# Auto-mpg.data

Eines d'ajuda a la presa de decisions

Oscar Galera i Alfaro
2on MEINF
UdG
17/11/2018

# Índex

# 1. Dades i variables

# 1.1. Dades i Variables

El conjunt de dades de treball conté **392 observacions i 12 variables** (de les quals V9, V10, V11 i V12 només s’utilitzaran per la representació) que són:

- V1. *mpg*: Consum en milles per gallon - **V. Quantitativa continua**
- V2. *cylinders*: Nombre de cilindres - **V. Quantitativa discreta**
- V3. *displacement*: Distància necessaria per frenar - **V. Quantitativa continua**
- V4. *horsepower*: Potència - **V. Quantitativa continua**
- V5. *weight*: Pes - **V. Quantitativa continua**
- V6. *acceleration*: Acceleració - **V. Quantitativa continua**
- V7. *model year*: Any del model - **V. Quantitativa discreta**
- V8. *model*: Nom del model - **V. Qualitativa**
- V9 *marca*: Marca - **V. Qualitativa**
- V10 *marca_1*: Marca del primer subconjunt - **V. Qualitativa**
- V11 *marca_2*: Marca del segon subconjunt - **V. Qualitativa**
- V12 *marca_3*: Marca del tercer subconjunt - **V. Qualitativa**

# 1.2. Preprocessament - Dades i Variables

Mostrar el tipus que ha inferit *R* per cada variable.

```
#Mostrar el tipus de les variables
str(dades)
```

```
## 'data.frame':    398 obs. of  12 variables:
##  $ mpg       : num  10 13 31 9 29 31.9 41.5 44.3 43.4 44 ...
##  $ cylinders : int  8 8 4 8 4 4 4 4 4 4 ...
##  $ displ     : num  307 350 119 304 90 89 98 90 90 97 ...
##  $ hp        : Factor w/ 94 levels "?","100.0","102.0",..: 43 26 78 41 67 68 72 51 51 53 ...
##  $ weight    : num  4376 4055 2720 4732 1937 ...
##  $ accel     : num  15 12 19.4 18.5 14.2 14 14.7 21.7 23.7 24.6 ...
##  $ model_year: int  70 76 82 70 76 79 80 80 80 82 ...
##  $ model     : Factor w/ 189 levels "'cuda","100",..: 57 56 161 7 151 151 151 151 89 147 ...
##  $ marca     : Factor w/ 32 levels "amc","audi","bmw",..: 8 8 8 14 32 32 32 32 32 32 ...
##  $ marca_1   : Factor w/ 11 levels "bmw","buick",..: 3 3 3 5 11 11 11 11 11 11 ...
##  $ marca_2   : Factor w/ 12 levels "audi","capri",..: 8 8 8 8 8 8 8 8 8 8 ...
##  $ marca_3   : Factor w/ 12 levels "amc","cadillac",..: 8 8 8 8 8 8 8 8 8 8 ...
```

Com es pot veure, la variable *hp* es tracta com una variable qualitativa quan se sap que només pot tenir valors numèrics, però per què?

# 1.2. Preprocessament - Dades i Variables

Es mostra la distribució de valors de *hp*

```r
#Mostrar la distribució de valors que pren la variable hp
levels(dades$hp)
```

```
##  [1] "?"      "100.0" "102.0" "103.0" "105.0" "107.0" "108.0" "110.0"
##  [9] "112.0" "113.0" "115.0" "116.0" "120.0" "122.0" "125.0" "129.0"
## [17] "130.0" "132.0" "133.0" "135.0" "137.0" "138.0" "139.0" "140.0"
## [25] "142.0" "145.0" "148.0" "149.0" "150.0" "152.0" "153.0" "155.0"
## [33] "158.0" "160.0" "165.0" "167.0" "170.0" "175.0" "180.0" "190.0"
## [41] "193.0" "198.0" "200.0" "208.0" "210.0" "215.0" "220.0" "225.0"
## [49] "230.0" "46.00" "48.00" "49.00" "52.00" "53.00" "54.00" "58.00"
## [57] "60.00" "61.00" "62.00" "63.00" "64.00" "65.00" "66.00" "67.00"
## [65] "68.00" "69.00" "70.00" "71.00" "72.00" "74.00" "75.00" "76.00"
## [73] "77.00" "78.00" "79.00" "80.00" "81.00" "82.00" "83.00" "84.00"
## [81] "85.00" "86.00" "87.00" "88.00" "89.00" "90.00" "91.00" "92.00"
## [89] "93.00" "94.00" "95.00" "96.00" "97.00" "98.00"
```

Sembla ser que la variable té valors absents, quants d'aquests valors té?

```r
#Nombre d'observacions amb valor abscent (?) en la variable hp
nrow(dades[dades$hp == "?",])
```

```
## [1] 6
```

# 1.2 Preprocessament - Dades i Variables

Algunes de les opcions per resoldre el problema de valors absents són

- Eliminar les observacions
- Assignar un valor ‘neutre’ (0, cadena buida...)
- Assignar un valor central (mitjana, mediana...)

En aquest cas, cal complir la propietat $N > 20p$ per l'anàlisi que es vol fer, i per aquest motiu s’opta per eliminar les observacions amb valor absent.

# 1.2. Preprocessament - Dades i Variables

Ara s’ha de indicar a *R* que tracti la variable com a numèrica.

```
#Convertir hp de qualitativa a quantitativa
dades$hp = as.numeric(dades$hp)
str(dades)
```

```
## 'data.frame':    392 obs. of  12 variables:
##  $ mpg        : num  10 13 31 9 29 31.9 41.5 44.3 43.4 44 ...
##  $ cylinders  : int  8 8 4 8 4 4 4 4 4 4 ...
##  $ displ      : num  307 350 119 304 90 89 98 90 90 97 ...
##  $ hp         : num  43 26 78 41 67 68 72 51 51 53 ...
##  $ weight     : num  4376 4055 2720 4732 1937 ...
##  $ accel      : num  15 12 19.4 18.5 14.2 14 14.7 21.7 23.7 24.6 ...
##  $ model_year: int  70 76 82 70 76 79 80 80 80 82 ...
##  $ model      : Factor w/ 189 levels "'cuda","100",..: 57 56 161 7 151 151 151 151 89 147 ...
##  $ marca      : Factor w/ 32 levels "amc","audi","bmw",..: 8 8 8 14 32 32 32 32 32 32 ...
##  $ marca_1    : Factor w/ 11 levels "bmw","buick",..: 3 3 3 5 11 11 11 11 11 11 ...
##  $ marca_2    : Factor w/ 12 levels "audi","capri",..: 8 8 8 8 8 8 8 8 8 8 ...
##  $ marca_3    : Factor w/ 12 levels "amc","cadillac",..: 8 8 8 8 8 8 8 8 8 8 ...
```

# 1.2. Preprocessament - Dades i Variables

La variable marca conté el nom de la marca per a cada vehicle, i les variables *marca_1*, *marca_2* i *marca_3* contenen l'agrupació de les marques en subconjunts disjunts i on cada variable té la categoria *other*, d'aquesta manera la variable *marca_1* conté *1/3* de les marques i la resta a *other* (de forma similar per *marca_2* i *marca_3*). Aquestes variables 'sintètiques' seran d'ajuda en les representacions.

```
#Categories de la variable marca_1
table(dades$marca_1)

##
##     bmw   buick   chevy    ford      hi   mazda mercury   other pontiac
##       2      17       3      48       1      12      11     273      16
## renault      vw
##       3       6
```

```
#Categories de la variable marca_2
table(dades$marca_2)

##
##           audi         capri        datsun         dodge          fiat
##              7             1            23            28             8
##          honda mercedes-benz         other       peugeot          saab
##             13             3           280             8             4
##        triumph    volkswagen
##              1            16
```

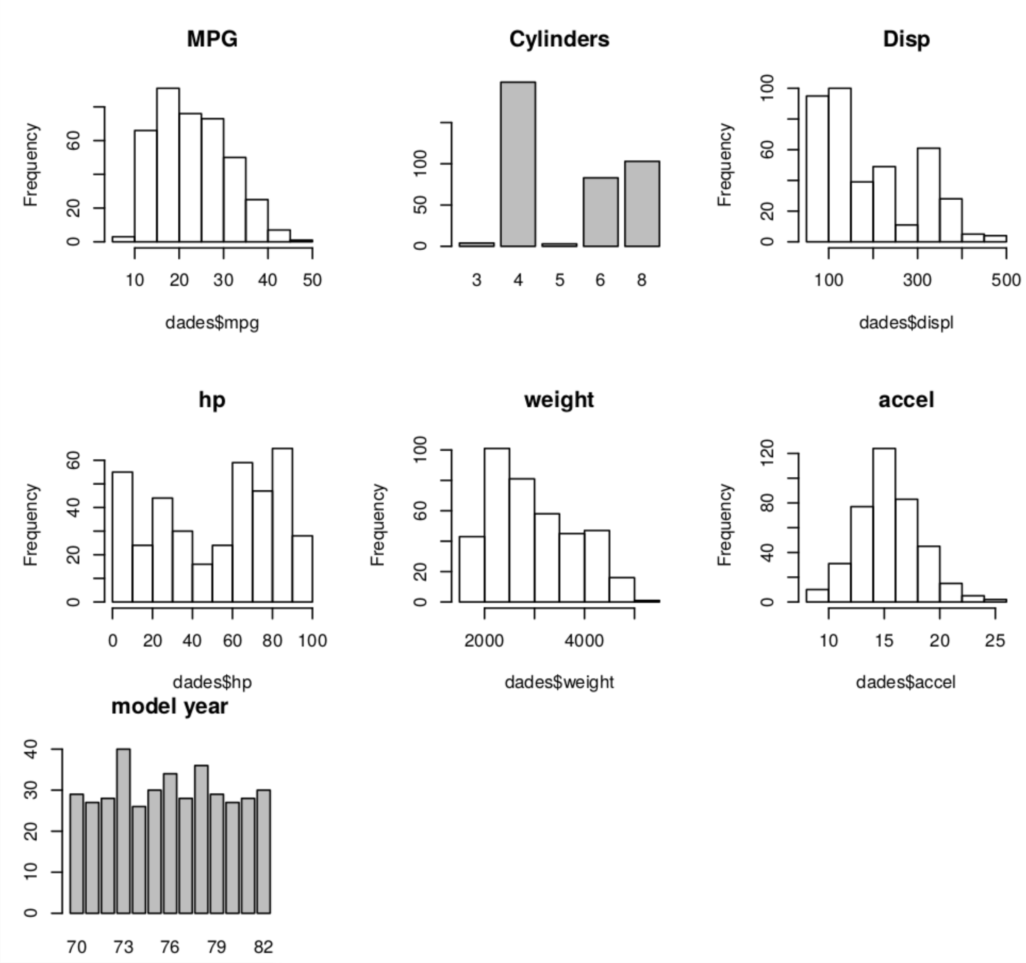
```
#Categories de la variable marca_3
table(dades$marca_3)

##
##        amc   cadillac  chevrolet   chrysler     nissan oldsmobile
##         27          2         44          6          1         10
##       opel      other   plymouth     subaru     toyota      volvo
##          4        231         31          4         26          6
```

