FIB
Departament d'Estadística
i Investigació Operativa
UNIVERSITAT POLITÈCNICA DE CATALUNYA
UPC

Models estadístics i ciència de dades
Exemples i exercicis

Bloc D – Probabilitat i Estadística

2024

Índex

1. Exemple compressor. Descriptiva, estimació puntual i per IC
2. Exemple de comparar 3 rendiments en experiment amb plantes
3. Exemple benzina i velocitat. Model lineal
4. Exemple cervesa alcohol. Model lineal
5. Exemple brillantor i durada. Model lineal
6. Exemple de transformació logarítmica en model lineal
7. Exemple Dijkstra. Comparar mitjanes de temps (logarítmics)
8. Exemple recorregut arbre. Model amb explicatives quantitatives i qualitatives
9. Exemple “iris” amb PCA i k-means
10. Exemple “Estudi Web Benchmark JetStream 2.0” amb PCA

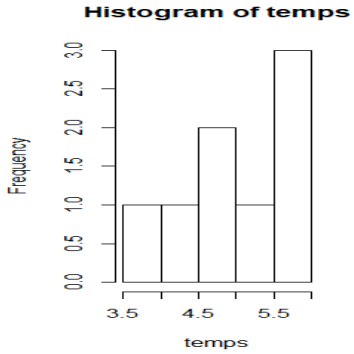
Exemple model lm (Y~1) . Compressor

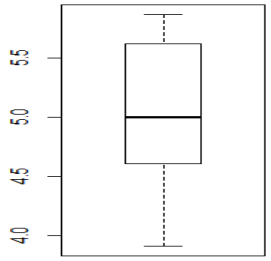
El fabricant d'un determinat compressor d'arxius assegura que arxius de 500Kb són comprimits en 5 segons. Aquest compressor ha estat testejat amb 8 arxius i el temps en segons necessari per a cada compressió han estat : 4.85,4.36,5.12,5.64,5.6,5.87,3.91,4.88

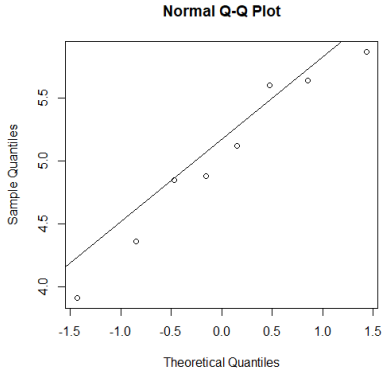
```
R: temps <- c(4.85,4.36,5.12,5.64,5.6,5.87,3.91,4.88)
```

Estadística Descriptiva:

```
R: hist(temps)   boxplot(tesmps)   qqnorm(temps) qqline(temps)
```







Estimació puntual:

```
R: mean(temps)     [1] 5.02875
   sd(temps)       [1] 0.6728285
   var(temps)      [1] 0.4526982
   summary(temps)
    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    3.910   4.728   5.000   5.029   5.610   5.870
```

Exemple model lm (Y~1) . Compressor

Compleixen que es comprimeixin en 5 sg en mitjana? (inferència sobre la mitjana)

```
lm(temps~1)
summary(lm(temps~1))
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.0287     0.2379   21.14 1.33e-07 ***

Residual standard error: 0.6728 on 7 degrees of freedom
```

IC ($\mathbf{5.0287} \pm t_{7,0.975}\ 0.2379$ = **[4.47,5.59]**

5 cau dins del IC, 5 segons de mitjana és un valor versemblant però no podem afirmar que sigui la mitjana real. La diferència entre la mitjana mostral i 5 (μ hipotètica) podria explicar-se per l'atzar. La part residual que el model no recull, o desviació dels residus, és 0.6728

Si la variància és superior a 0.22 segons^2 el compressor té una qualitat insuficient.
És suficient o no? (inferència sobre la desviació)

$$IC(\sigma^2, 0.95) = \left(\frac{s^2(n-1)}{\chi^2_{n-1,1-\alpha/2}}, \frac{s^2(n-1)}{\chi^2_{n-1,\alpha/2}}\right) = \left(\frac{0.67^2(8-1)}{\chi^2_{7,0.975}}, \frac{0.67^2(8-1)}{\chi^2_{7,0.025}}\right) = \left(\frac{3.14}{16.01}, \frac{3.14}{1.69}\right) = (0.20,1.86)$$

0.22 cau dins del IC, per tant té una qualitat suficient (almenys no hi ha motius per dubtar-ne).

Exemple model lm (Y~G) . Rendiment 3 plantes

PlantGrowth {datasets}

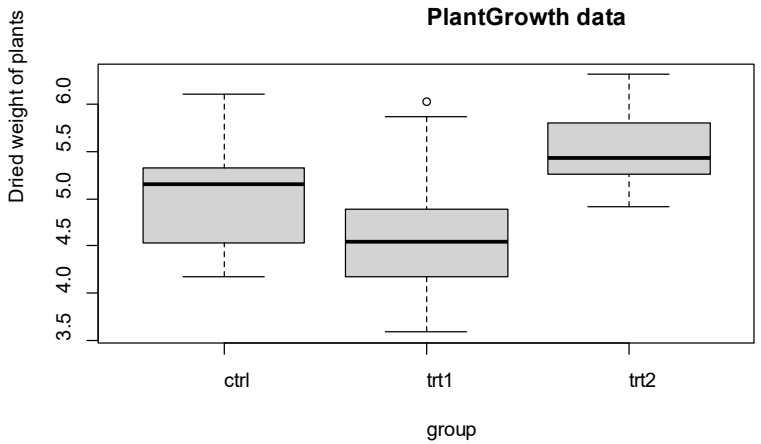
Results from an Experiment on Plant Growth

Resultats d'un experiment per comparar els rendiments
(mesurats pel pes sec de les plantes)
obtinguts sota un control i dues condicions de tractament diferents

```
require(stats); require(graphics)

boxplot(weight ~ group, data = PlantGrowth,
main = "PlantGrowth data", ylab = "Dried
weight of plants", col = "lightgray")

mod=lm(weight~group,data=PlantGrowth)
summary(mod)
```




```
Residuals:
    Min      1Q  Median      3Q     Max
-1.0710 -0.4180 -0.0060  0.2627  1.3690

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.0320     0.1971  25.527   <2e-16 ***
grouptrt1    -0.3710     0.2788  -1.331   0.1944
grouptrt2     0.4940     0.2788   1.772   0.0877 .

Residual standard error: 0.6234 on 27 degrees of freedom
Multiple R-squared: 0.2641, Adjusted R-squared:  0.2096
F-statistic: 4.846 on 2 and 27 DF,  p-value: 0.01591
```

```
library(emmeans)
emmeans(mod, ~group)
```

group	emmean	SE	df	lower.CL	upper.CL
ctrl	5.03	0.197	27	4.63	5.44
trt1	4.66	0.197	27	4.26	5.07
trt2	5.53	0.197	27	5.12	5.93

```
pairs(emmeans(mod, ~group))
```

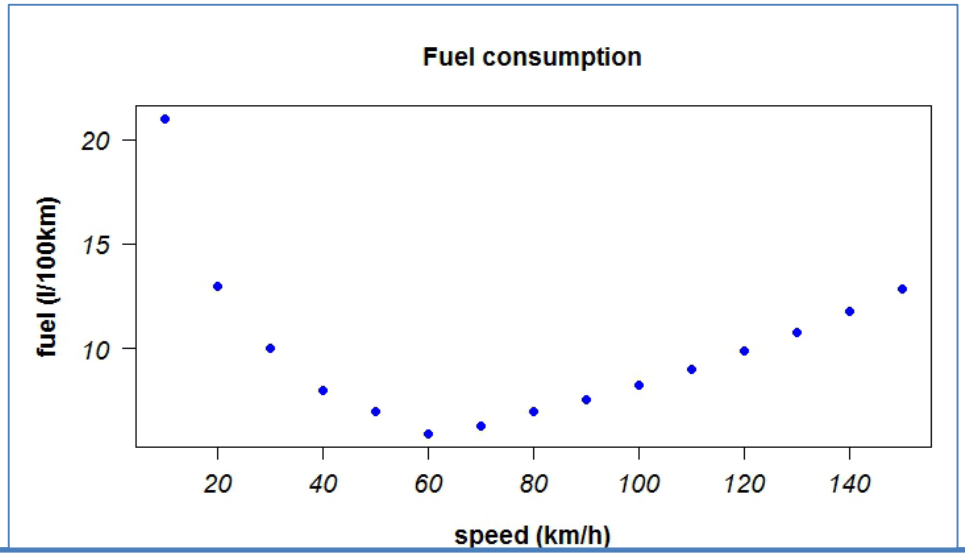
contrast	estimate	SE	df	t.ratio	p.value
ctrl - trt1	0.371	0.279	27	1.331	0.3909
ctrl - trt2	-0.494	0.279	27	-1.772	0.1980
trt1 - trt2	-0.865	0.279	27	-3.103	0.0120

