> summary(model_pcr)

Call:
lm(formula = mpg ~ PC1 + PC2, data = scores)

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
```

FIGURE 3.8 – summary du regression de ACP

3.1 Test d'Anova

```
Analysis of Variance Table

Model 1: mpg ~ PC1 + PC2
Model 2: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     29 195.60
2     21 147.49  8     48.11 0.8562 0.5665
```

FIGURE 3.9 – Projection des véhicules colorés par cylindres

3.2 Interprétation Statistique

3.2.1 Comparaison des R^2

— La différence entre le R^2 du modèle ACP (**0,82**) et celui de la régression multiple (**0,83**) est **négligeable** ($< 1\%$).
— Le **R^2 ajusté** favorise légèrement le modèle ACP (**0,81** contre **0,79** pour le modèle multiple), suggérant un possible *sur-ajustement* du modèle complet.

3.2.2 Test ANOVA

— p-value = **0,5665** $> 0,05 \Rightarrow$ **Aucune différence significative** entre les modèles
— Le modèle ACP montre une performance **équivalente** au modèle complet malgré :
  — Une complexité réduite (2 composantes vs 10 variables)
  — L'absence de problèmes de multicolinéarité

Implications :
— La réduction de dimension par ACP **préserve l'information pertinente** tout en simplifiant l'analyse
— Le modèle complet n'apporte pas d'amélioration significative malgré sa complexité accrue
— Le choix peut se faire sur des critères pratiques :
  — ACP pour la visualisation et l'analyse des structures latentes
  — Régression multiple pour l'interprétation physique directe (quand la multicolinéarité est contrôlée)

Chapitre 4

Comparaison des approches : régression classique vs régression sur ACP

4.1 Régression avec les variables d'origine

Avantages :
- Interprétation directe des coefficients : chaque variable garde sa signification initiale.
- Analyse précise de l'effet de chaque prédicteur sur la variable réponse.
- Mise en œuvre simple et compréhensible.

Inconvénients :
- Risque de multicolinéarité, pouvant fausser les estimations.
- Modèle instable en présence de nombreuses variables.
- Risque de surapprentissage avec un échantillon réduit.

4.2 Régression sur les composantes principales (ACP)

Avantages :
- Réduction de la dimension tout en conservant l'information essentielle.
- Suppression de la multicolinéarité grâce à l'orthogonalité des composantes.
- Modèle plus stable et robuste en ne conservant que les composantes pertinentes.

Inconvénients :
- Perte d'interprétabilité : les composantes sont des combinaisons de variables.
- Nécessite une transformation préalable des données (centrage, réduction).
- Significativité statistique ne garantit pas une compréhension intuitive.

4.3 Conclusion

Le choix dépend des objectifs :

— **Régression classique** : adaptée si l'interprétation des variables est primordiale.
— **Régression sur ACP** : préférable pour un modèle prédictif robuste en cas de corrélations élevées entre variables.

Annexe

```r
library(readr)
library(ggplot2)
library(gridExtra)
library(FactoMineR)
library(corrplot)
library(MASS)
library(pls)
```

Listing 4.1 – Chargement des packages

```r
mtcars <- read_csv("mtcars.csv")
data("mtcars")
View(mtcars)
head(mtcars)
summary(mtcars)
str(mtcars)
dim(mtcars)
```

Listing 4.2 – Chargement et exploration des données

4.3.1 Analyse exploratoire visuelle

```r
pairs(mtcars)
cor_matrix <- cor(mtcars)
corrplot(cor_matrix, method = "circle")
```

Listing 4.3 – Diagrammes de paires et corrélations

```r
colors <- c("#1f77b4", "#ff7f0e", "#2ca02c", "#d62728", "
   #9467bd",
            "#8c564b", "#e377c2", "#7f7f7f", "#bcbd22", "#17
               becf", "#aec7e8")
plot_list <- list()
for (i in 1:ncol(mtcars)) {
  plot_list[[i]] <- ggplot(mtcars, aes(x = .data[[names(
     mtcars)[i]]])) +
    geom_histogram(fill = colors[i %% length(colors) + 1],
       color = "white", bins = 10, alpha = 0.8) +
    labs(title = paste("Distribution de", names(mtcars)[i]))
       +
```