# 2. Anàlisi descriptiu bàsic

# 2.1. Centralitat - Anàlisi descriptiu bàsic

Amb aquest anàlisi es vol veure els valors de centralitat i dispersió per cada variable. Primer de tot, es mostra la distribució de valors per cada variable a través de diagrames de barres i histogrames.

```
hNumerics <- function(){
  #Agrupar els gràfics en tripletes
  attach(mtcars)
  par(mfrow=c(2,3))

  hist(x = dades$mpg, main="MPG")
  barplot(table(dades$cylinders), main="Cylinders")
  hist(x = dades$displ, main = "Disp")
  hist(x = dades$hp, main="hp")
  hist(x = dades$weight, main="weight")
  hist(x = dades$accel, main="accel")
  barplot(table(dades$model_year), main="model year")
}
hNumerics()
```

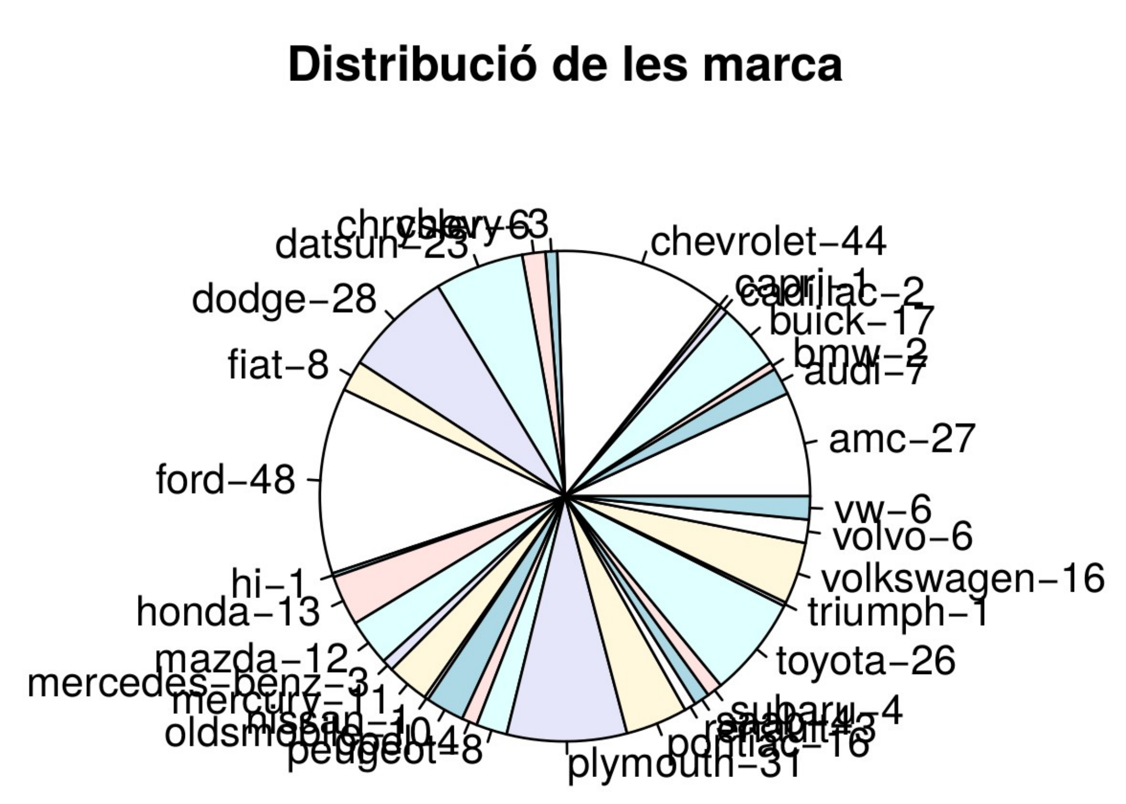



Es destaca que la variable ‘*Accel*’ té un alt grau de simetria, les variables ‘*mpg*’, ‘*displ*’ i ‘*weight*’ tenen biaix a la dreta, i en el cas de ‘*Model year*’ la seva distribució és força uniforme.

# 2.1. Centralitat - Anàlisi descriptiu bàsic

Es vol veure com es distribueixen les marques dels vehicles en les observacions.

```r
pintarPie <- function (dades, titol){
  taula <- table(dades)
  etiquetes <- paste(names(taula), "-", taula, sep="")
  pie(taula, labels = etiquetes, main=titol)
}
pintarPie(dades$marca, "Distribució de les marca")
```

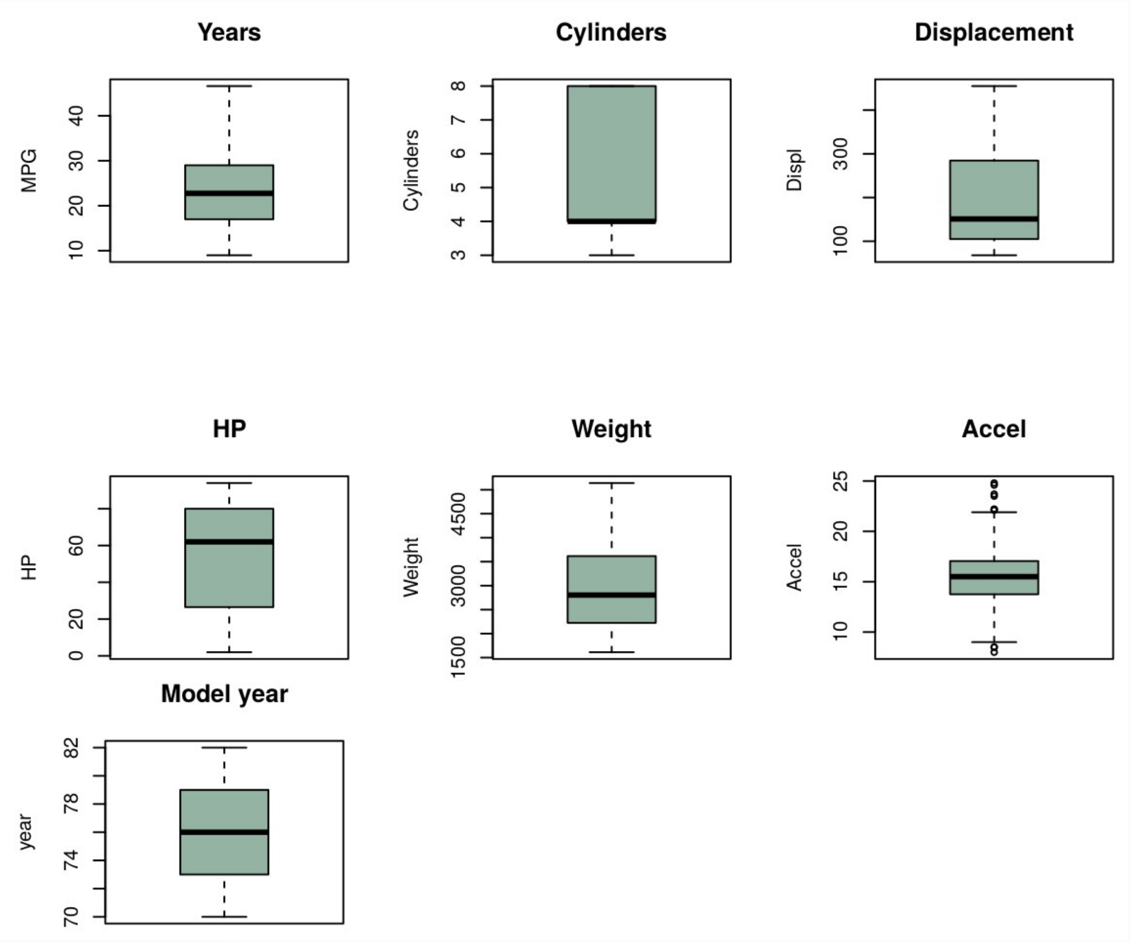



Les marques amb més representación són: *ford* (48), *chevrolet* (44) i *plymouth* (31).

# 2.2. Dispersió - Anàlisi descriptiu bàsic

Per veure la dispersió en les variables, és molt útil utilitzar diagrames de caixa.

```
bplotNumerics <- function(){
  attach(mtcars)
  par(mfrow=c(2,3))

  #Bloxplots
  boxplot(dades$mpg, col=rgb(0.3, 0.5, 0.4, 0.6), ylab="MPG", main="Years")
  boxplot(dades$cylinders, col=rgb(0.3, 0.5, 0.4, 0.6), ylab="Cylinders", main="Cylinders")
  boxplot(dades$displ, col=rgb(0.3, 0.5, 0.4, 0.6), ylab="Displ", main="Displacement")
  boxplot(dades$hp, col=rgb(0.3, 0.5, 0.4, 0.6), ylab="HP", main="HP")
  boxplot(dades$weight, col=rgb(0.3, 0.5, 0.4, 0.6), ylab="Weight", main="Weight")
  boxplot(dades$accel, col=rgb(0.3, 0.5, 0.4, 0.6), ylab="Accel", main="Accel")
  boxplot(dades$model_year, col=rgb(0.3, 0.5, 0.4, 0.6), ylab="year", main="Model year")
}
bplotNumerics()
```




Gràcies aquests diagrames, és fàcil veure com hi ha un clar biaix a la dreta en la variable '*cylinders*', i també com hi ha simetria en *'Accel'* i *'Model year'.*

També destacar que hi ha dades atípiques inferiors (per sota de **1*Q* – 1.5*RIQ***) i superiors (per sobre de **3*Q* + 1.5*RIQ***) per la variable 'Accel'

# 2.3. Resum - Anàlisi descriptiu bàsic

Per acabar aquest resum bàsic de les variables, es poden calcular diferents estadístics de centralitat i dispersió robustos (les dades atípiques influeixen molt en el seu valor) i no robustos. En aquest cas es calcularan:

| Robustos – Centralitat | No Robustos – Centralitat | Robustos – Dispersió | No Robustos – Dispersió |
|---|---|---|---|
| Mediana | Mitjana | *RIC* | Desviació estàndard |
| Mitjana Retallada | | *MAD* | |
| Mitjana *Winsoritzada* | | | |

Table 1: Variable mpg

| Estadistic | Valor |
|---|---|
| **C. robustos** | |
| Mediana | 22.750000 |
| Mit. Retallada | 22.750000 |
| Mit. Winsoritzada | 22.750000 |
| **C. NO robustos** | |
| Mitjana | 23.445918 |
| **D. robustos** | |
| RIC | 12.000000 |
| MAD | 46.600000 |
| **D. NO robustos** | |
| D. Estàndard | 7.805008 |

