3.Estadística Descriptiva Bivariant

Estadística
Grau en Matemàtiques

Josep A. Sanchez
Dept. Estadística i I.O.(UPC)

UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

Observem, per a cada individu de la mostra dues variables:

X	Y
x_1	y_1
x_2	y_2
x_3	y_3
x_4	y_4
:	:
x_n	y_n

Objectiu: Esbrinar quin tipus de relació existeix entre elles si n'hi ha alguna.

Si existeix relació, vol dir que:

Depenent del valor d'una de les variables canvia la distribució de l'altra

Estadística Descriptiva Bivariant

Considerem les tres situacions possibles:

- Categòrica-Categòrica (ej. Tipus de cotxe-Pais de Fabricació)
- Categòrica-Numèrica (ej. Grau que es cursa-Nota d'Estadística)
- Numèrica-Numèrica (ej. Número de paraules en un text-Mida del fitxer de word en Mb)

Resums Numèrics (cat-cat)

Taules de contingència:

Tabulació de la mostra d'acord a dues variables categòriques creuades. Suposem que la variable A té m nivells i la variable B en té k

	B_1	B_2	B_3	$\ldots$	B_k
A_1	n_{11}	n_{12}	n_{13}	$\ldots$	n_{1k}
A_2	n_{21}	n_{22}	n_{23}	$\ldots$	n_{2k}
A_3	n_{31}	n_{32}	n_{33}	$\ldots$	n_{3k}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ldots$	$\vdots$
A_m	n_{m1}	n_{m2}	n_{m3}	$\ldots$	n_{mk}

on $\sum_{i=1}^{m}\sum_{j=1}^{k} n_{ij} = n$

Resums Numèrics (cat-cat)

Taules de contingència:

- Marginal per files: $n_{i.} = \sum_{j=1}^{k} n_{ij}$
- Marginal per columnes: $n_{.j} = \sum_{i=1}^{m} n_{ij}$

	B_1	B_2	B_3	$\ldots$	B_k	
A_1	n_{11}	n_{12}	n_{13}	$\ldots$	n_{1k}	$n_{1.}$
A_2	n_{21}	n_{22}	n_{23}	$\ldots$	n_{2k}	$n_{2.}$
A_3	n_{31}	n_{32}	n_{33}	$\ldots$	n_{3k}	$n_{3.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ldots$	$\vdots$	$\vdots$
A_m	n_{m1}	n_{m2}	n_{m3}	$\ldots$	n_{mk}	$n_{4.}$
	$n_{.1}$	$n_{.2}$	$n_{.3}$	$\ldots$	$n_{.k}$	n

Observació: La marginal per files és la taula de freqüències de la variable A i la marginal per columnes és la taula de freqüències de la variable B

Resums Numèrics (cat-cat)

La representació d'una taula de contingència en termes relatius es pot calcular de tres maneres:

- Freqüència relativa global: $f_{ij} = \frac{n_{ij}}{n}$.
  - És la freqüència de individus de la mostra que tenen simultàniament les categories A_i i B_j (freqüència conjunta).
- Freqüència relativa per files: $f_{ij} = \frac{n_{ij}}{n_{i.}}$
  - És la freqüència de individus de la mostra amb la categoria A_i que tenen la categoria B_j (freqüència condicional).
- Freqüència relativa per columnes: $f_{ij} = \frac{n_{ij}}{n_{.j}}$
  - És la freqüència de individus de la mostra amb la categoria B_j que tenen la categoria A_i (freqüència condicional).

Resums Numèrics (cat-cat)

```
table(Centre,System)
```

```
##          System
## Centre    Linux Mac-OS Windows
##   ETSEIB     12     11       8
##   FIB        18     13      18
##   FME        16      8      16
```

```
prop.table(table(Centre,System))
```

```
##          System
## Centre          Linux      Mac-OS     Windows
##   ETSEIB 0.10000000 0.09166667 0.06666667
##   FIB    0.15000000 0.10833333 0.15000000
##   FME    0.13333333 0.06666667 0.13333333
```

Resums Numèrics (cat-cat)

```
prop.table(table(Centre,System),margin=1)
```

```
##          System
## Centre          Linux    Mac-OS   Windows
##   ETSEIB 0.3870968 0.3548387 0.2580645
##   FIB    0.3673469 0.2653061 0.3673469
##   FME    0.4000000 0.2000000 0.4000000
```

```
prop.table(table(Centre,System),margin=2)
```

```
##          System
## Centre          Linux    Mac-OS   Windows
##   ETSEIB 0.2608696 0.3437500 0.1904762
##   FIB    0.3913043 0.4062500 0.4285714
##   FME    0.3478261 0.2500000 0.3809524
```

Representacions Gràfiques (cat-cat)

Una variable categòrica es pot representar gràficament amb un diagrama de barres.