Exemple model lm (Y~X) . Benzina i velocitat

- Una equació com $Y = b_0 + b_1 \cdot X$ pot relacionar-nos dues variables com el consum de benzina i la velocitat (dades a la taula)
- Així, tenim un model per previsions del **consum** (Y) segons la **velocitat** (X):

$$Y = 11.058 - 0.01466 \cdot X$$

- *Què vol dir el coeficient –0.01466? Realment podem esperar menys consum amb més velocitat, veient el gràfic?*
- A més, no oblidem que el consum de benzina no depèn només de la velocitat.

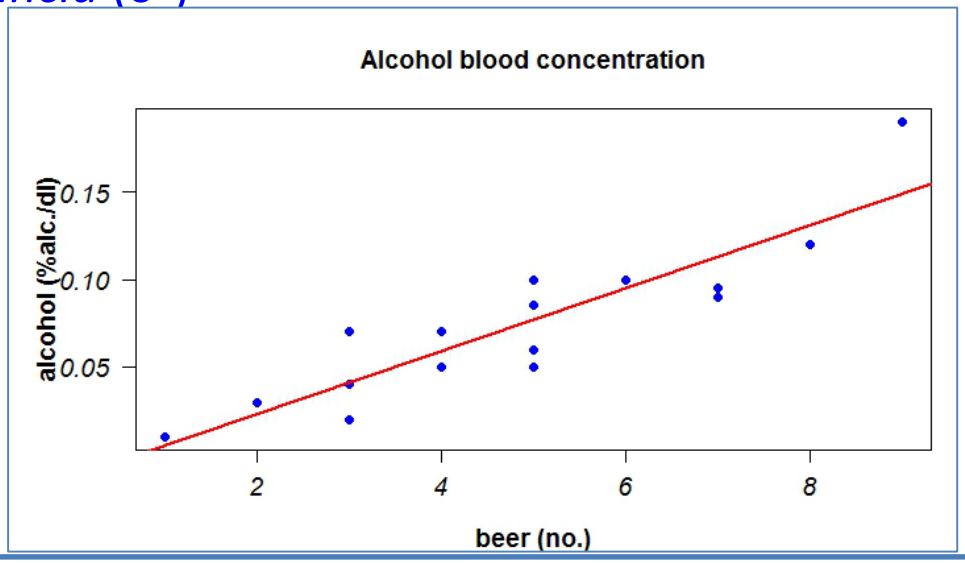


speed (km/h)	fuel (l/100 km)
10	21
20	13
30	10
40	8
50	7
60	5.9
70	6.3
80	6.95
90	7.57
100	8.27
110	9.03
120	9.87
130	10.79
140	11.77
150	12.83

Exemple model lm (Y~X) . Cervesa i alcohol

- Un estudi ha sol·licitat a 16 voluntaris que es prengui una quantitat determinada (aleatòriament) de cervesa, mesurada en llaunes, i es mesura l'alcohol a la sang trenta minuts després [%alc. /dl sang].
- Un model simple és ajustar-hi una recta, que implica dos paràmetres: *pendent* (β_1) i *constant* (β_0) a l'origen
- Al voltant tenim una certa dispersió que requereix un tercer paràmetre: la *variància* (σ^2)

Source: The Basic Practice of Statistics. 4th ed. David S. Moore. Example 24.7



cerveses	alcohol
5	0.100
2	0.030
9	0.190
8	0.120
3	0.040
7	0.095
3	0.070
5	0.060
3	0.020
5	0.050
4	0.070
6	0.100
5	0.085
7	0.090
1	0.010
4	0.05

Exemple model lm (Y~X) . Cervesa i alcohol

cerveses	alcohol
5	0.100
2	0.030
9	0.190
8	0.120
3	0.040
7	0.095
3	0.070
5	0.060
3	0.020
5	0.050
4	0.070
6	0.100
5	0.085
7	0.090
1	0.010
4	0.05

Càlculs dels estadístics convencionals:

$\bar{y} = 0.07375$ $\quad s_Y^2 = 0.0019483$ $\quad s_{XY} = 0.08675$

$\bar{x} = 4.8125$ $\quad s_X^2 = 4.829167$ $\quad r_{XY} = \frac{s_{XY}}{s_X s_Y} = 0.894338$

Resultats de la regressió:

$b_1 = \frac{s_{XY}}{s_X^2} = r_{XY} \cdot \frac{s_Y}{s_X} = 0.01796$ $\quad b_0 = \bar{Y} - b_1\bar{X} = -0.0127$ $\quad s = \sqrt{\frac{\sum(e_i^2)}{n-2}} = 0.0204$

Model amb R:

```
> lm(alc ~ n.cerv)    # (alc és Y, n.cerv és X)
Call:
lm(formula = alc ~ n.cerv)
Coefficients:
(Intercept)        n.cerv
   -0.01270       0.01796
```

Variància de l'error amb R:

```
sum(lm(alc~n.cerv)$resid^2)/14
```

Exemple model lm (Y~X) . Cervesa i alcohol

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.0127     0.0126   -1.00     0.33
n.cerv        0.0180     0.0024    7.48  3.0e-06 ***
---
Signif.codes:0‘***’ 0.001‘**’ 0.01‘*’ 0.05‘.’ 0.1‘ ’ 1