Table 2: Variable disp

| Estadistic | Valor |
|---|---|
| **C. robustos** | |
| Mediana | 151.000 |
| Mit. Retallada | 151.000 |
| Mit. Winsoritzada | 151.000 |
| **C. NO robustos** | |
| Mitjana | 194.412 |
| **D. robustos** | |
| RIC | 170.750 |
| MAD | 455.000 |
| **D. NO robustos** | |
| D. Estàndard | 104.644 |

Table 3: Variable hp

| Estadistic | Valor |
|---|---|
| **C. robustos** | |
| Mediana | 62.00000 |
| Mit. Retallada | 62.00000 |
| Mit. Winsoritzada | 62.00000 |
| **C. NO robustos** | |
| Mitjana | 52.16071 |
| **D. robustos** | |
| RIC | 53.25000 |
| MAD | 94.00000 |
| **D. NO robustos** | |
| D. Estàndard | 29.49805 |

Table 4: Variable weight

| Estadistic | Valor |
|---|---|
| **C. robustos** | |
| Mediana | 2803.5000 |
| Mit. Retallada | 2803.5000 |
| Mit. Winsoritzada | 2803.5000 |
| **C. NO robustos** | |
| Mitjana | 2977.5842 |
| **D. robustos** | |
| RIC | 1389.5000 |
| MAD | 5140.0000 |
| **D. NO robustos** | |
| D. Estàndard | 849.4026 |

Table 5: Variable accel

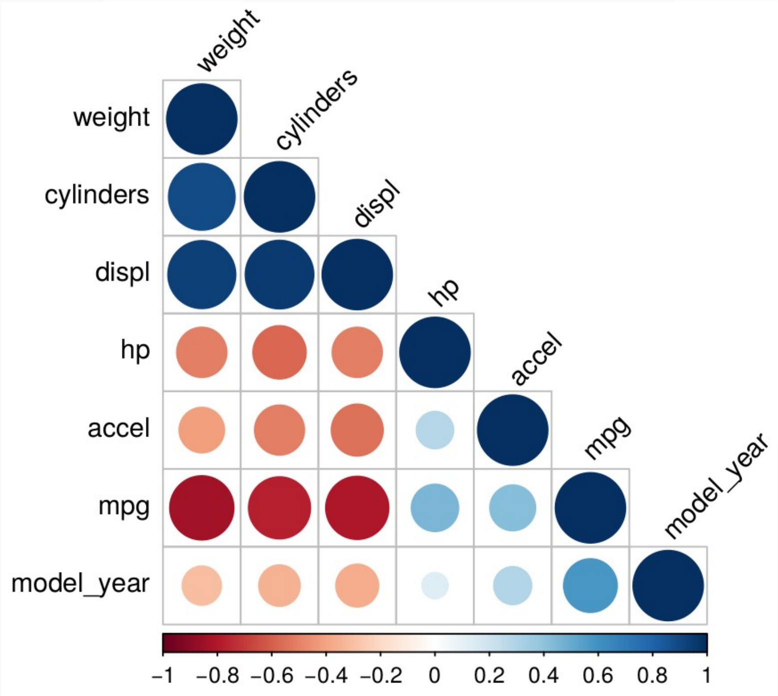
| Estadistic | Valor |
|---|---|
| **C. robustos** | |
| Mediana | 15.500000 |
| Mit. Retallada | 15.500000 |
| Mit. Winsoritzada | 15.500000 |
| **C. NO robustos** | |
| Mitjana | 15.541327 |
| **D. robustos** | |
| RIC | 3.250000 |
| MAD | 24.800000 |
| **D. NO robustos** | |
| D. Estàndard | 2.758864 |

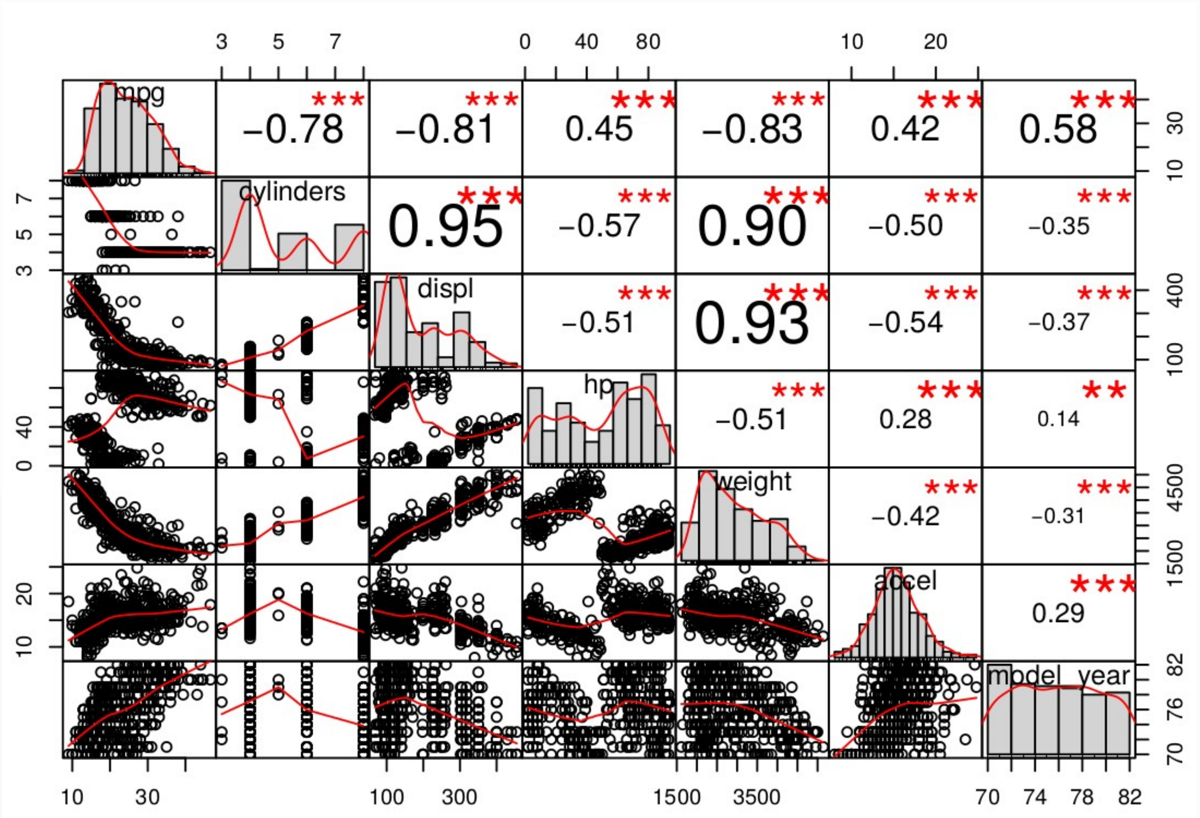
# 3. Anàlisi exploratori amb **ACP**

# 3.1. *ACP* – Anàlisi exploratori

L'anàlisi de components principals o *ACP* permet descriure un conjunt de dades, resumint-lo i reduir la seva dimensionalitat. En aquest cas, s'usarà per interpretar la relació que hi ha entre les variables.

Primer es vol veure la relació que hi ha entre les variables, per això es fa la matriu de correlacions.



# *3.1. ACP* – Anàlisi exploratori

Ara es vol veure si val la pena fer l'ACP o no, per això cal comprovar que com a mínim hi hagui dues variable amb una variància diferent per a que es puguin resumir en els eixos vectorials que es volen buscar.

$$H_0 : s_1^2 = s_2^2 \wedge s_3^2 ... \wedge s_k^2$$

$$H_1 : s_i^2 \neq s_j^2 | parella(i, j)$$

```
library(psych)
cortest.bartlett(cor.mat, n=100 )
```

```
## $chisq
## [1] 680.7449
##
## $p.value
## [1] 1.547461e-130
##
## $df
## [1] 21
```

Com que s'obté un *p-value* pròxim a *0 (*1,547 x $10^{-130}$) es pot rebutjar la hipòtesi nul·la (homocedasticitat) amb un nivell de confiança del 95%

# *3.1. ACP* – Anàlisi exploratori

També es pot calcular el coeficient de *Kaiser-Mayer-Olkin*, el qual descriu heterocedasticitat per valors superiors a 0,5

```
library(psych)
KMO(cor.mat)
```

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = cor.mat)
## Overall MSA =  0.8
```

Com que en aquest cas s'ha obtingut un valor de 0,8 es reafirma.

# *3.1. ACP* – Anàlisi exploratori

Per calcular l'ACP es pot utilitzar la funció *PCA* de la llibreria *FactoMineR.*

```
library(FactoMineR)
res <- PCA(dades.PCA, scale.unit=TRUE, ncp=7, graph=FALSE)
```

Com que s'han utilitzat set variables, hi ha set eixos vectorials, que són:

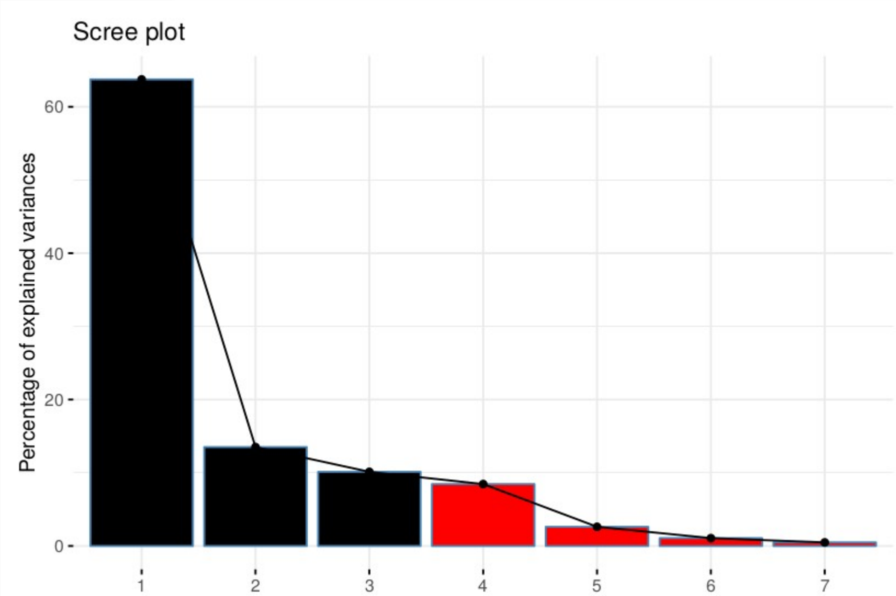
```
res$eig
```

```
##        eigenvalue percentage of variance cumulative percentage of variance
## comp 1 4.46178005             63.7397149                          63.73971
## comp 2 0.94578664             13.5112377                          77.25095
## comp 3 0.70881007             10.1258581                          87.37681
## comp 4 0.59180203              8.4543147                          95.83113
## comp 5 0.18316342              2.6166203                          98.44775
## comp 6 0.07471212              1.0673160                          99.51506
## comp 7 0.03394567              0.4849382                         100.00000
```

# ***3.1. ACP*** **– Anàlisi exploratori**

Quants d’aquests eixos hem d’utilitzar? Per això es pot representar en un diagrama de barres la quantitat de variabilitat que captura cada variable.