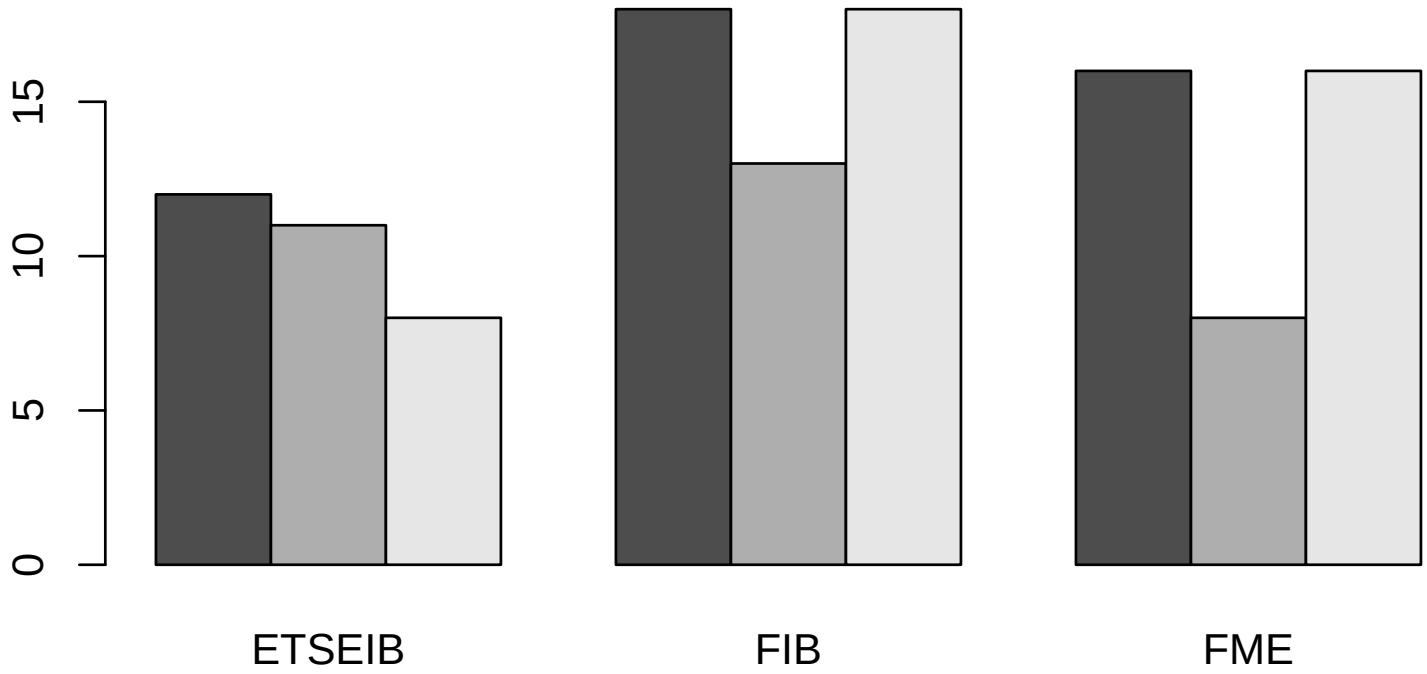
L'altre variable categòrica indueix una segmentació/partició de la mostra.

Combinant ambdues consideracions, la representació gràfica bivariant pot ser:

- **Diagrama de barres agrupades**: la segmentació es representa amb barres juntes
- **Diagrama de barres apilades**: la segmentació dona lloc a una divisió en les barres

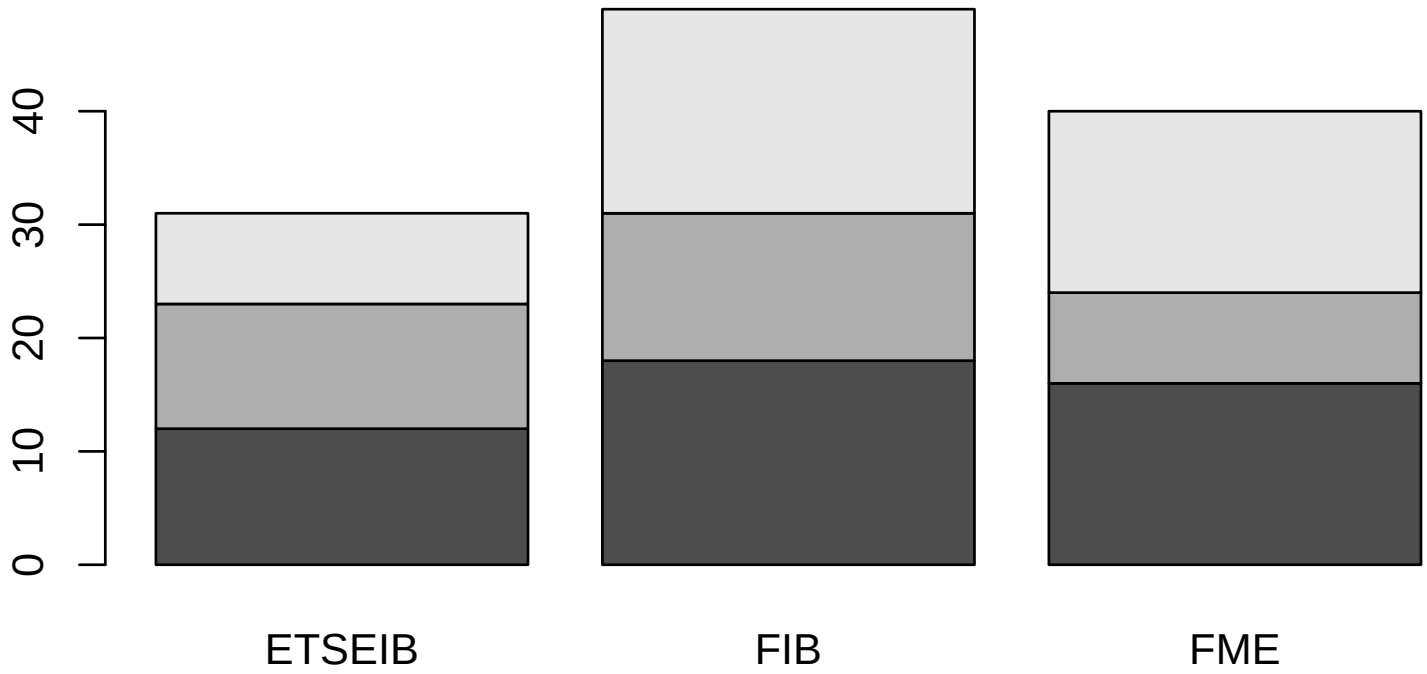
Representacions Gràfiques (cat-cat)

```
barplot(table(System,Centre),beside=T)
```