Residual standard error: 0.0204 on 14 degrees of freedom
Multiple R-squared:  0.8,       Adjusted R-squared: 0.786
F-statistic: 55.9 on 1 and 14 DF,  p-value: 2.97e-06
```

$IC_{95\%}$: $IC(\beta_1, 95\%) = b_1 \mp t_{n-2,0.975} \cdot s_{b_1} = 0.018 \mp 2.15 \cdot 0.0024 = [0.013, 0.023]$

(Cada cervesa de més incrementa el contingut d'alcohol per decilitre de sang en un valor que pot estar entre 0.0128% i 0.0231%, amb un 95% de confiança)

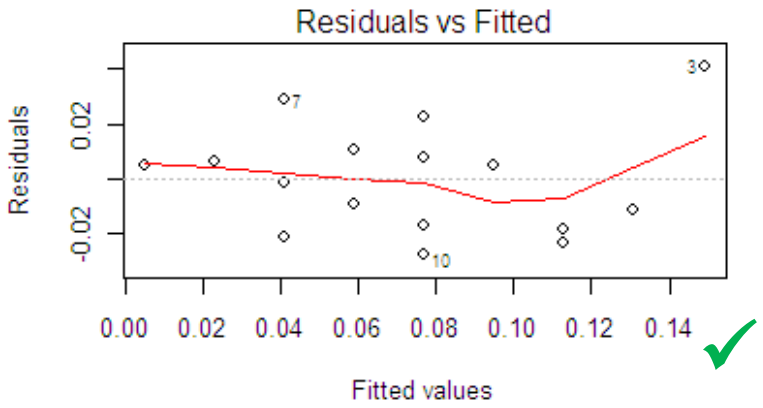
Conclusió pràctica: No és versemblant que el coeficient del pendent sigui 0

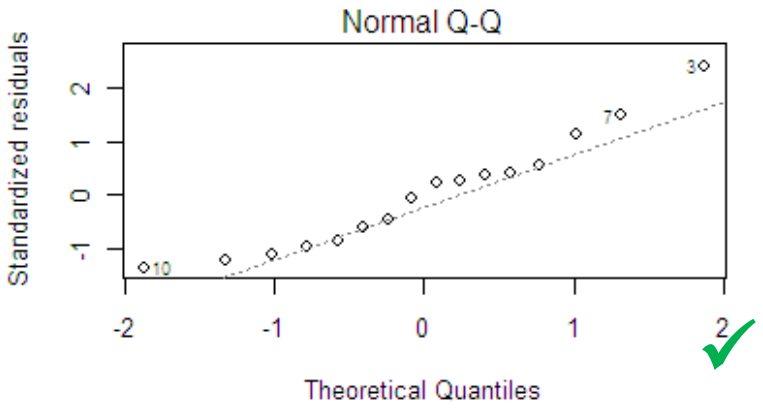
$IC_{95\%}$: $IC(\beta_0, 95\%) = b_0 \mp t_{n-2,0.975} \cdot s_{b_0} = 0.0127 \mp 2.15 \cdot 0.0126 = [-0.040, 0.014]$

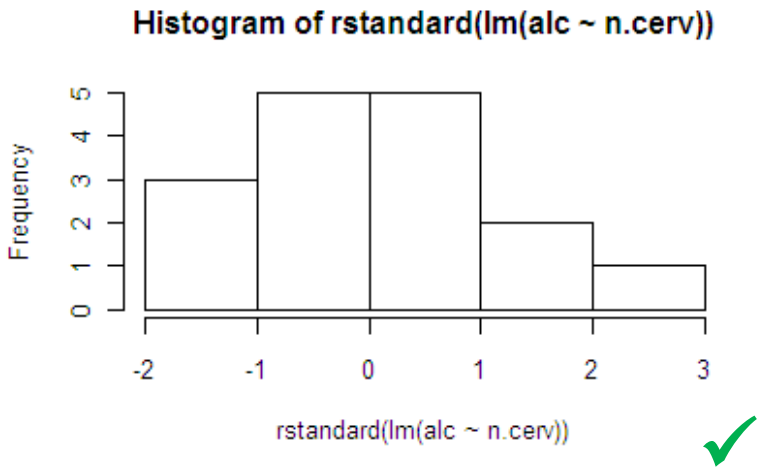
(És versemblant que el terme independent sigui 0. No es pot rebutjar que la recta passi per l'origen, pel punt (0,0). A 0 llaunes de cervesa li correspon una quantitat d'alcohol en sang de 0.0%)

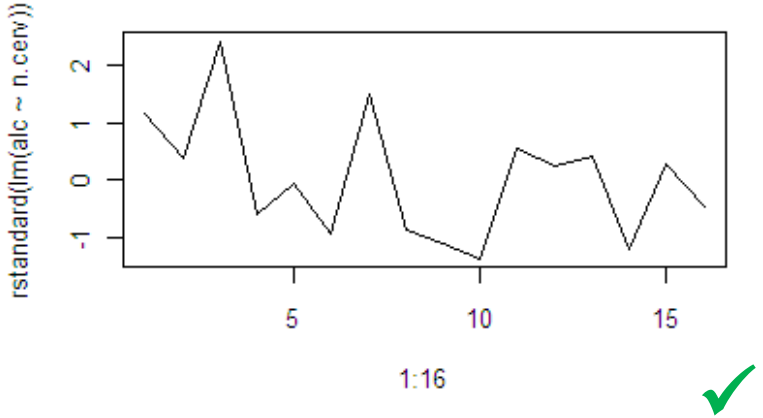
Exemple model lm (Y~X) . Cervesa i alcohol (validació)

```
##-- Exemple de les cerveses
par(mfrow=c(2,2))
plot(lm(alc~n.cerv),c(2,1))                    # QQ-Norm i Standard Residuals vs. Fitted
hist(rstandard(lm(alc~n.cerv)))                # Histograma dels residus estandaritzats
plot (1:16,rstandard(lm(alc~n.cerv)),type="l") # Ordre dels residus estandaritzats
```









Encara que són poques dades, validem (donem *encara* per bones) les 4 premisses

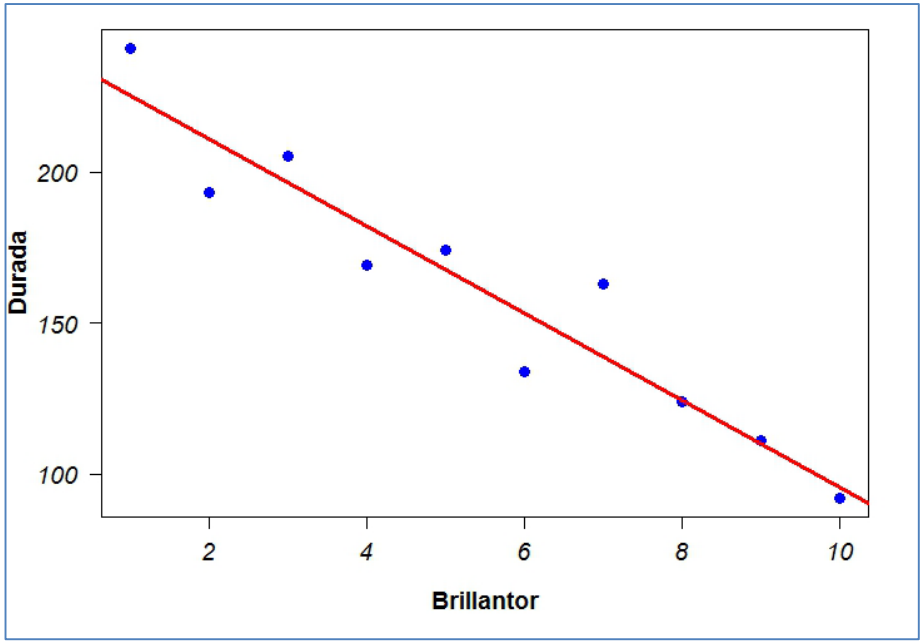
Exemple model lm (Y~X) . Brillantor i durada

La pantalla de l'ordinador portàtil és l'element que consumeix més energia del sistema. Per estudiar l'impacte que el nivell de brillantor de la pantalla (que l'usuari pot graduar) té en la durada de la bateria, treballant amb tasques quotidianes, es mesura el temps que l'ordinador triga des que arrenca amb la bateria totalment carregada fins que avisa per manca d'energia suficient per continuar. Els resultats obtinguts figuren a continuació:

Brillantor (X)	1	2	3	4	5	6	7	8	9	10
Durada (Y)	241	193	205	169	174	134	163	124	111	92

Varia la durada de la bateria segons el nivell de brillantor?

```
> plot(Durada~Brillantor)
> abline(lm(Durada~Brillantor))
```



Exemple model lm (Y~X) . Brillantor i durada

$$\left.\begin{array}{c}\bar{y}=160.6\\ s_y^2=2106.044\\ \bar{x}=5.5\\ s_x^2=9.167\\ s_{xy}=-132.11\\ r_{xy}=s_{xy}/(s_x s_y)=-0.95\end{array}\right\}\rightarrow\left\{\begin{array}{c}b_1=\frac{s_{xy}}{s_x^2}=r_{xy}\cdot\frac{s_y}{s_x}=-14.41\\ b_0=\bar{y}-b_1\bar{x}=239.9\\ s^2=\frac{\sum e_i^2}{n-2}=227.3\end{array}\right.$$