```
library("factoextra")
fviz_screeplot(res, ncp=7, barfill=c(rep(1,3),rep(2,4)))
```




Tot i que només amb els **dos primers eixos** ja s’obté un **77,25% de la inèrcia**, per tenir més marge en els posteriors exemples, **s’ha decidit utilitzar els tres primers eixos** que acumulen el **87,37% de la inèrcia total**.

# *3.1. ACP* – Anàlisi exploratori

Ara es vol saber quins són les variables que més han contribuït en l’elecció d’aquests eixos.

```
res[["var"]][["contrib"]]
```

```
##                 Dim.1      Dim.2       Dim.3      Dim.4      Dim.5
## mpg         17.971051  4.1289649  5.46536743  0.6485902 58.2654241
## cylinders   19.973778  2.4770942  0.06248921  2.5388709 20.6006934
## displ       20.484216  0.8996707  0.03251121  6.5770159  9.2294562
## hp           8.555233 22.3649622  2.06135113 66.1509514  0.1105271
## weight      19.212769  2.4323281  2.30976764 10.0299317  0.8491002
## accel        8.102681  2.0046966 85.29490851  1.8815834  1.8451240
## model_year   5.700273 65.6922833  4.77360487 12.1730567  9.0996750
##                  Dim.6      Dim.7
## mpg         12.8340210  0.6865811
## cylinders   33.6956323 20.6514423
## displ        0.3971941 62.3799362
## hp           0.5061954  0.2507799
## weight      49.9390296 15.2270741
## accel        0.3106581  0.5603486
## model_year   2.3172694  0.2438378
```

Pel primer eix aquestes variables són: ‘*displ*’, ‘*cylinders*’, ‘*weight*’ i ‘*mpg*’, pel segon eix: ‘*model_year*’ i ‘*hp*’ i així successivament.

# *3.1. ACP* – Anàlisi exploratori

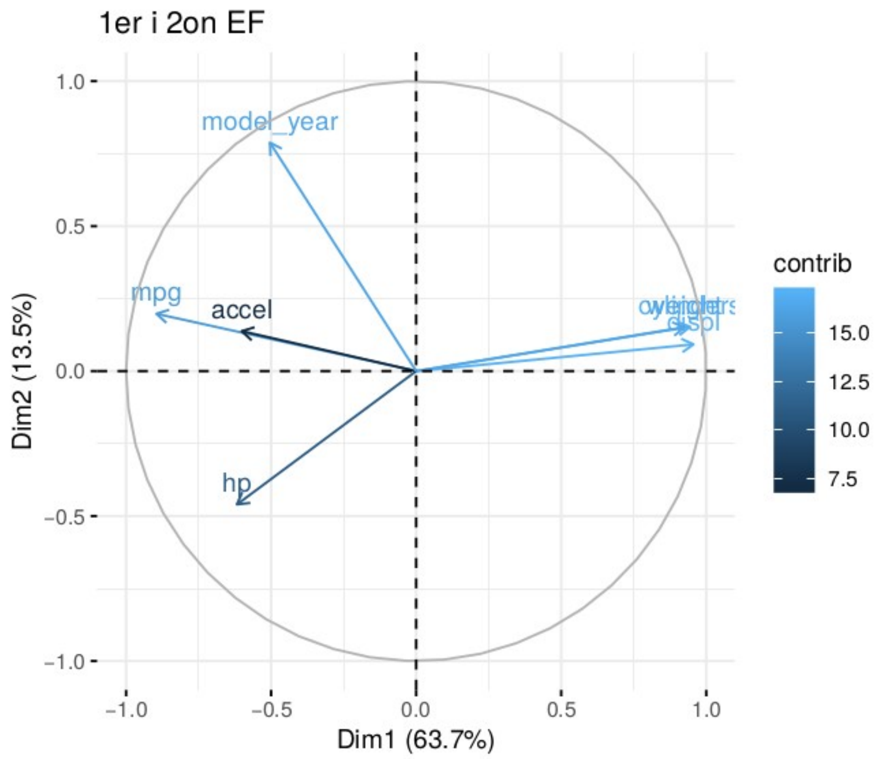
Ara es vol conèixer com de ben representades estan les variables en els eixos trobats, això ve reflexat en el *cos²*.

```
res[["var"]][["cos2"]]