Representacions Gràfiques (cat-cat)

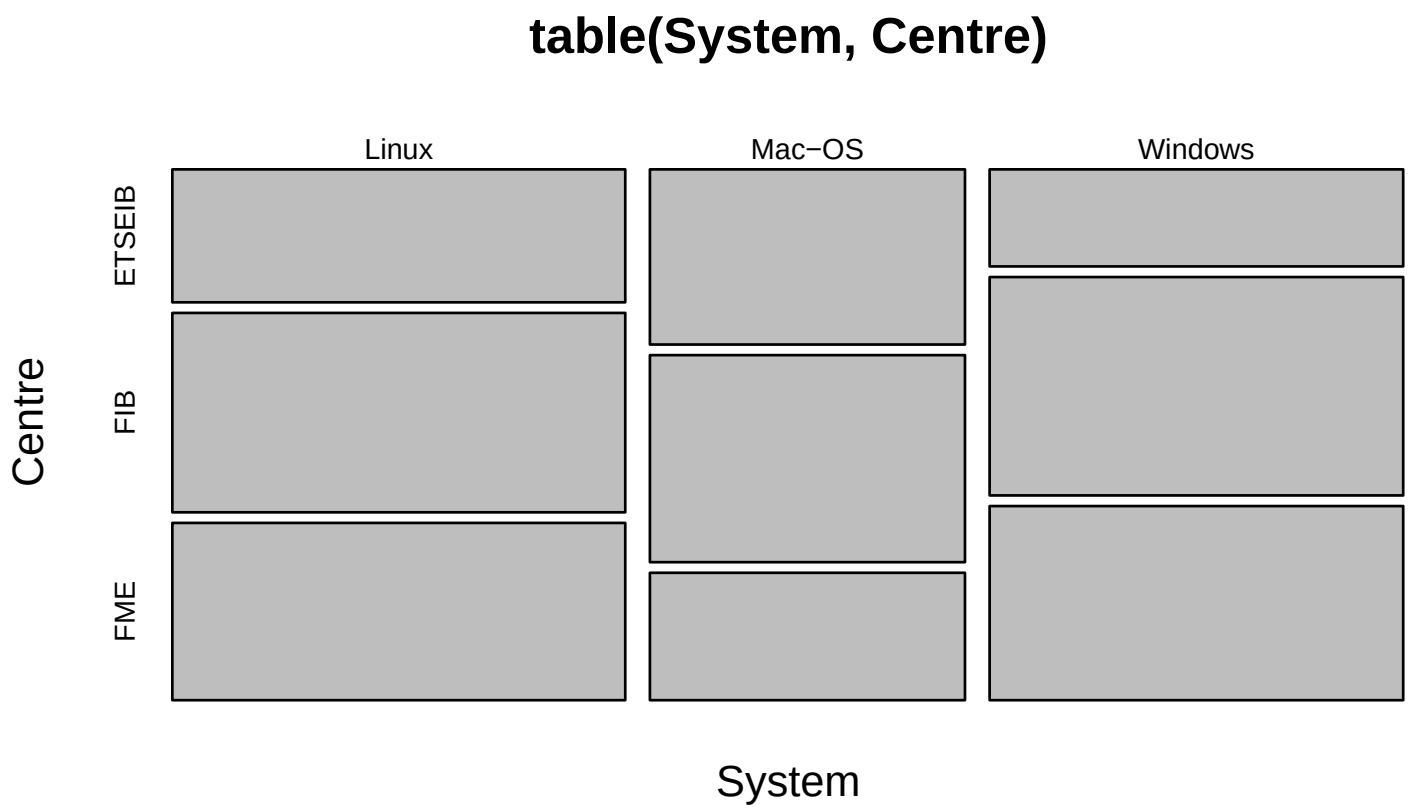
```
barplot(table(System,Centre))
```



Representacions Gràfiques (cat-cat)

Una representació habitual per una taula de contingència és el *Mosaic plot*

```
mosaicplot(table(System,Centre))
```



Resums Numèrics (cat-num)

Càlcul d'estadístics per grups:

- La variable categòrica indueix una segmentació de la mostra en grups
- Dins de cada grup es calculen els estadístics per a la variable numèrica

Permet comparar la distribució de la variable numèrica entre els diferents grups.