```
> summary(lm(Durada~Brillantor,datos))
Call:
lm(formula = Durada ~ Brillantor, data = datos)
Residuals:
     Min       1Q   Median       3Q      Max
-19.3939 -10.8500   0.1364   7.8258  24.0182
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   239.87      10.30  23.290 1.23e-08 ***
Brillantor    -14.41       1.66  -8.683 2.41e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 15.08 on 8 degrees of freedom
Multiple R-squared:  0.9041,            Adjusted R-squared:
0.8921
F-statistic: 75.39 on 1 and 8 DF,  p-value: 2.411e-05
```

Recta resultant: $\hat{y}_i = 239.9 - 14.41x_i$

Interpretació de b_1: Per cada grau de brillantor augmentat, la bateria dura uns 14.4 minuts menys.

IC$_{95\%}$: $IC(\beta_1, 95\%) = b_1 \mp t_{n-2,0.975} \cdot s_{b_1} = -14.41 \mp 2.3 \cdot 1.66 = [-18.23, -10.59]$

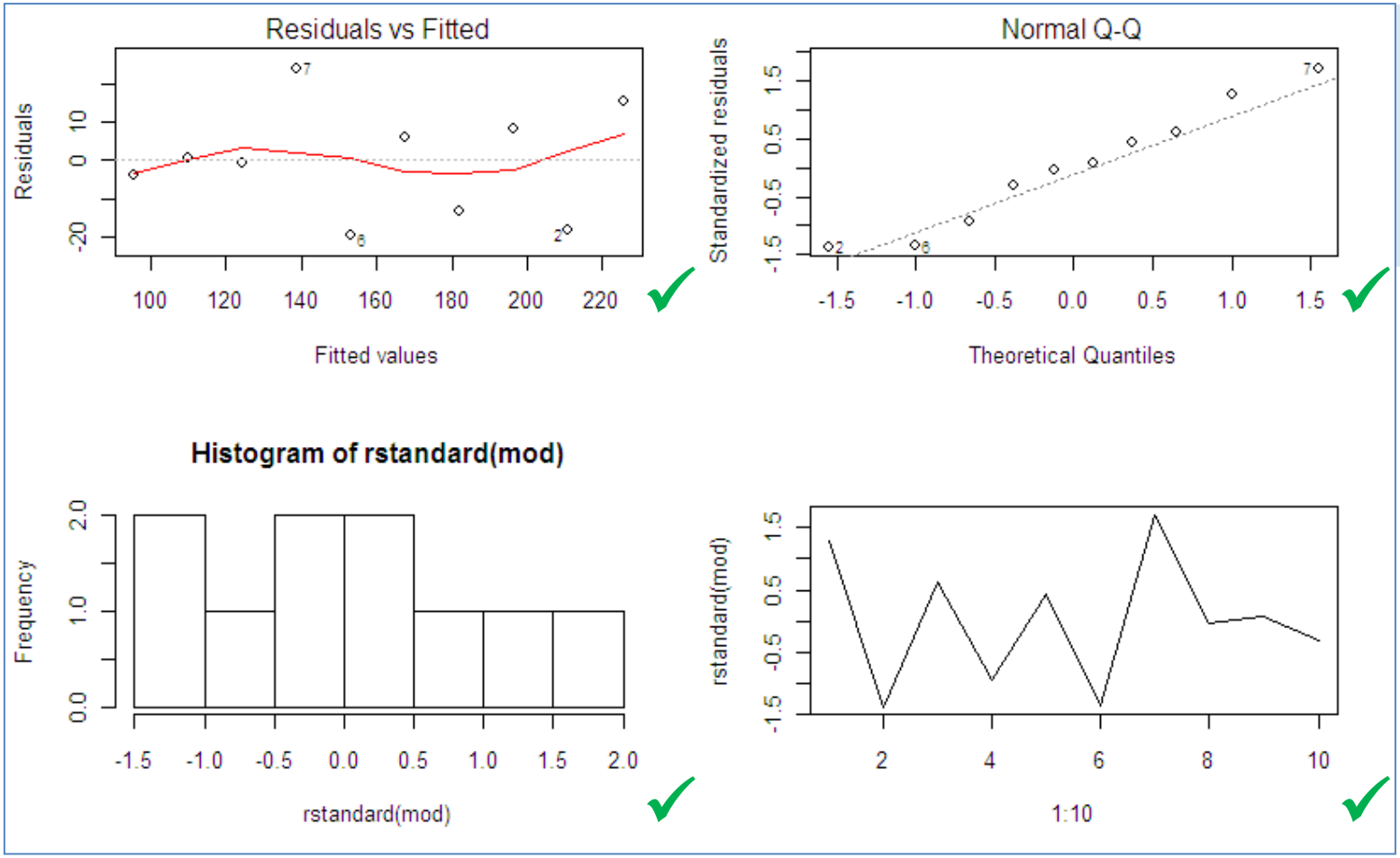
(Cada grau que pugem la brillantor de la pantalla significa entre uns 10 i 18 minuts menys de durada de la bateria)

Interpretació de b_0: Amb un grau de brillantor nul (sense usar la pantalla), la bateria durarà unes 4 hores (239.9 minuts)

Interpretació de la s: la desviació residual és 15.1. Podem esperar fluctuacions d'uns quinze minuts respecte les previsions de durada en funció de la brillantor que ens doni el model

Exemple model lm (Y~X) . Brillantor i durada (validació)

```
##-- Exemple de la pantalla d'ordinador
par(mfrow=c(2,2))
plot(lm(Durada ~ Brill),c(2,1))                    # QQ-Norm i Standard Residuals vs. Fitted
hist(rstandard(lm(Durada ~ Brill)))                # Histograma dels residus estandaritzats
plot (1:10,rstandard(lm(Durada ~ Brill)),type="l") # Ordre dels residus
```



Encara que són poques dades, podem validar (o donar *encara* per bones) les 4 premisses

Exemple model lm (Y~X) . Brillantor i durada (predicció)

Les dades de l'exemple de la pantalla d'ordinador

Brillantor (X)	1	2	3	4	5	6	7	8	9	10
Durada (Y)	241	193	205	169	174	134	163	124	111	92

Havíem trobat que la recta estimada era:

$\hat{y}_i = 239.9 - 14.41x_i$

Quina durada podem esperar per a pantalles de brillantor 7.5?

$\bar{x} = 5.5$
$s_x^2 = 9.167$
$s^2 = 227.3$

	Valor esperat	Valors individuals
Estimació puntual	$\hat{y}_h = 239.9 - 14.41 \cdot 7.5 = \mathbf{131.8}$ `new <- data.frame("X"=7.5)` `predict(lm(Y~X),new)` -> *131.8*	$\hat{y}_h = 239.9 - 14.41 \cdot 7.5 = \mathbf{131.8}$ `new <- data.frame("X"=7.5)` `predict(lm(Y~X),new)` -> *131.8*

Estimació per interval	`predict(lm(Y~X),new,int="confidence")`
	`fit lwr upr`
	`1 131.8 118.4 145.2`
		`predict(lm(Y~X),new,int="prediction")`
		`fit lwr upr`
		`1 131.8 94.5 169.0`

Conclusió	**Per a les pantalles** de brillantor de 7.5 podem esperar una **durada mitjana** entre 118.4 i 145.2 min. amb una confiança del 95%	**Per a una pantalla** de brillantor 7.5 podem esperar una **durada** entre 94.5 i 169.0 min. amb una confiança del 95%

Veure gràfics de pags. 191-192 a *Estadística per a enginyers informàtics*. Ed UPC

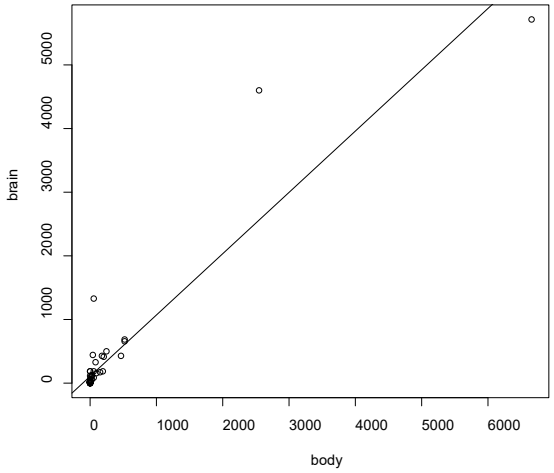
Exemple. Transformació logarítmica en model lineal

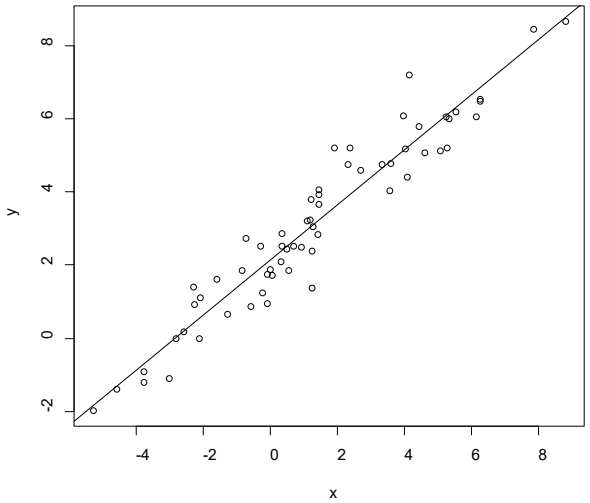
Transformació logarítmica

mammals: Brain and Body Weights for 62 Species of Land Mammals