```
    theme_minimal()
}
grid.arrange(grobs = plot_list, ncol = 3)
```

Listing 4.4 – Histogrammes par variable

```
boxplot(mtcars, main = "Boxplots des variables de mtcars",
    col = "lightblue", las = 2)
```

Listing 4.5 – Boxplot global

4.3.2 Normalité des variables actives

```
vars_actives <- c("mpg", "disp", "hp", "drat", "wt", "qsec")
data_active <- mtcars[, vars_actives]
shapiro_results <- lapply(data_active, shapiro.test)
for (i in 1:length(shapiro_results)) {
  cat("Variable :", names(shapiro_results)[i], "\n")
  cat("p-value  :", round(shapiro_results[[i]]$p.value, 4), "
     \n")
  cat("Conclusion :", ifelse(shapiro_results[[i]]$p.value <
     0.05, "Non normale", "Normale"), "\n\n")
}
```

Listing 4.6 – Test de Shapiro-Wilk

4.4 Régression linéaire classique

4.4.1 Modèle complet et modèle sélectionné

```
full_model <- lm(mpg ~ cyl + disp + hp + drat + wt + qsec +
    vs + am + gear + carb, data = mtcars)
summary(full_model)

modele_selection <- stepAIC(full_model, direction = "both",
    trace = FALSE)
summary(modele_selection)
```

Listing 4.7 – Régression linéaire classique

```
par(mfrow = c(1, 2))
plot(full_model, which = 1, main = "R sidus du mod le
    complet")
plot(modele_selection, which = 1, main = "R sidus du mod le
     s lectionn ")
```

Listing 4.8 – Comparaison graphique des résidus

```r
nouveau_vehicule <- data.frame(
  mpg = 35, cyl = 4, disp = 95, hp = 80, drat = 4.1,
  wt = 2.0, qsec = 18.5, vs = 1, am = 1, gear = 5, carb = 1
)
mtcars_ajout <- rbind(mtcars, nouveau_vehicule)
modele_complet_ajout <- lm(mpg ~ ., data = mtcars_ajout)
modele_reduit_ajout <- lm(mpg ~ wt + qsec + am, data = mtcars
   _ajout)
summary(modele_complet_ajout)
summary(modele_reduit_ajout)
```

Listing 4.9 – Ajout d'un nouveau véhicule

4.5 Régression sur composantes principales (PCA)

4.5.1 Analyse en composantes principales

```r
acp_data <- scale(mtcars[, -1])
res_acp <- prcomp(acp_data, center = TRUE, scale. = TRUE)
summary(res_acp)
```

Listing 4.10 – Centrage et ACP

4.5.2 Visualisations

```r
fviz_eig(res_acp, addlabels = TRUE, barfill = "#2E9FDF")
fviz_pca_var(res_acp, col.var = "contrib", gradient.cols = c(
   "#4E79A7", "#59A14F", "#E15759"))
```

Listing 4.11 – Scree Plot et cercle des corrélations

```r
fviz_contrib(res_acp, choice = "var", axes = 1, fill = "#4
   E79A7")
fviz_contrib(res_acp, choice = "var", axes = 2, fill = "#59
   A14F")
```

Listing 4.12 – Contributions des variables

```r
fviz_pca_ind(res_acp,
             col.ind = as.factor(mtcars$cyl),
             addEllipses = TRUE,
             ellipse.type = "confidence")
```

Listing 4.13 – Projection des individus

4.5.3 Régression sur les composantes principales

```
scores <- as.data.frame(res_acp$x[, 1:2])
colnames(scores) <- c("PC1", "PC2")
scores$mpg <- mtcars$mpg
model_pcr <- lm(mpg ~ PC1 + PC2, data = scores)
summary(model_pcr)
```

Listing 4.14 – Modèle PCR

```
ggplot(scores, aes(x = PC1, y = mpg)) +
  geom_point(aes(color = as.factor(mtcars$cyl))) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "R gression: mpg ~ PC1")
```

Listing 4.15 – Visualisation de la régression

```
par(mfrow = c(2, 2))
plot(model_pcr)
```

Listing 4.16 – Diagnostic du modèle PCR

4.5.4 Comparaison des modèles

```
anova(model_pcr, full_model)
```

Listing 4.17 – Test ANOVA entre PCR et régression classique