- Si els estadístics són semblants, no hi ha relació entre les dues variables
- Si hi ha diferències clares, les dues variables estan relacionades

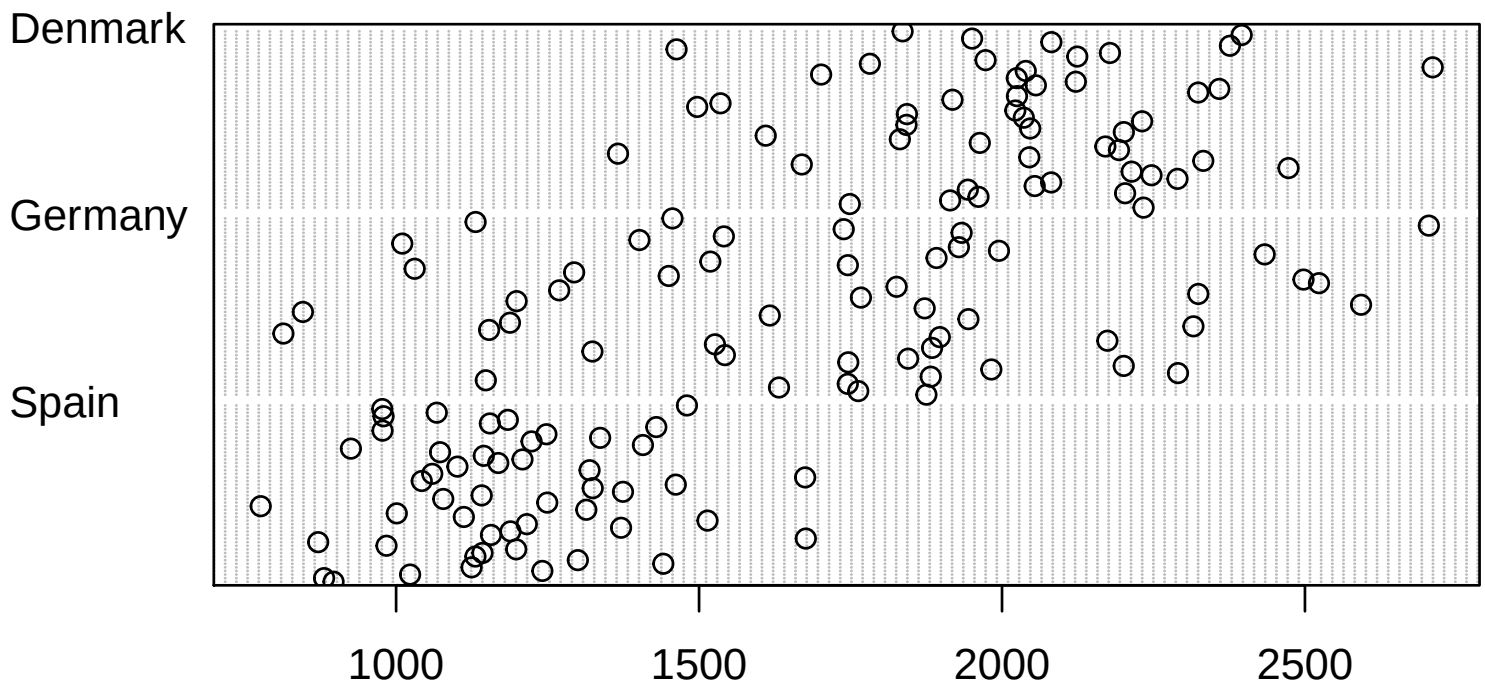
Resums Numèrics (cat-num)

```
by(Sous,Pais,summary)
```

```
## Pais: Denmark
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1366    1861    2042    2025    2203    2711
## ------------------------------------------------
## Pais: Germany
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     814    1414    1754    1728    1942    2704
## ------------------------------------------------
## Pais: Spain
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   776.5  1061.5  1162.6  1187.4  1317.9  1676.2
```