```
library(MASS)
with(mammals, plot(body, brain))
r1 = lm(brain~body, mammals)
abline(r1)
r1
```

```
     Coefficients:
    (Intercept)          body
        91.0044        0.9665
```

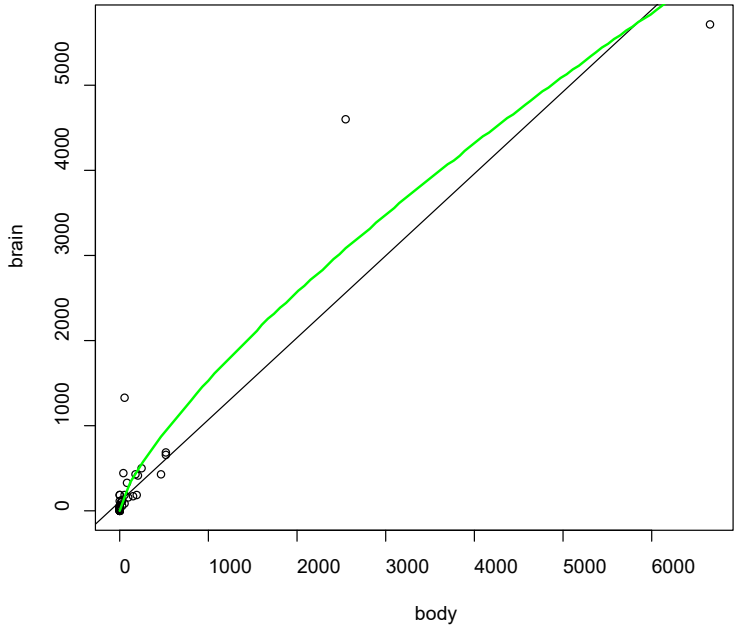





```
y = log(mammals$brain)
x = log(mammals$body)
plot(x, y)
r2 = lm(y~x)
abline(r2)
r2
```

```
    Coefficients:
   (Intercept)            x
        2.1348        0.7517
```

```
with(mammals, plot(body, brain))
rng = range(mammals$body)
X = seq(rng[1], rng[2], len=100)
Y = exp(r2$coef[1]) * X^r2$coef[2]
lines(X, Y, col='green', lwd=2)
```



Exemple. Temps algoritme Dijkstra. Comparar temps logarítmics

Temps i nombre de nodes de graf en l'algoritme de Dijkstra (X: nombre nodes, Y: temps amb transformació logarítmica

$\bar{y}_1 = \mathbf{3.082}$ $s_1^2 = 0.90$ $\bar{y}_2 = \mathbf{4.910}$ $s_2^2 = 1.79$ $\bar{y}_3 = \mathbf{6.83}$ $s_3^2 = 0.79$ $\bar{y} = 4.94$ $s^2 = 3.5 \rightarrow s = 1.87$

(la descriptiva indica usar el logaritme del temps per complir la premissa de normalitat, `lgt<-log(temps)`)

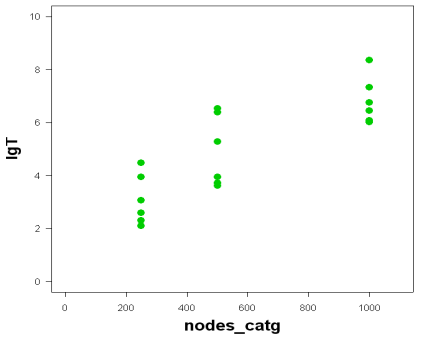
temps	x_i (nodes)	y_i (lgt)
10.07	250	2.31
88.23	250	4.48
13.33	250	2.59
21.33	250	3.06
8.17	250	2.10
51.94	250	3.95
51.42	500	3.94
589.93	500	6.38
678.58	500	6.52
194.42	500	5.27
41.26	500	3.72
36.97	500	3.61
632.70	1000	6.45
1510.21	1000	7.32
862.64	1000	6.76
437.03	1000	6.08
4230.18	1000	8.35
407.49	1000	6.01

El gràfic (mostres en verd) indica un canvi més gran cap a grup 3 que entre els 1 i 2.

El model ho quantifica:

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         3.0817     0.4394   7.014 4.18e-06 ***
as.factor(nodes)500 1.8250     0.6214   2.937   0.0102 *
as.factor(nodes)1000 3.7467    0.6214   6.030 2.31e-05 ***

Residual standard error: 1.076 on 15 degrees of freedom
Multiple R-squared:  0.708,     Adjusted R-squared:  0.669
F-statistic: 18.18 on 2 and 15 DF,  p-value: 9.789e-05
```



El model recull un 70.8 % de la variabilitat total de la variable resposta. La part residual és 1.076

L'estimació de la mitjana de referència, la del grup 1, és 3.082 amb IC 3.082±qt(0.975,15)*0.4394 -> [2.15,4.02]

L'estimació del canvi de la mitjana del 1r al 2n grup és 1.825 amb IC 1.825±qt(0.975,15)*0.6214 -> [0.5,3.15]

L'estimació del canvi de la mitjana del 2n al 3r grup és 3.7467 amb IC 3.7467±qt(0.975,15)*0.6214 -> [2.42,5.07]

A partir de l'estimació de referència i de les dels canvis obtenim les estimacions de les tres mitjanes:

3.082 3.082+1.825 -> **4.91** 3.082+3.7467 -> **6.83**

Exemple. Temps algoritme Dijkstra. Comparar temps logarítmics

Temps i nombre de nodes de graf en l'algoritme de Dijkstra (X: nombre nodes, Y: temps amb transformació logarítmica

$\bar{y}_1 = 3.082 \;\; s_1^2 = 0.90 \qquad \bar{y}_2 = 4.910 \;\; s_2^2 = 1.79 \qquad \bar{y}_3 = 6.83 \;\; s_3^2 = 0.79 \qquad \bar{y} = 4.94 \;\; s^2 = 3.5 \rightarrow S = 1.87$

x_i (nodes)	y_i (lgt)
250	2.31
250	4.48
250	2.59
250	3.06
250	2.10
250	3.95
500	3.94
500	6.38
500	6.52
500	5.27
500	3.72
500	3.61
1000	6.45
1000	7.32
1000	6.76
1000	6.08
1000	8.35
1000	6.01

Abans hem obtingut els IC dels canvis del paràmetre mitjana poblacional dels temps logarítmics entre els grafs de 250 nodes i els de 500 nodes; i entre grafs de 250 i 1000 nodes.