##                Dim.1       Dim.2        Dim.3      Dim.4        Dim.5
## mpg        0.8018288 0.039051199 0.0387390745 0.00383837 0.1067209428
## cylinders  0.8911860 0.023428026 0.0004429298 0.01502509 0.0377329344
## displ      0.9139606 0.008508965 0.0002304427 0.03892291 0.0169049876
## hp         0.3817157 0.211524825 0.0146110643 0.39148267 0.0002024452
## weight     0.8572315 0.023004635 0.0163718656 0.05935734 0.0015552410
## accel      0.3615238 0.018960152 0.6045788985 0.01113525 0.0033795922
## model_year 0.2543336 0.621308839 0.0338357919 0.07204040 0.0166672758
##                   Dim.6        Dim.7
## mpg        0.0095885694 2.330646e-04
## cylinders  0.0251747217 7.010271e-03
## displ      0.0002967521 2.117529e-02
## hp         0.0003781893 8.512893e-05
## weight     0.0373105084 5.168933e-03
## accel      0.0002320993 1.902141e-04
## model_year 0.0017312811 8.277239e-05
```

Pel primer eix aquestes variables són: ‘*displ*’, ‘*cylinders*’, ‘*weight*’ i ‘*mpg*’, pel segon eix: ‘*model_year*’ i ‘*hp*’ i així successivament.

# 3.2. *Primer i segon eix, ACP* – Anàlisi exploratori

Es projecten les variables sobre els dos primers eixos factorials per buscar relacions entre aquestes




En el primer eix es mostren els vehicles més pesats (*weight*) que solen tenir una major cilindrada (*cylinder*), i que tendeixen a consumir més combustible per milla recorreguda (*mpg*). També es pot veure com a mesura que augmenta el *pes* i *cilindrada* dels cotxes, aquest necessiten més espai per frenar (*disp*).

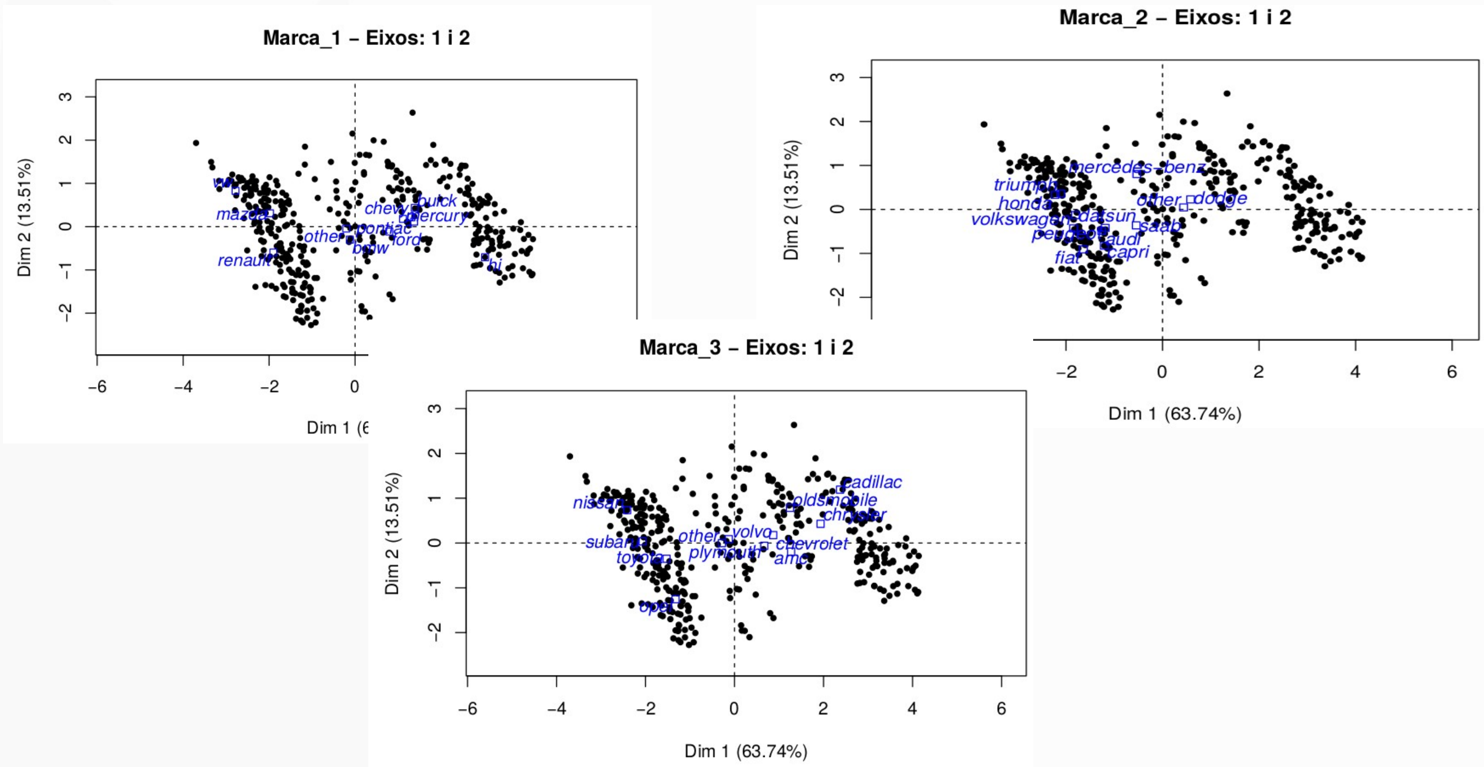
**Relació entre weight, cylinders, disp, accel i mpg.**

El segon eix explica que a mesura que augmenta l'any del model (*model_year)*, hi ha tendència a disminuir la potència (*hp)*, disminuir el consum (*mpg)* i augmentar l'acceleració (accel).

**Relació entre model_year, hp, mpg i accel.**

# 3.2. *Primer i segon eix, ACP* – Anàlisi exploratori

Ara es volen projectar les marques sobre els eixos factorials per veure les tendències que segueixen cada una d'aquestes marques. Per això s'utilitzaran les variables *marca_1, marca_2* i *marca_3*

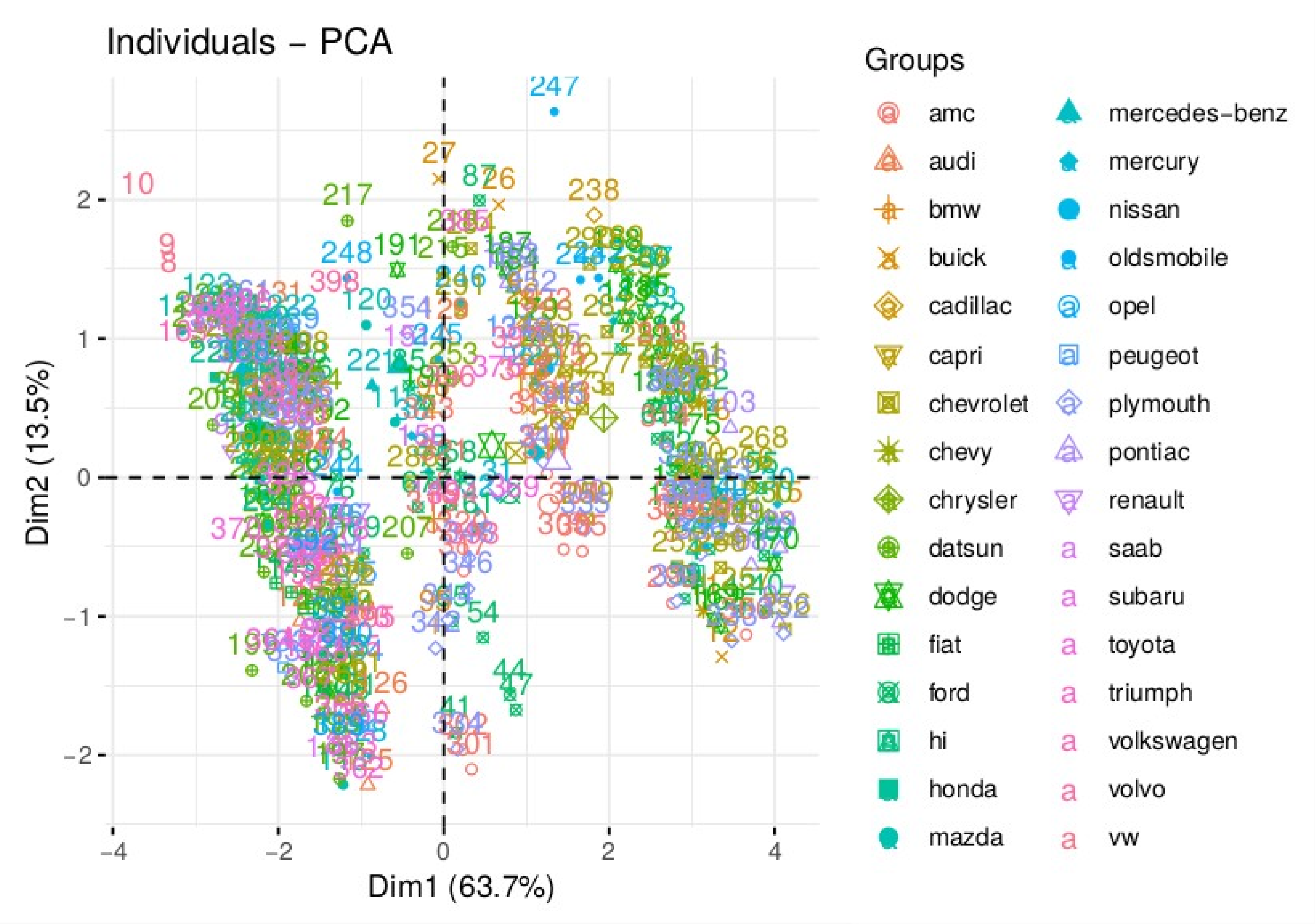



<u>Primer eix factorial</u> es pot deduir que les marques ***vw***, ***mazda***, ***honda***, ***triumph***, ***nissan***, ***renault*** i ***subaru*** tendien a fabricar **vehicles poc pesants**, **amb pocs cilindres** i amb un **consum baix**. Les marques ***volvo***, ***bmw*** i ***ford*** buscaven un cert **equilibri entre pes i consum**, i finalment, fabricants com ***hi***, ***chrysler*** o ***cadillac*** tendien a treballar de forma contraria i **fabricaven cotxes més pesants i que consumien més**.

<u>Segon eix factorial</u>, es pot veure com les marques ***vw*** i ***nissan*** tenen representació de **veichles mes moderns** (*model year*) mentre que els vehicles de les marques ***fiat***, ***capri*** o ***opel*** **són més antics**.