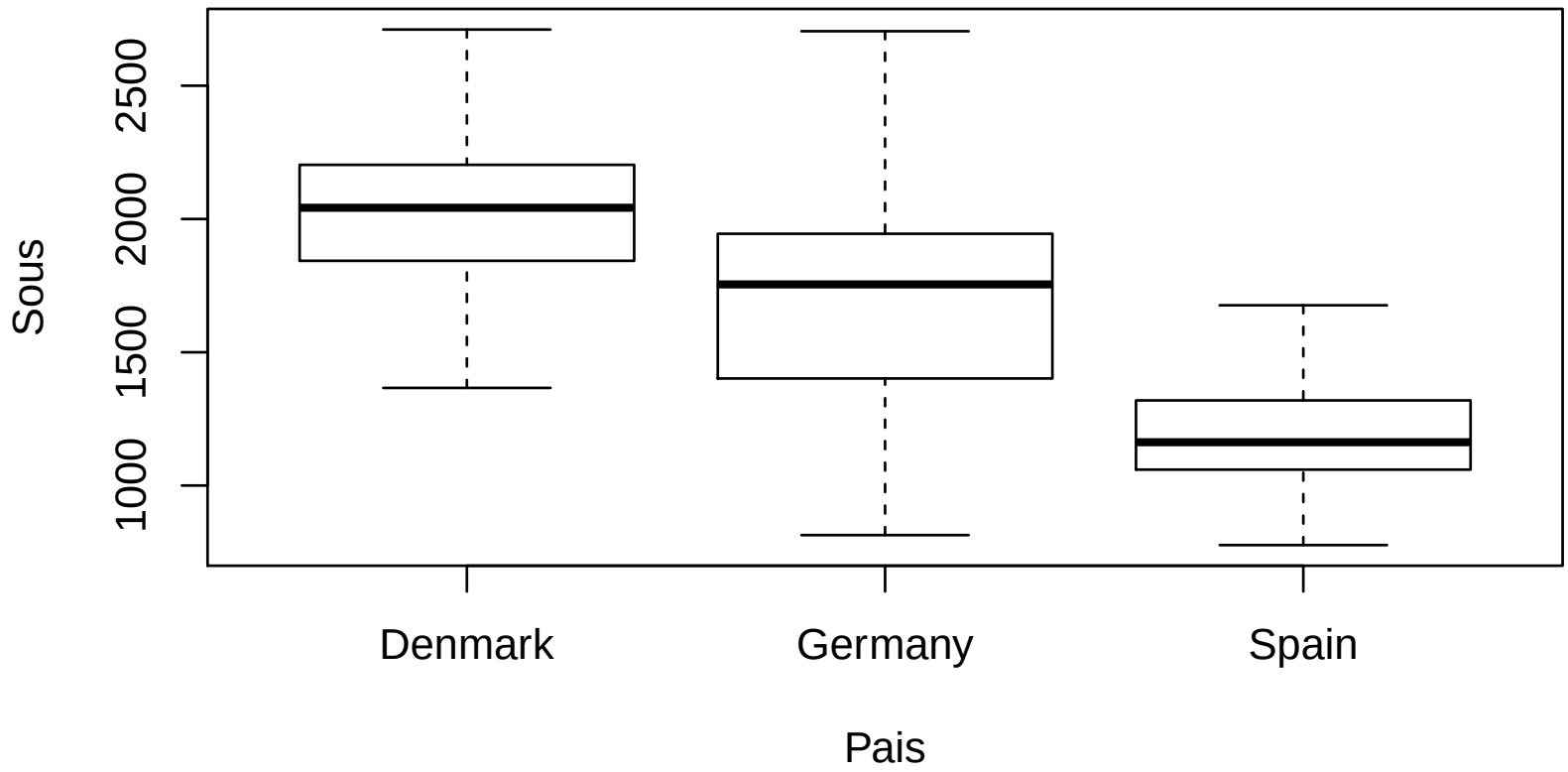
Representacions gràfiques (cat-num)

```
dotchart(Sous,groups=factor(Pais))
```



Representacions gràfiques (cat-num)

```
boxplot(Sous~Pais)
```



Resums Numèrics (num-num)

Càlcul de mesures:

- **Covariància mostral**:

$$S_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

- **Coeficient de correlació mostral**: mesura el grau de relació lineal entre dues variables

$$r_{XY} = \frac{S_{XY}}{S_X S_Y}$$

Resums Numèrics (num-num)

Propietats:

1. $r_{XY} \in [-1, 1]$
2. $r_{XY} = 0 \Rightarrow$ No hi ha relació lineal entre X i Y
3. $r_{XY} = \pm 1 \Rightarrow$ Relació lineal perfecta entre X i Y ($X = \pm Y$)

Observació: Ambdues mesures tenen les seves corresponents versions teòriques, si es coneix la distribució de les variables de les qual provenen les dades.

$$\sigma_{XY} = E\left[(X - E[X])(Y - E[Y])\right]$$

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

Resums Numèrics (num-num)

Conjunt de dades `mtcars`: consum (mpg), Potència(hp) i Pes (wt)

Covariància:

```
cov(mtcars[,c("mpg","hp","wt")])
```

```
##              mpg          hp         wt
## mpg   36.324103 -320.73206 -5.116685
## hp  -320.732056 4700.86694 44.192661
## wt    -5.116685   44.19266  0.957379
```

Correlació:

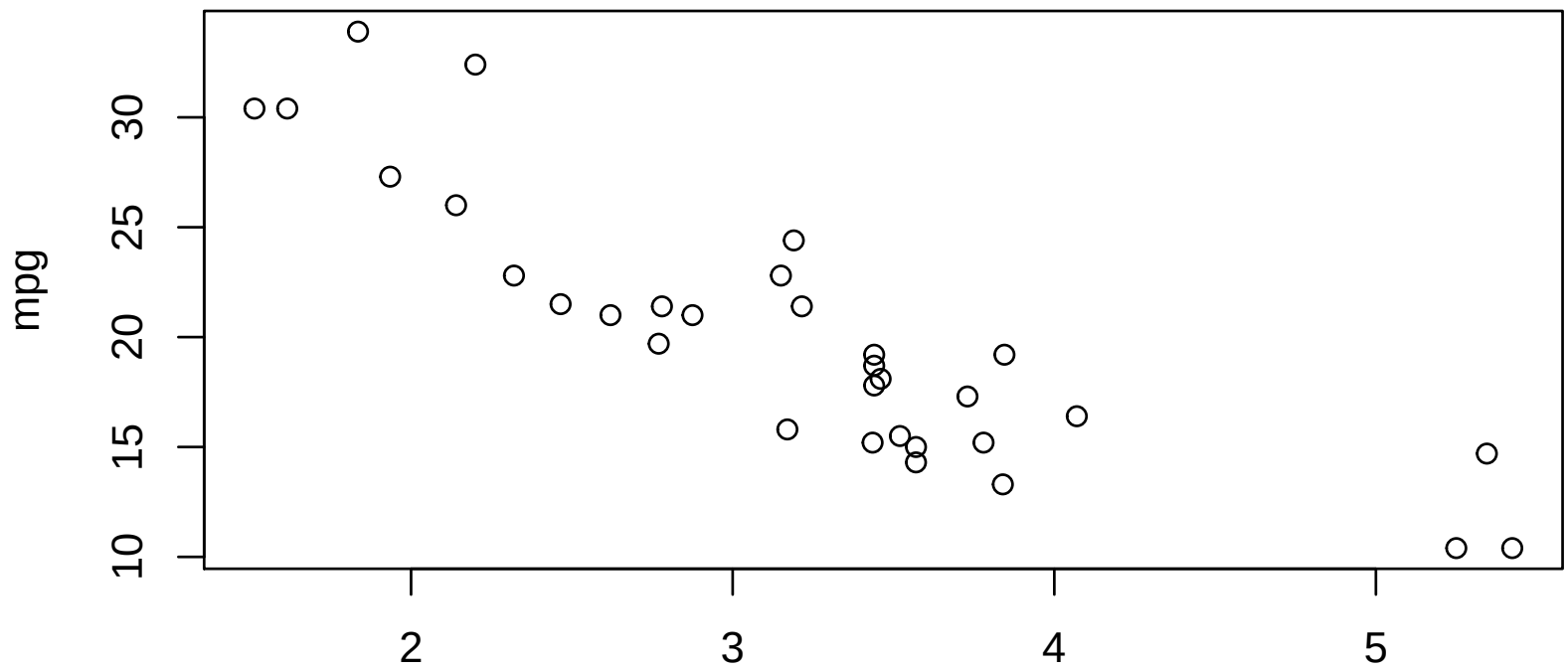
```
cor(mtcars[,c("mpg","hp","wt")])
```

```
##              mpg         hp         wt
## mpg  1.0000000 -0.7761684 -0.8676594
## hp  -0.7761684  1.0000000  0.6587479
## wt  -0.8676594  0.6587479  1.0000000
```

Representacions Gràfiques (num-num)

Diagrama de punts bivariant (*scatterplot*): Representació en el pla, on les coordenades venen determinades pels valors d'ambdues variables.