Per exemple, el canvi en mitjana en el temps logarítmic entre grafs de 250 i 500 nodes (el canvi entre el grup 1 i el 2) és 1.825 ($\widehat{\vartheta_2}$ = 1.825 del model amb resposta logarítmica)

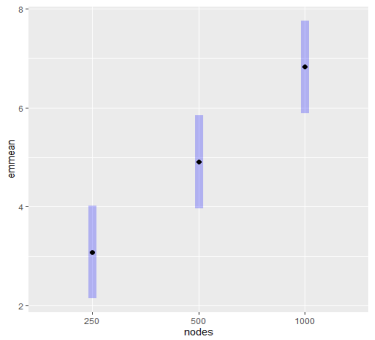
Per tant, el canvi en mitjana en el temps entre grafs de 250 i 500 és exp(1.825)= $e^{1.825}$ = 6.2

També podem trobar l'IC per a cadascun dels paràmetres:

```
emmeans(lm(lgt~as.factor(nodes)),~nodes)
```

```
nodes emmean    SE df lower.CL upper.CL
  250   3.08 0.439 15     2.15     4.02
  500   4.91 0.439 15     3.97     5.84
 1000   6.83 0.439 15     5.89     7.76
Confidence level used: 0.95
```

```
plot(emmeans(…))
```



Els IC de μ_1 μ_2 i μ_3 estan separats (veure gràfic amb IC de color blau), per tant hi ha un increment en les mitjanes esperades del logaritme del temps entre grafs de 250, 500 i 1000 nodes

Desfent logaritmes, per exemple a partir del IC al 95% del temps logarítmic per grafs de 250 nodes, obtenim l'IC del temps per grafs de 250 nodes:

[exp(2.15)= $e^{2.15}$, exp(4.02)= $e^{4.02}$] -> [8.58 , 55.7]

Exemple. Temps recorre arbres. Models amb quantitatives i qualitatives

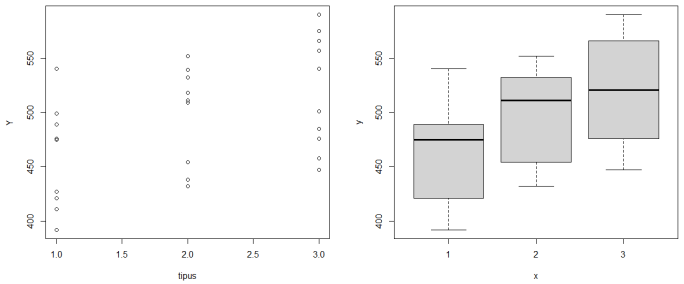
```
Temps <- c(392,421,540,475,427,411,476,489,499,454,509,432,539,552,518,511,438,532,447,590,566,557,540,501,575,458,476,485)
metode <- c(1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,2,3,3,3,3,3,3,3,3,3,3)  # mètodes: preordre, inordre i postordre
nodes <- c(100,140,200,160,120,130,170,180,190,150,160,100,210,200,180,170,140,190,120,190,180,170,160,150,200,110,130,140)
```

Per diverses mides d'arbres (en nombre de nodes) recollim el temps de recórre'ls (o linealitzar-los) usant els tres mètodes indicats. Per explicar la variable de resposta del temps, provem a estimar usant 3 models: model comparant les mitjanes dels tres mètodes, model relacionant amb el nombre de nodes, i model usant mètode i node

summary(lm(Temps~as.factor(metode)))

```
Coefficients:
                    Estimate  Std. Error t value Pr(>|t|)
(Intercept)          458.89       16.29   28.171   <2e-16 ***
as.factor(metode)2    39.44       23.04    1.712   0.0992 .
as.factor(metode)3    60.61       22.45    2.699   0.0123 *

Residual standard error: 48.87 on 25 degrees of freedom
Multiple R-squared:  0.2292,    Adjusted R-squared:  0.1675
F-statistic: 3.716 on 2 and 25 DF,  p-value: 0.03864
```

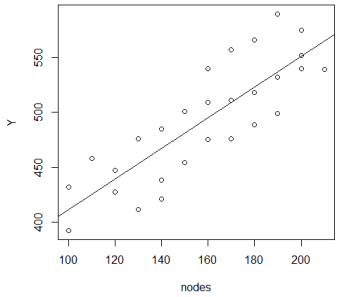


Aquest primer model només explica un 22.92 % de la variabilitat

summary(lm(Temps~nodes))

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 271.2801    29.2824   9.264 1.02e-09 ***
nodes         1.3996     0.1812   7.726 3.38e-08 ***

Residual standard error: 30.06 on 26 degrees of freedom
Multiple R-squared:  0.6966,    Adjusted R-squared:  0.6849
F-statistic: 59.69 on 1 and 26 DF,  p-value: 3.379e-08
```



Aquest segon model explica quasi un 70% (69.66) de la variabilitat. El model és l'equació de la recta Temps = b0+b1*nodes = 271.3+1.4*nodes
Per 0 nodes el temps és 271.3 (seria un temps fixe) i per cada node de més, en el temps podem esperar un augment de 1.4 unitats de temps

Exemple. Temps recorre arbres. Models amb quantitatives i qualitatives

```
summary(lm(Temps~nodes+as.factor(metode)))
```

```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       239.78109   15.88211  15.098 9.43e-14 ***
nodes               1.41868    0.09699  14.628 1.87e-13 ***
as.factor(metode)2 22.10498    7.56022   2.924  0.00743 **
as.factor(metode)3 59.82295    7.27785   8.220 1.95e-08 ***