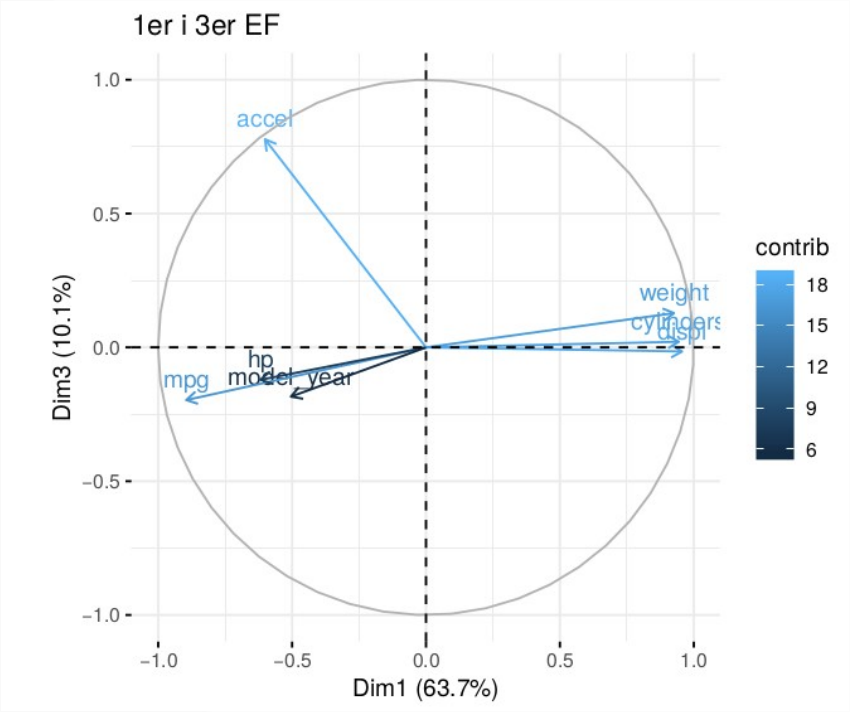
# ***3.2. Primer i segon eix, ACP*** **– Anàlisi exploratori**

*Projecció de tots els vehicles sobre els dos primers eixos factorials, utilitzant diferents símbols i colors en funció de la marca del vehicle.*



# *3.3. Primer i tercer eix, ACP* – Anàlisi exploratori

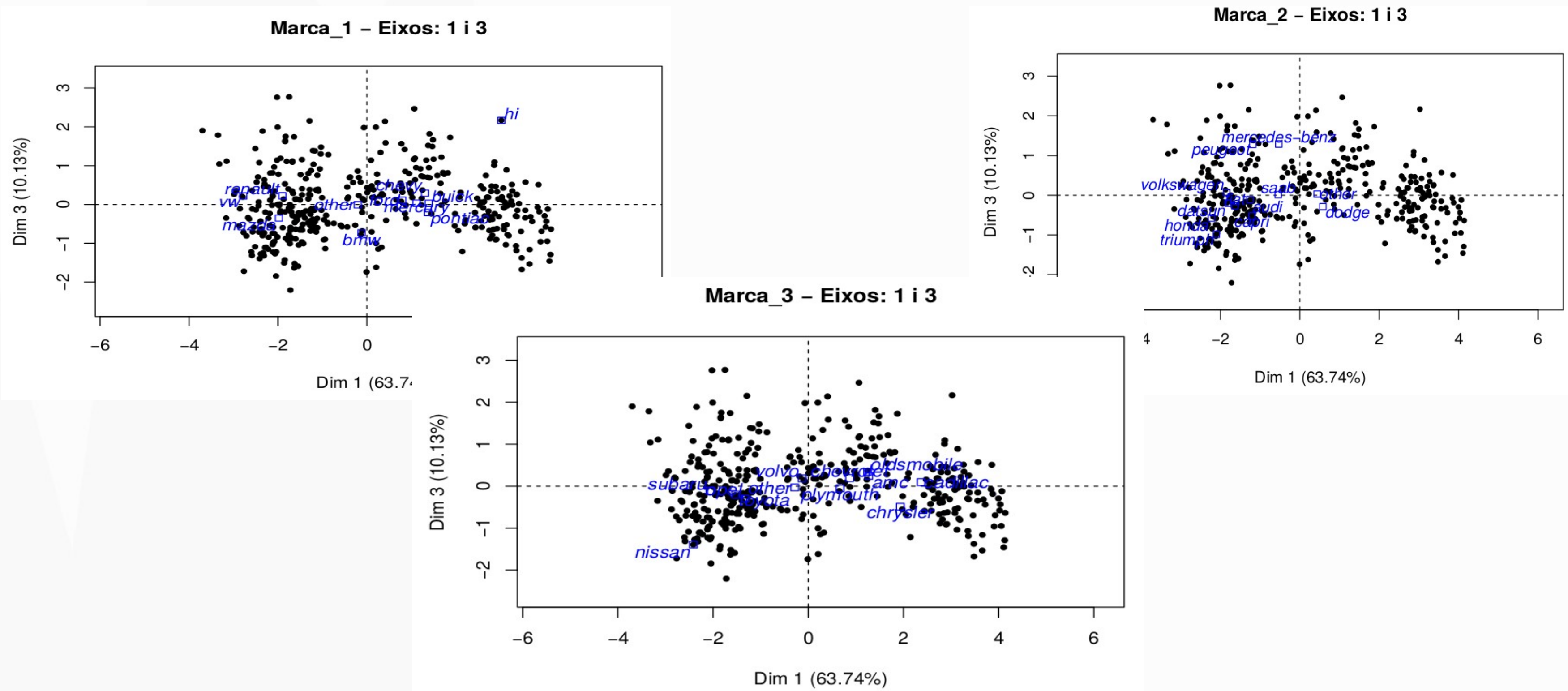
Com que la variable '***accel***' no estan ben representada en cap dels dos primers eixos, però si ho estan millor en el tercer eix. Es torna a dibuixar la circumferència però aquesta vegada amb el primer i tercer eix.




Amb aquest gràfic es pot intuir que a mesura que **augmenta l'acceleració** ('accel') **augmenta el consum**, ja que es poden recorre menys milles per galó ('*mpg*')

# *3.3. Primer i tercer eix, ACP* – Anàlisi exploratori

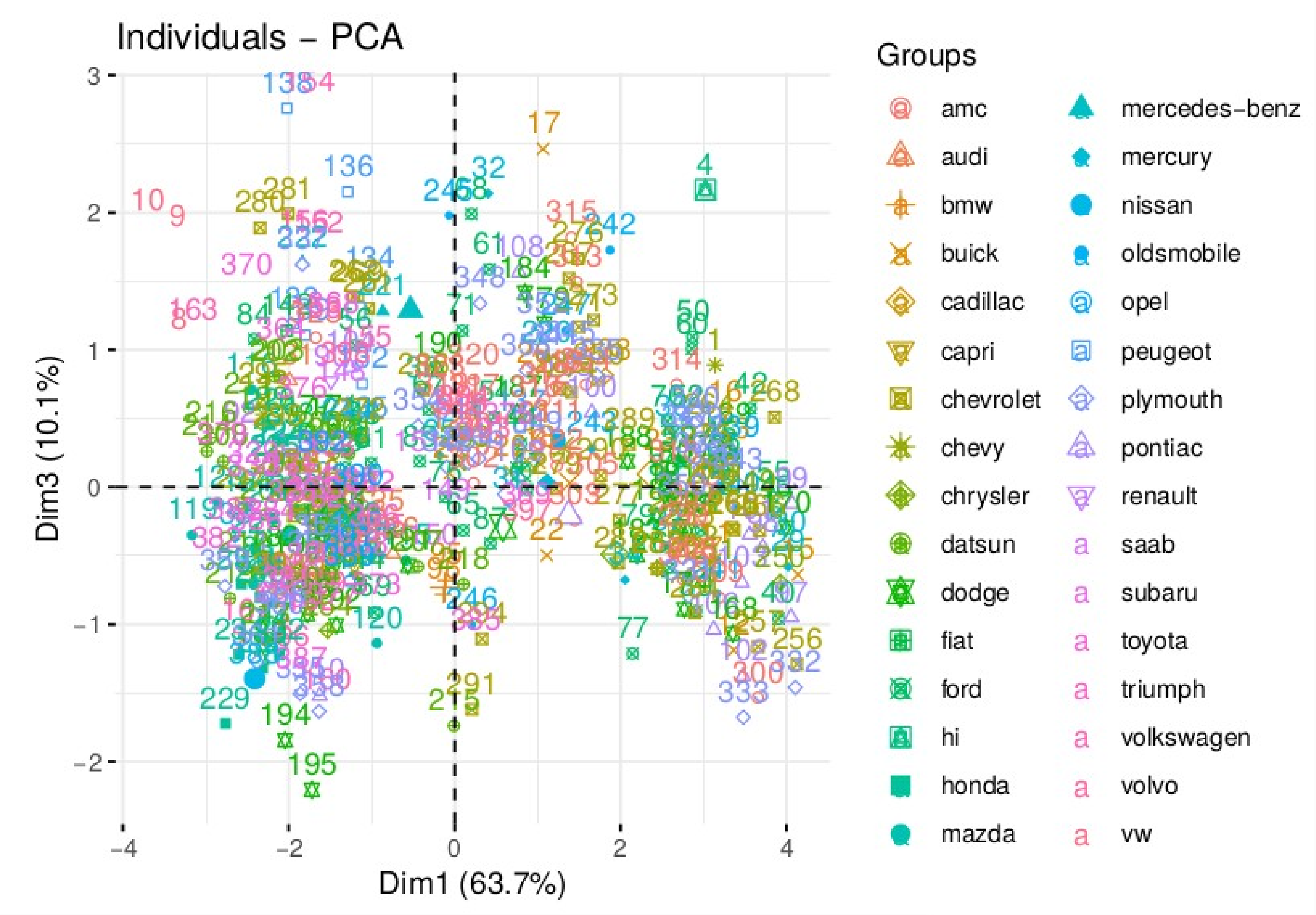
Ara es volen projectar les marques sobre el primer i el tercer eix factorial per veure les tendències que segueixen cada una d'aquestes marques. Per això es tornen a utilitzar les variables *marca_1, marca_2* i *marca_3*




Amb aquests tres gràfics es pot veure com les marques ***hi***, ***mercedes-benz*** i ***peugeot*** **tendien a fabricar cotxes amb més acceleració** (*accel*), mentre que les marques ***triumph***, ***bmw*** o ***nissan*** **fabricaven cotxes amb menys acceleració**.

# *3.3. Primer i tercer eix, ACP* – Anàlisi exploratori

*Projecció de tots els vehicles sobre el primer i el tercer eix factorial, utilitzant diferents símbols i colors en funció de la marca del vehicle.*



# 4. Anàlisi predictiu amb **RL**

# 4.1. RL - Anàlisi predictiu

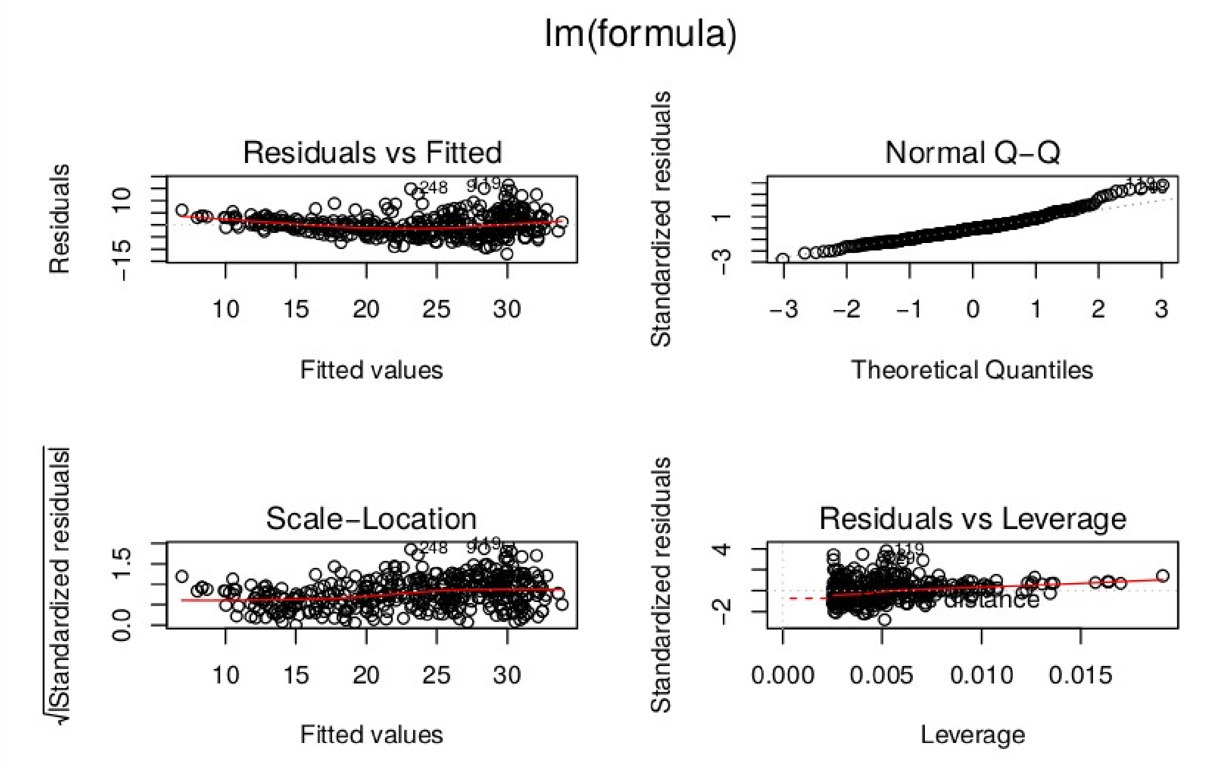
Per fer aquest anàlisi, es farà una regressió lineal simple per predir el valor de la variable MPG, per això es començarà el model amb la variable més correlacionada amb aquesta (weight) i s'aniran afegint variables mentre això millori el model.

Per aquest models de regressió lineal s'han de complir les següents propietats:

1) Independència (les observacions de la mostra s'han extret amb independència entre elles).
2) Linealitat (que la variable resposta es pot ajustar linealment a partir de les variables explicatives).
3) Normalitat en els errors (els errors segueixen una distribució normal).
4) Homocedasticitat (variància és constant).

# 4.2. mpg ~ weight - Anàlisi predictiu

Per comprovar el tercer i quart punt es poden mostrar les gràfiques referents als residus en la relació mpg ~ weight.




El gràfic *Residuals* vs *Fitted* mostra les variàncies dels residus (punt 4) mentre que el gràfic *Normal Q-Q* mostra la distribució dels errors (Punt 3).

# 4.2. mpg ~ weight - Anàlisi predictiu

Per assegurar que es compleix l'homocedasticitat, es pot aplicar un test *Breuch Pagane* on les hipòtesis són:

$$H_0 : Hi\ ha\ homocedasticitat$$

$$H_1 : No\ hi\ ha\ homocedasticitat$$