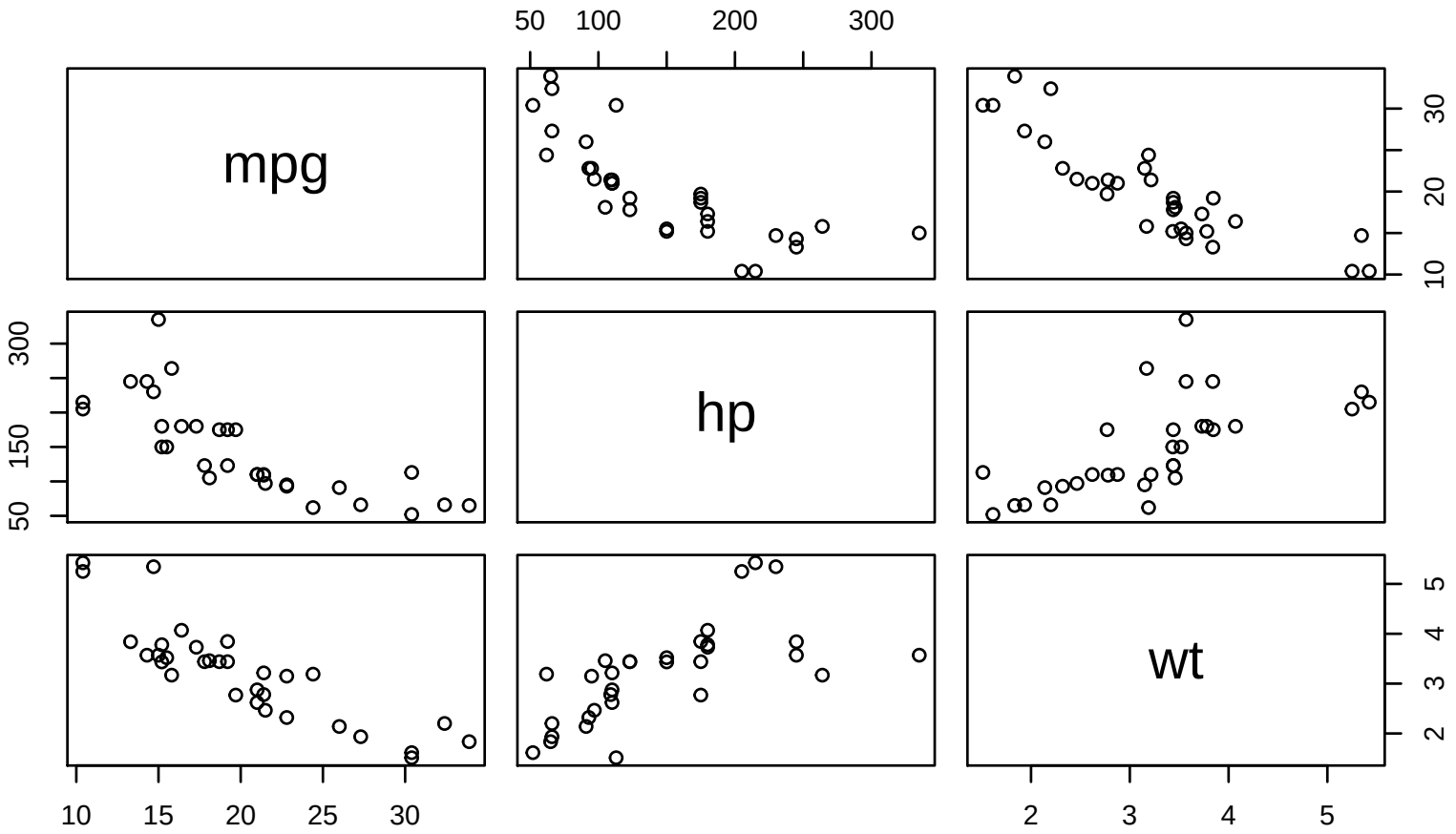
```
plot(mpg~wt,mtcars)
```



Representacions Gràfiques (num-num)

Matrix plot

```
pairs(~mpg+hp+wt,mtcars)
```



Representacions Gràfiques (num-num)

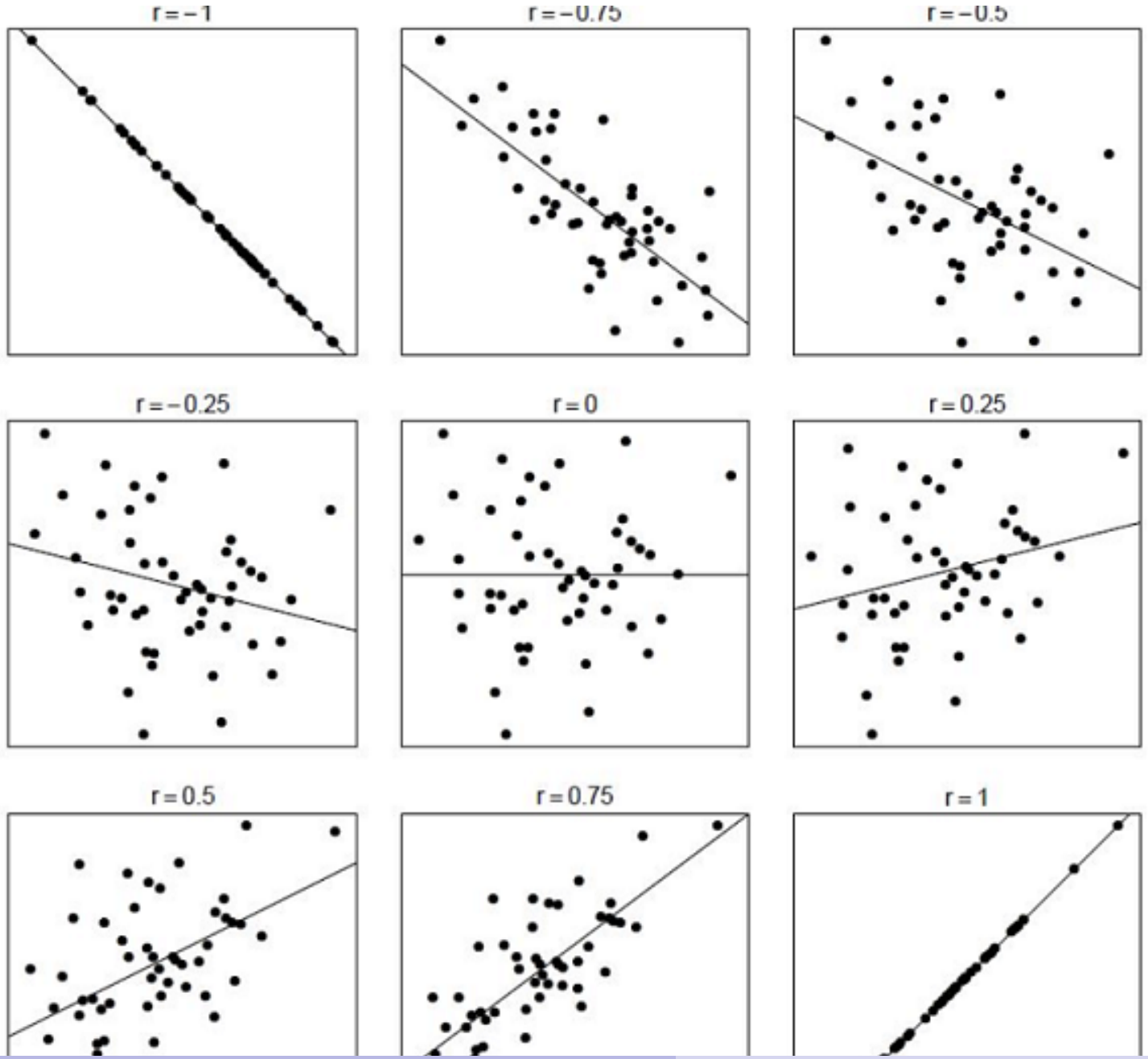
Quan fem el *scatter plot* té sentit pensar en la recta $Y = \beta_0 + \beta_1 X$ que "millor" ajusta el núvol de punts en el sentit de que minimitza:

$$\sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

Aquesta recta s'anomena **recta de regressió**

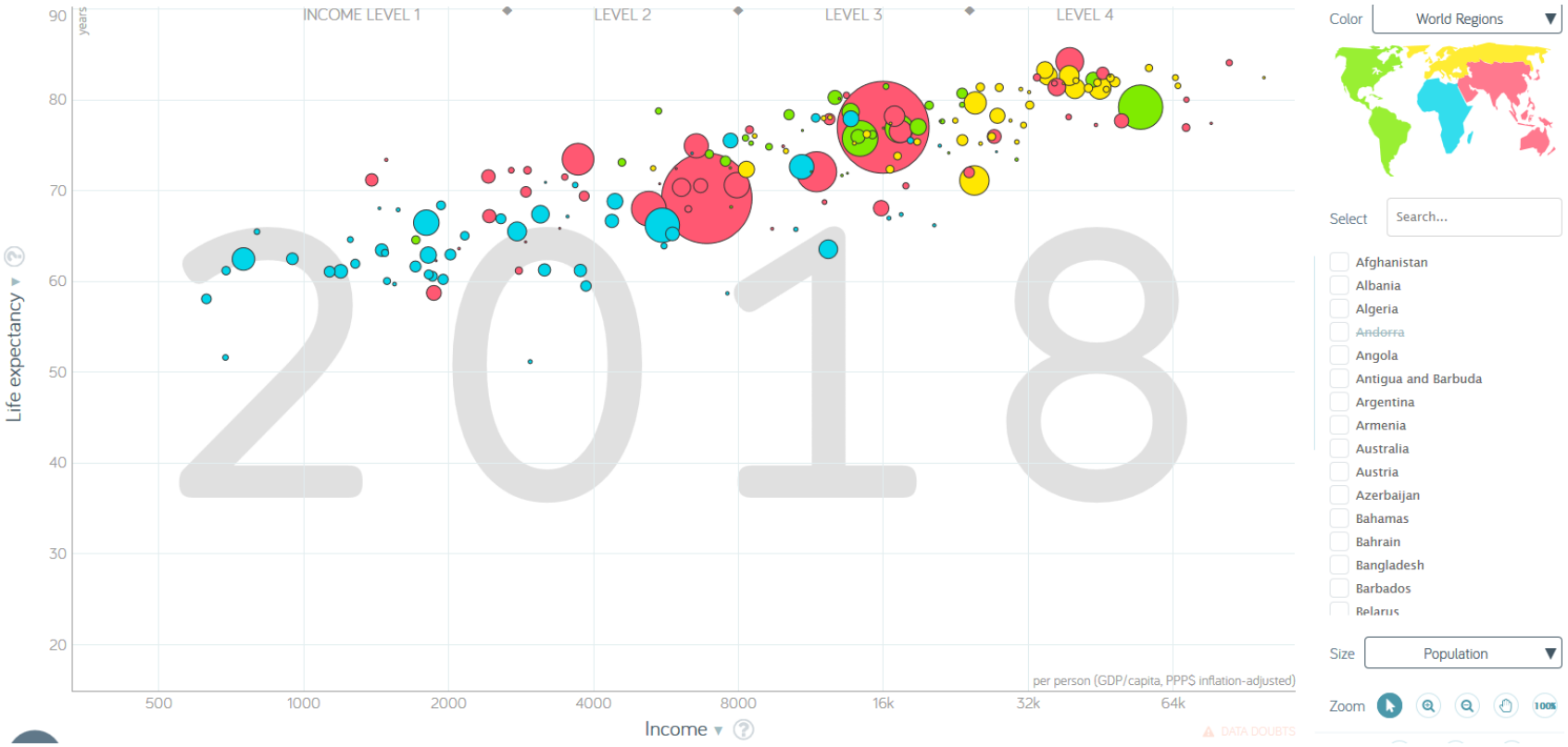
La pendent de la recta de regressió està molt relacionada amb el coeficient de correlació.

Representacions Gràfiques (num-num)



Representacions Gràfiques Multivariants

http://www.gapminder.org



En aquest gràfic es representen 4 variables simultàniament