Residual standard error: 15.84 on 24 degrees of freedom
Multiple R-squared:  0.9223,    Adjusted R-squared:  0.9125
F-statistic: 94.91 on 3 and 24 DF,  p-value: 1.892e-13
```

El model recull un 92.23 % de la variabilitat total de la variable resposta. La part residual és 15.84 (bastant inferior a la dels models anteriors)

Ara l'intercept és 239.78 amb error estàndard 15.88, per tant l'IC 239.78 $\pm$qt(0.975,24)*15.88211 -> [207.001, 272,56]

L'estimació del pendent és 1.42 amb IC 1.42$\pm$qt(0.975,24)*0.09699 -> [1.22, 1.62]

L'estimació del canvi de la mitjana del 1r al 2n mètode és 22.105 amb IC 22.105$\pm$qt(0.975,24)*7.56 -> [6.5, 37.71]

L'estimació del canvi de la mitjana del 1r al 3r mètode és 59.823 amb IC 59.823$\pm$qt(0.975,24)*7.27785 -> [44.8, 74.84]

Exemple “iris” amb PCA i k-means

PCA. Principal Component Analysis

[Y1, Y2, ... Yk] ----> $[\Psi 1, \Psi 2][\Psi 3 \dots \Psi k]$

K variables (Y1...Yk) es poden graficar en 2D (de 2 en 2) o 3D, però no fer gràfic de totes (si k>3 dimensions) on veure totes les relacions. Si es fa un canvi de base, es transformen les k columnes en unes k noves variables (components).

Si es fa un canvi de base que concentri la variabilitat en les primeres components, llavors gràfics en 2D de les primeres components (**components principals,** $\Psi 1\ \Psi 2$) són un bon resum de les dades globals, permetent reduir la dimensionalitat.

HCPC. Hierachical Clustering on Principal Components

Algorisme d'agrupament jeràrquic obtenint un **arbre on els nodes són agrupacions d’observacions**: les fulles representen cada observació com a grup individual, i a l’arrel un sol grup conté totes les observacions.
Comença identificant les dues observacions més properes i crea un nou cas, i itera fins al nivell més alt de la jerarquia.
S’obtenen bons resultats identificant els més propers a partir de “distàncies” en les components principals.

K-means

L'algorisme K-means és un mètode d'agrupament que té com a objectiu la **partició de les observacions en k grups** en el qual cada observació pertany al grup més proper a la mitjana.
Comença a partir de “k centres inicials” i itera re-assignant les observacions entre els k grups segons les “distàncies” al centre (mitjana) fins convergir.

La definició de “distància” o mètrica usada influeix en els resultats i determina variants d’aquestes tècniques

Exemple “iris” amb PCA i k-means

Iris és un conjunt disponible en R (i molt usat per comparar tècniques) amb mides de pètals i sèpals de 150 flors de 3 espècies diferents (setosa, virgínica i versicolor).

A continuació trobareu alguns resultats usant aquestes dades:

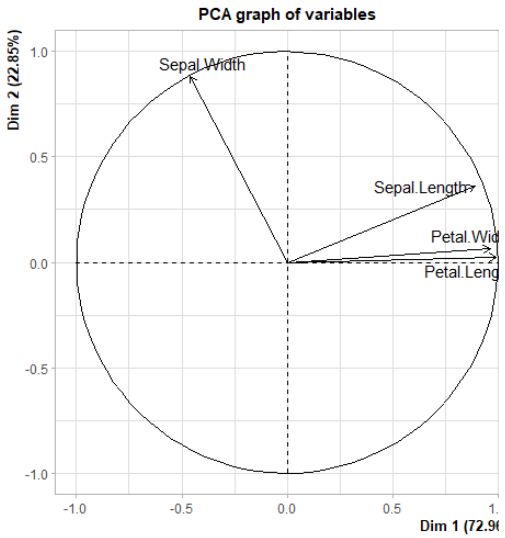
- Un PCA obtenint en **un gràfic de les dues dimensions principals** del canvi de base (pla factorial), amb la **representació de les variables i de les observacions**.
- Un HCPC a partir del pla factorial anterior, obtenint un **arbre amb les agrupacions** de les observacions fins l’arrel, i amb **salts entre nivells proporcionals a les diferències entre grups** ajuntats.
- Una **taula indicant % d’encert** comptant observacions d’una mateixa espècie que forma part d’una mateixa agrupació en l’arbre (en les 3 agrupacions del nivell corresponent)
- Una taula, com l’anterior, indicant % d’encerts comptant observacions d’una mateixa espècie que forma part d’un mateix grup dels 3 obtinguts per kmeans

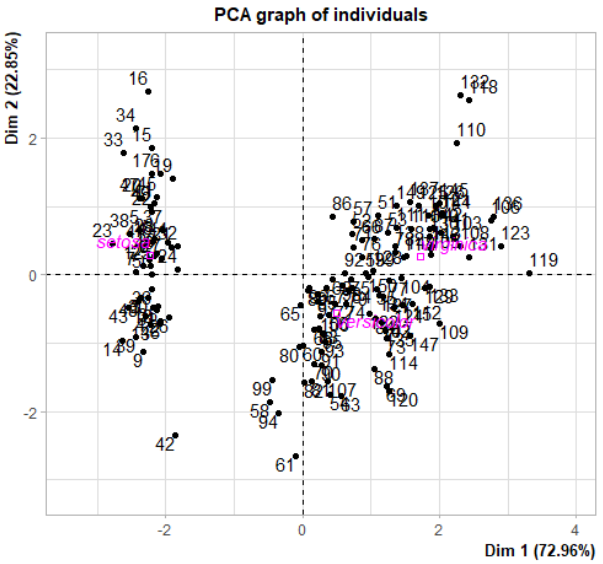
(a més es poden representar els 3 grups de k-means sobre les 2 primeres components principals)

En aquest cas es confronta una variable inicial (els tres possibles valors de l’espècie) amb una variable obtinguda a l’assignar un dels tres grups per HCPC o k-means. Per obtenir % d’encerts més fiables s’usen tècniques de dividir les dades en una part d’aprenentatge i una altra de test on avaluar els encerts.

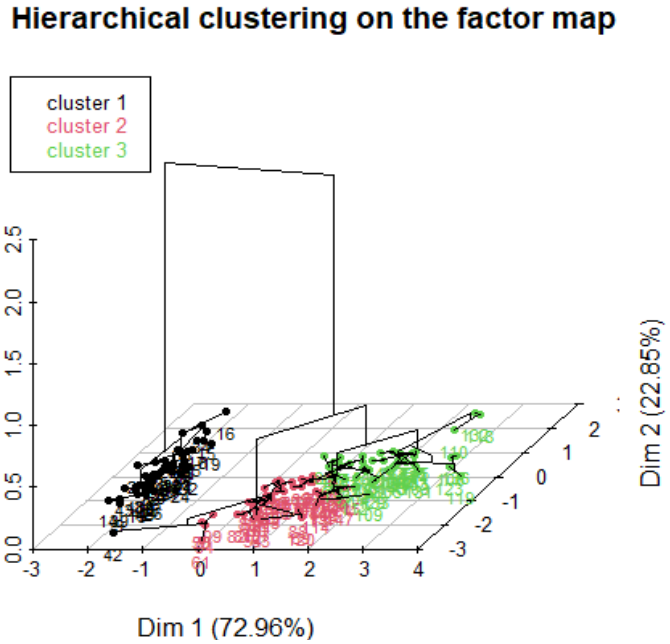
Ex "iris" (PCA i HCPC)

```
library(FactoMineR)
data(iris)
PCA(iris,quali.sup=c(5: 5))
```






```
pca.iris <- PCA(iris,quali.sup=c(5: 5))
HCPC(pca.iris)
```




```
hcpc.iris <- HCPC(pca.iris)
table(hcpc.iris$data.clust[,ncol(hcpc.iris$data.clust)],iris$Species)
```

	setosa	versicolor	virginica	
1	50	0	0	50
2	0	39	14	53
3	0	11	36	47
	50	50	50	

(50+39+36=125 de 150 encerts -> 83.3% d'encert)

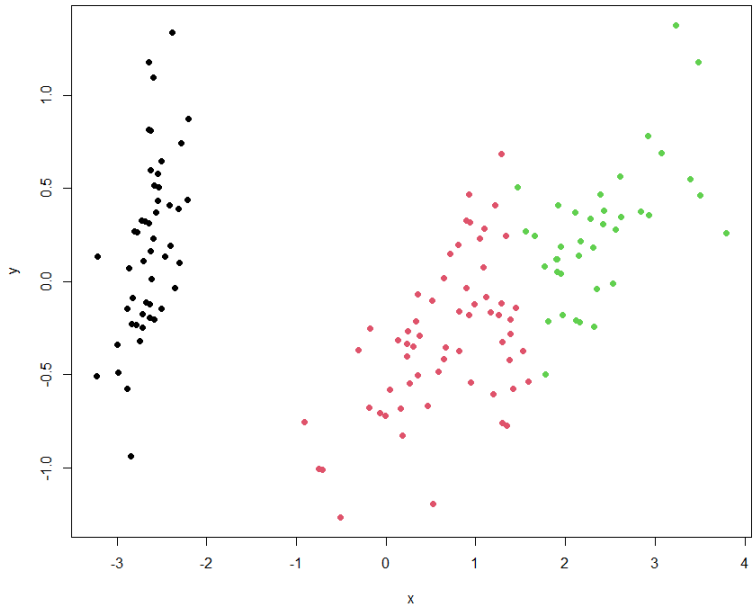
Ex “iris” (k-means)

```
km.iris <- kmeans(iris2,centers=3,nstart=10)
table(km.iris$cluster,iris$Species)
```

	setosa	versicolor	virginica	
1	50	0	0	50
2	0	48	14	53
3	0	2	36	38
	50	50	50	

(50+48+36=134 de 150 encerts -> 89.3% d’encert)
(el resultat pot dependre dels centres inicials i de les repeticions)