```r
#install.packages("lmtest")
library(lmtest)
bptest(mpg~weight,data=dades)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  mpg ~ weight
## BP = 22.456, df = 1, p-value = 2.15e-06
```

Com que el *p-value* del test és molt inferior al nivell de significació (α = 0.05), amb un nivell de confiança del 95% **es pot rebutjar la hipòtesi nul·la i considerar que no hi ha homocedasticitat** (hi ha heterocedasticitat).

# 4.2. mpg ~ weight - Anàlisi predictiu

Per comprovar la normalitat en els errors, es pot aplicar un test de Shapiro Wilk i un altre d' Anderson-Daling on en tots dos casos, les hipòtesis plantejades són:

$$H_0: \textit{Hi ha normalitat en els residus}$$

$$H_1: \textit{No hi ha normalitat en els residus}$$

```r
res <- lm(mpg~weight, data = dades)
#Shapiro Wilk test
shapiro.test(residuals(res))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(res)
## W = 0.96938, p-value = 2.525e-07
```

```r
#Anderson-Daling test
#install.packages("nortest")
library(nortest)
ad.test(residuals(res))
```

```
##
##  Anderson-Darling normality test
##
## data:  residuals(res)
## A = 2.6013, p-value = 1.428e-06
```

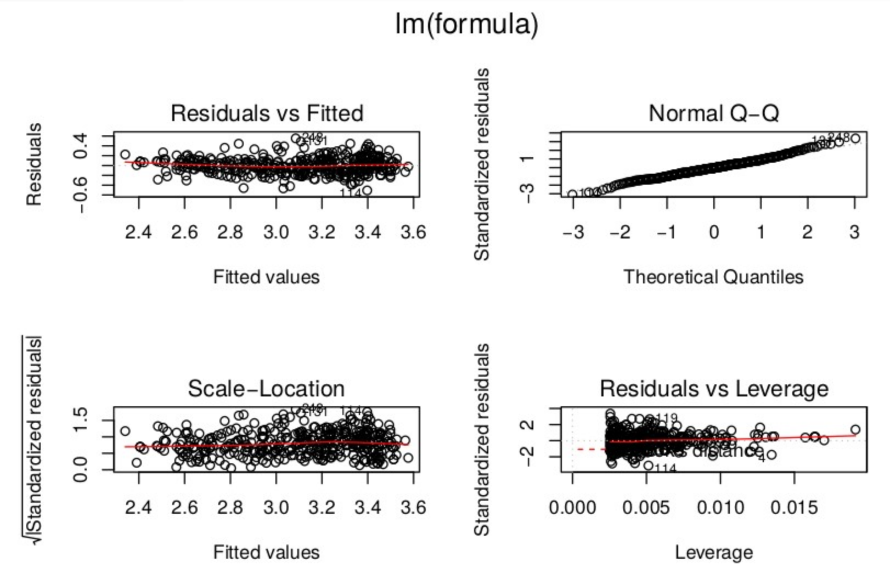
En el test de *Shapiro Wilk* s'ha obtingut un *p-value* de $2{,}038 \cdot 10^{-6}$ i en el test d'*Anderson-Daling* un *p-value* de $2.169 \cdot 10^{-6}$, i per tant, en tots dos casos **es pot descartar la hipòtesi nul·la amb un nivell de confiança superior al 95%, i es pot dir que els residus no segueixen una distribució normal.**

# 4.2. mpg ~ weight - Anàlisi predictiu

Com que s'ha demostrat que en aquest cas no es compleix ni el 3er ni el 4rt punt anunciats, la regressió que s'obtindria seria de baixa qualitat i no es podria donar un interval de confiança que acotés la predicció.

Observant la variable *mpg* es pot veure que hi ha un biaix a la dreta, per aquest motiu pot ser que aplicant una correcció logarítmica sobre la variable resposta s'aconsegueixi normalitat i constància en els residus.

```
#Correcció logarítmica en la variable mpg
dades$log_mpg <- log(dades$mpg)
```



# 4.2. mpg ~ weight - Anàlisi predictiu

Es tornen a aplicar el testos de Breuch Pagane i Shapiro-Wilk

```
#Homocedasticitat
bptest(log_mpg~weight,data=dades)

##
##  studentized Breusch-Pagan test
##
## data:  log_mpg ~ weight
## BP = 2.6409, df = 1, p-value = 0.1041
```

```
#Normalitat en els errors
shapiro.test(residuals(res_log))

##
##  Shapiro-Wilk normality test
##
## data:  residuals(res_log)
## W = 0.99375, p-value = 0.1059
```

Ara s'obté un *p-value* de 0,1041 pel test de *Breuch Pagane*, i per tant, **amb un nivell de confiança del 95% no hi ha suficients evidències per rebutjar la hipòtesi nul·la, i per això s'assumeix homocedasticitat**. Per altre banda, en el test de *Shapiro Wilk* s'obté un *p-value* de 0,1059, i per tant, **amb un nivell de confiança del 95% no es pot rebutjar la hipòtesi nul·la, i per tant, s'assumeix que els errors segueixen una distribució normal.**

# 4.2. mpg ~ weight - Anàlisi predictiu

Ara es consulten els coeficients del model de regressió, i es mostra en un *scatter plot* les amb les dues variables del model i la recta de regressió

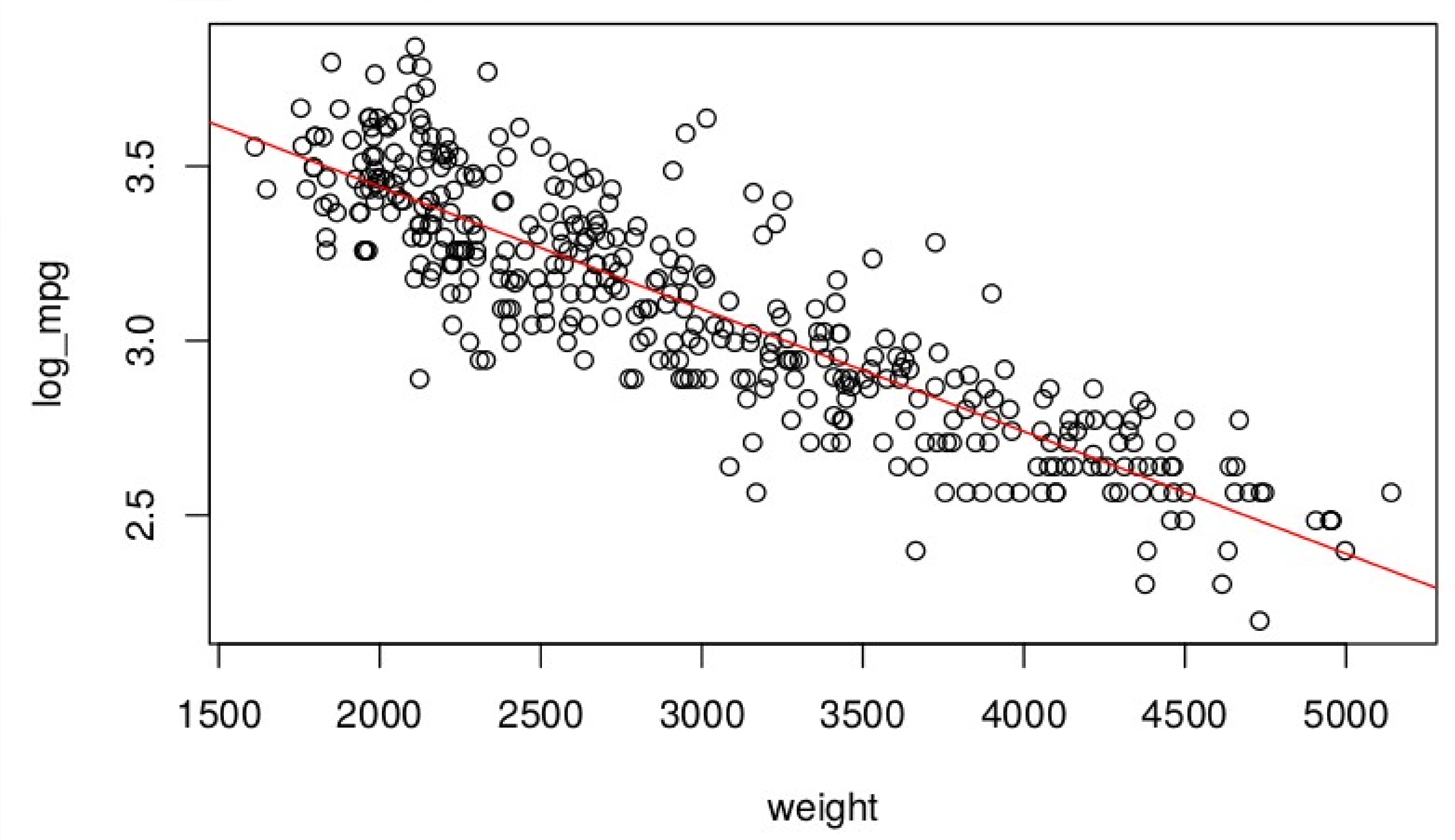
```
#Dades resultants de la regressió
summary(res_log)
```

```
##
## Call:
## lm(formula = log_mpg ~ weight, data = dades)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.50716 -0.09966 -0.00621  0.09973  0.55239
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.142e+00  3.031e-02  136.66   <2e-16 ***
## weight      -3.505e-04  9.790e-06  -35.81   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1644 on 390 degrees of freedom
## Multiple R-squared:  0.7668, Adjusted R-squared:  0.7662
## F-statistic:  1282 on 1 and 390 DF,  p-value: < 2.2e-16
```

$$\hat{y} = 4.142 + -3.505 \cdot 10^{-4} \cdot x_1 + \epsilon$$

$$mpg = 4.142 + -3.505 \cdot 10^{-4} \cdot weight + \epsilon$$

# 4.2. mpg ~ weight - Anàlisi predictiu



# 4.3. mpg ~ weight + accel - Anàlisi predictiu

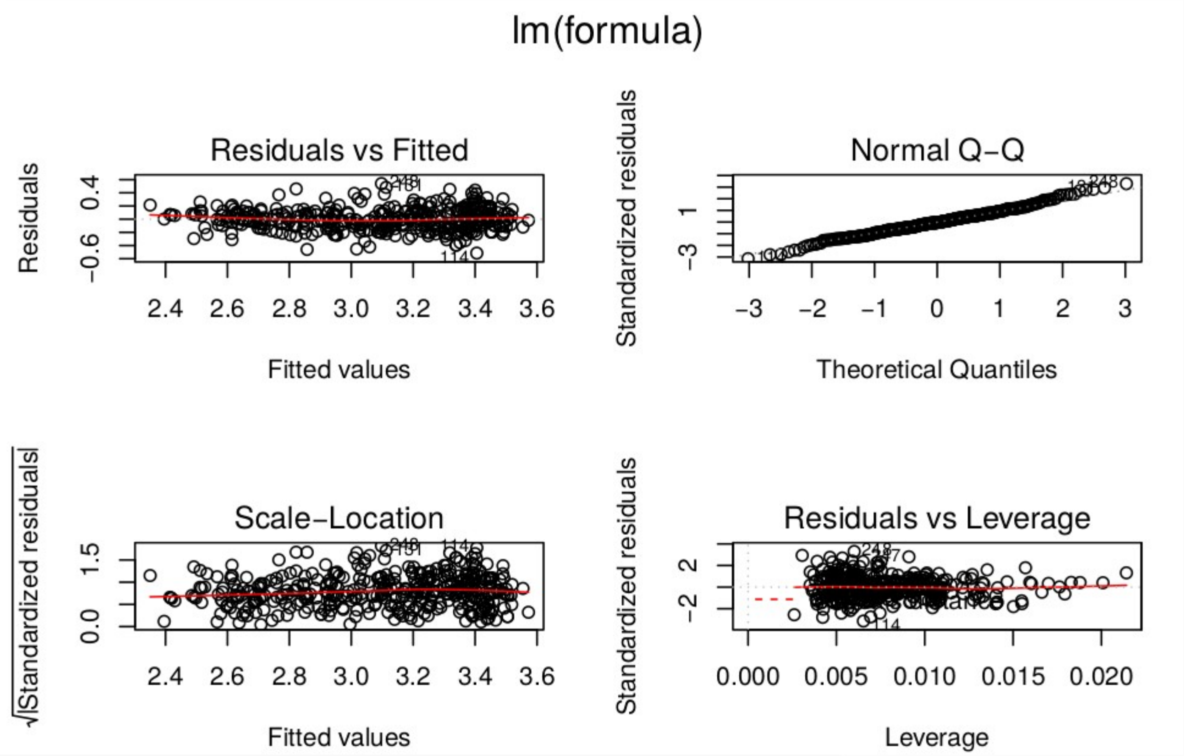
Amb la intenció de millorar el model, es poden afegir noves variables (passant d'una regressió lineal simple a una regressió lineal múltiple) per això, **s'ha de provar d'afegir aquelles variables poc relacionades amb les que ja inclou el model (que intentin explicar el que encara no està explicat) i que estiguin el màxim de correlacionades amb la variable resposta.**

Es prova d'afegir la **variable *hp*** i es tornen a fer els testos de homocedasticitat i normalitat d'errors (cal tornar a fer-los per assegurar que no es perd qualitat amb les noves variables).

# 4.3. mpg ~ weight + accel - Anàlisi predictiu

Es prova d'afegir la variable *hp* i es tornen a fer els testos de homocedasticitat i normalitat d'errors (cal tornar a fer-los per assegurar que no es perd qualitat amb les noves variables).

```r
#Nova regressió
res_log <- lm(log_mpg ~ weight + hp, dades)

#Gràfics
parametres_reg(log_mpg ~ weight + hp, dades)
```



# 4.3. mpg ~ weight + accel - Anàlisi predictiu

```
#Homocedasticitat
bptest(res_log)

##
##  studentized Breusch-Pagan test
##
## data:  res_log
## BP = 3.1615, df = 2, p-value = 0.2058
#Normalitat en els errors
shapiro.test(residuals(res_log))

##
##  Shapiro-Wilk normality test
##
## data:  residuals(res_log)
## W = 0.99327, p-value = 0.07722
```

En el test de constància en els errors s'obté un *p-value* de 0,2058, i per tant, **es continua assumint homocedasticitat,** i en el test de normalitat dels residus s'obté un *p-value* de 0,07722 **i es segueix assumint normalitat en els residus.**

$$\hat{y} = 4.098 + -3.431 \cdot 10^{-4} \cdot x_1 + 4.2 \cdot 10^{-4} \cdot x_2 + \epsilon$$

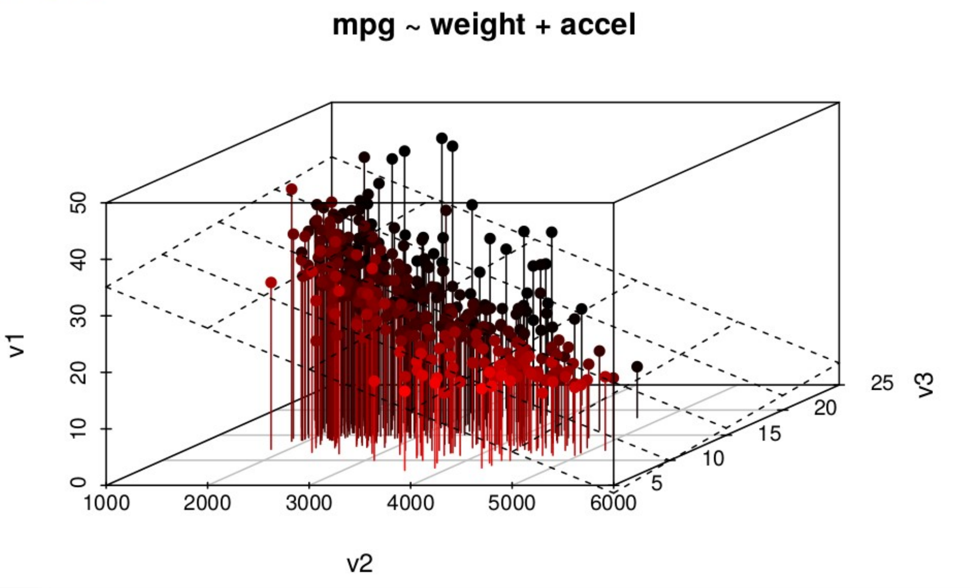
$$mpg = 4.098 + -3.431 \cdot 10^{-4} \cdot weight + 4.2 \cdot 10^{-4} \cdot hp + \epsilon$$

# 4.3. mpg ~ weight + accel - Anàlisi predictiu

```r
summary(res_log)
```

```
##
## Call:
## lm(formula = log_mpg ~ weight + hp, data = dades)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.51505 -0.10357 -0.00648  0.09882  0.54000
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.098e+00  4.573e-02  89.622   <2e-16 ***
## weight      -3.431e-04  1.136e-05 -30.206   <2e-16 ***
## hp           4.200e-04  3.271e-04   1.284      0.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1643 on 389 degrees of freedom
## Multiple R-squared:  0.7678, Adjusted R-squared:  0.7666
## F-statistic:   643 on 2 and 389 DF,  p-value: < 2.2e-16
```

```r
library(ggplot2)
plot_reg_3d <- function(v1,v2,v3){
  library(scatterplot3d)
  attach(mtcars)
  s3d <-scatterplot3d(v2,v3,v1, pch=16, highlight.3d=TRUE,
    type="h", main="mpg ~ weight + accel")
  fit <- lm(v1 ~ v2+v3)
  s3d$plane3d(fit)
}
plot_reg_3d(dades$mpg, dades$weight, dades$accel)
```



# 4.3. mpg ~ weight + accel - Anàlisi predictiu

Aquest valor ara és de 0,7666 (en vers al 0,7662 obtingut només amb la variable *weight*), per veure si aquesta millora és significativa, es pot aplicar un test *ANOVA* on les hipòtesis són:

$$H_0 : \beta_2 = 0$$

$$H_1 : \beta_2 \neq 0$$

```
anova(lm(mpg~weight, dades), lm(mpg~weight+hp, dades))
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ weight
## Model 2: mpg ~ weight + hp
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    390 7321.2
## 2    389 7297.7  1    23.577 1.2568  0.263
```

Com que el *p-value* és 0,263, **amb un 95% de confiança no es pot rebutjar la hipòtesi nul·la, i per tant, es considerà que la contribució de la variable *hp* al model és 0.**

$$mpg = 4.142 + -3.505 \cdot 10^{-4} \cdot weight + \epsilon$$

# 4.4. Eficiència del model, RL - Anàlisi predictiu

Ara es vol comprovar de forma empírica com de bé, el model prediu el consum a partir del pes i per això s'utilitza la tècnica del *k-fold cross validation*. Com que s'aconsella utilitzar un 80% de les dades (314 observacions) per construir el model i el 20% per testejar-lo (78 observacions), el valor del k serà 5.

```
library(DAAG)

#k-fold cross validation amb k=5
kfold = cv.lm(data=dades, lm(log_mpg ~ weight, dades), m=5, printit = FALSE)
```

Es calcula l'error comés en la validació

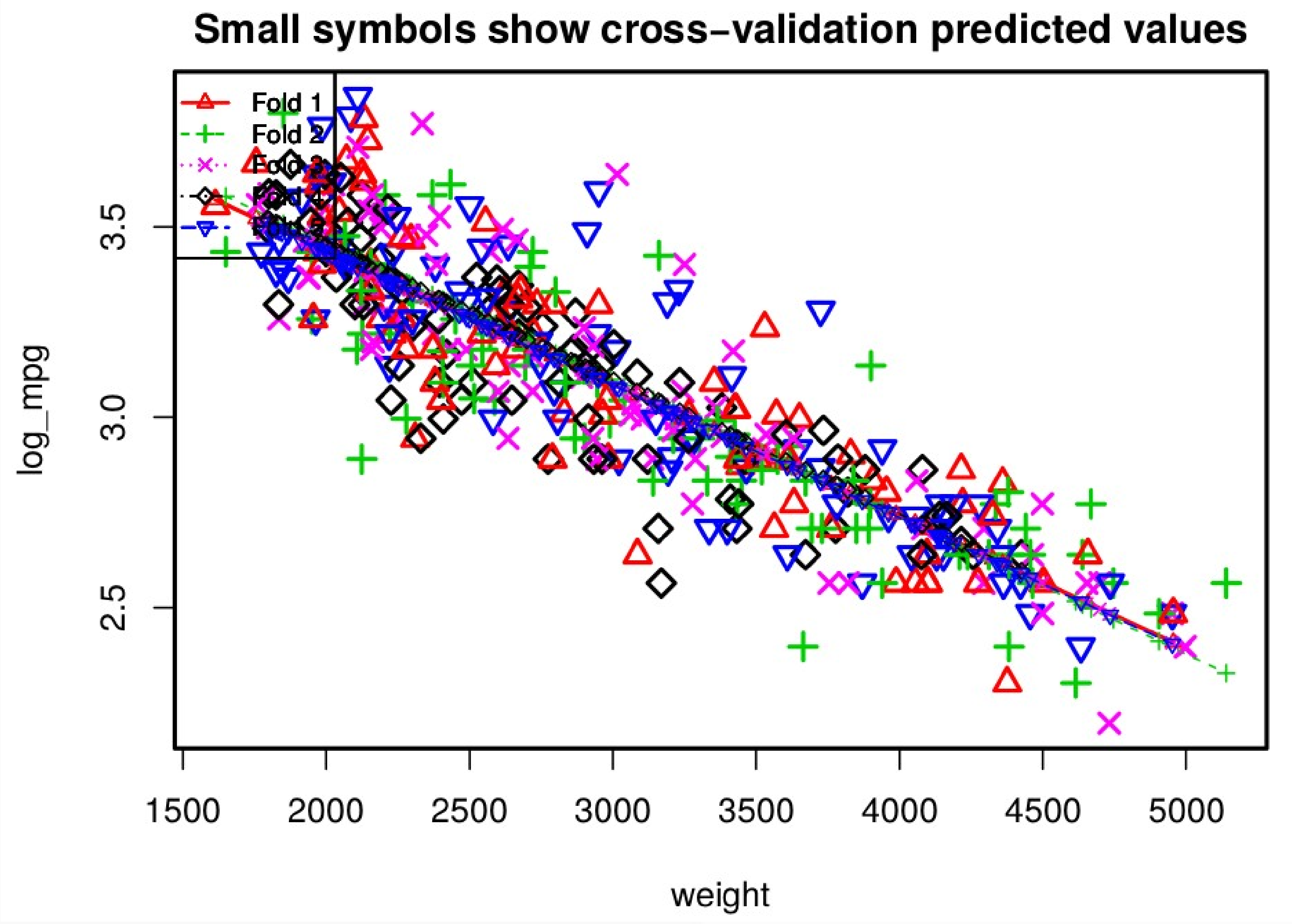
```
errors_log = mean(sqrt(sum(kfold$log_mpg-kfold$cvpred)^2))

#En unitats mpg
errors = exp(errors_log)

#mitja de la variable mpg
mitja_real = mean(dades$mpg)

#Error relatiu
(mitja_real-errors)/mitja_real
```

```
## [1] 0.950176
```

```
remove(kfold); remove(errors_log); remove(errors); remove(mitja_real)
```

L'error comés en aquesta predicció ronda 1,179 milles per galó, tenint en compte que la mitja de totes les observacions és 23,45, i per tant s'ha comés un error relatiu del 0.95.

# 4.4. Eficiència del model, RL - Anàlisi predictiu



# Gràcies

# Preguntes?