```
pr.comp <- princomp(iris2)
x <- pr.comp$scores[,1]
y <- pr.comp$scores[,2]
plot(x,y,pch=19,col=km.iris$cluster)
```



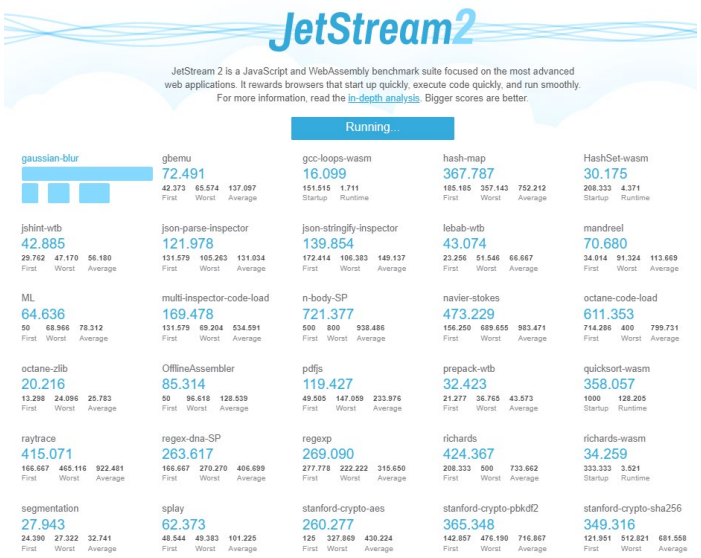
Exemple. Estudi Web Benchmark JetStream 2.0

JetStream és un índex que mesura el rendiment d'un navigador web, en base a 64 subtests orientats a diferents *workloads*.

Objectiu inicial: comparar navigadors (Edge vs Google Chrome)

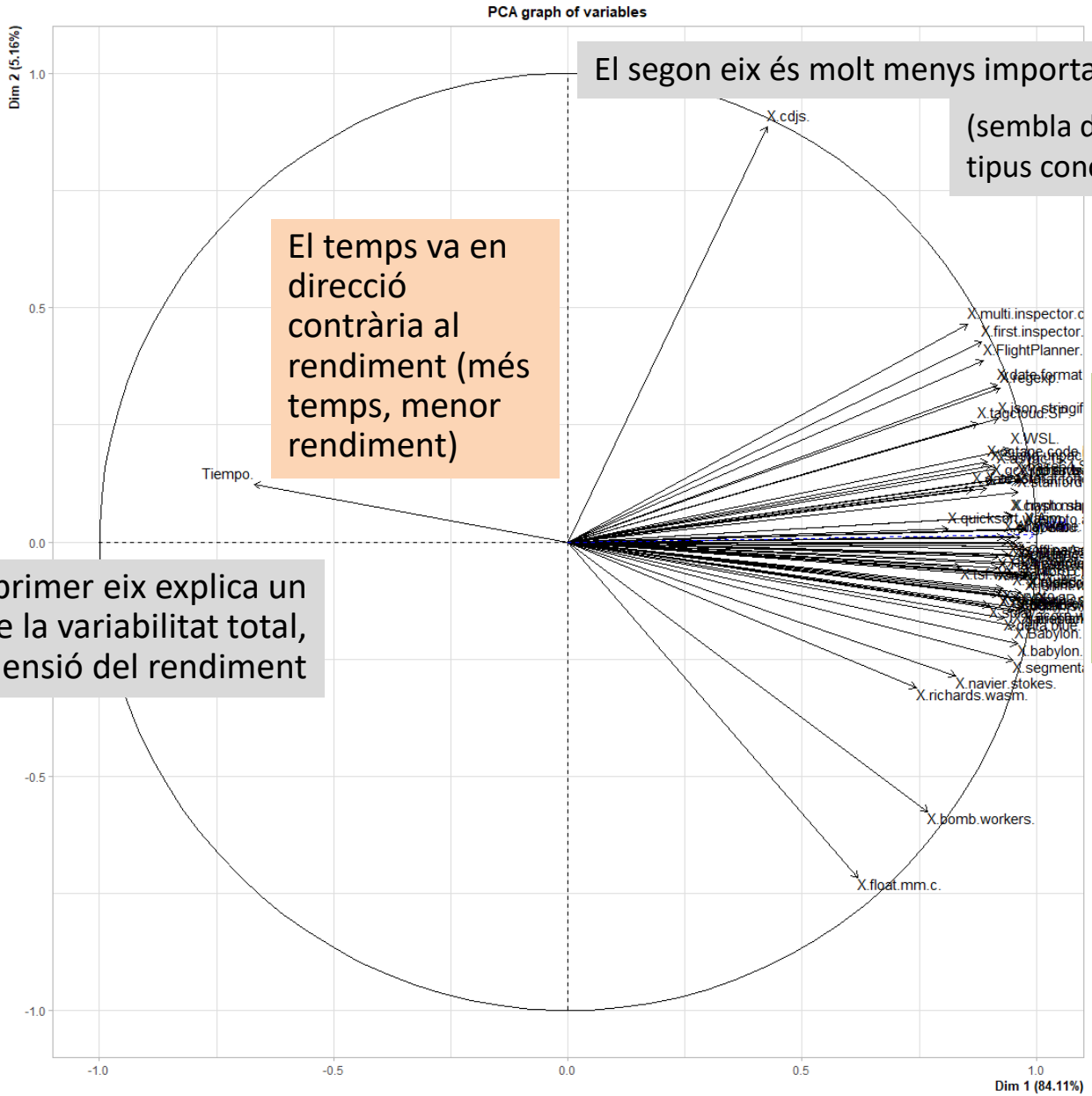
Tres persones utilitzen el benchmark en tres SO diferents (Windows, Linux, iOS). Compte, hardware diferent.

Anàlisi post hoc per estudiar relacions entre els 64 subtests, per **Components Principals** (PCA).




```
load(url("https://www-eio.upc.edu/teaching/pe/DADES/jetstream.Rdata"))
dim(C)
[1] 60 70

library(FactoMineR)
pca = PCA(C[,-c(3:4)], quali.sup=1:2, quanti.sup=3)
```



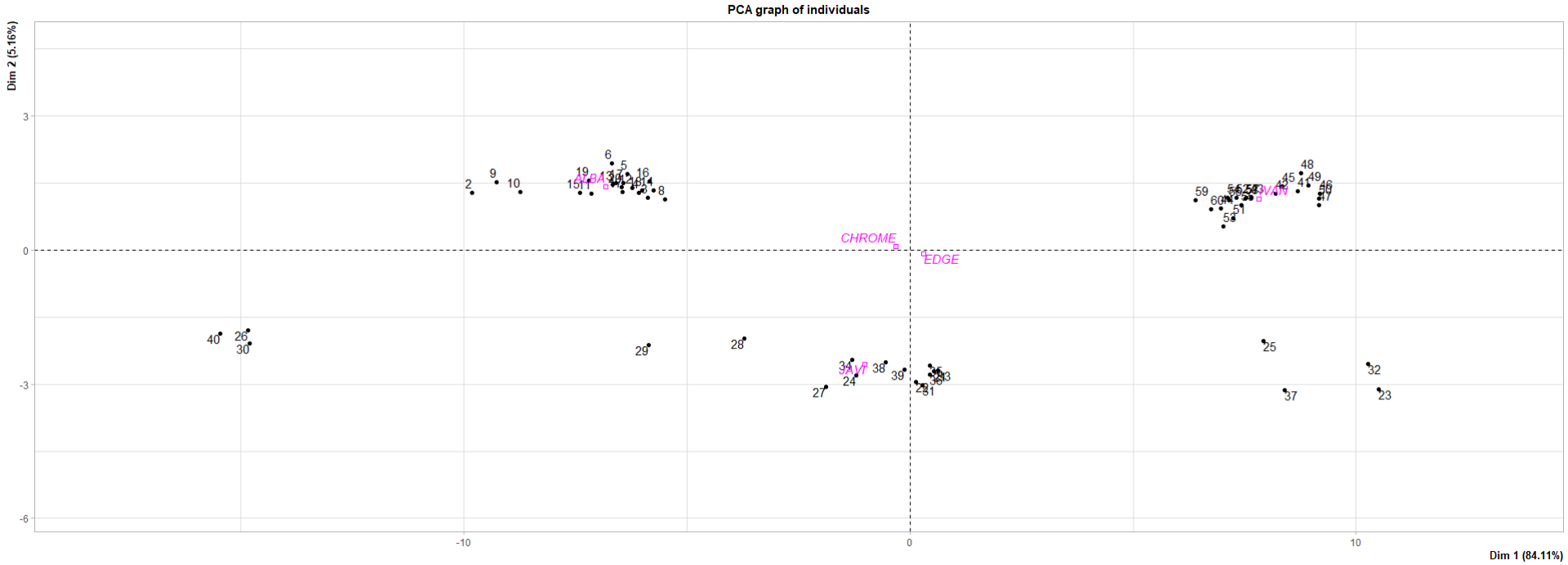
El segon eix és molt menys important. Explica un 5.16%

(sembla distingir entre alguns tipus concrets de test)

El temps va en direcció contrària al rendiment (més temps, menor rendiment)

La major part dels tests són redundants: no aporten informació nova realment

El primer eix explica un 84.11% de la variabilitat total, i és la dimensió del rendiment



Sobre el pla factorial de les dues primeres components hem projectat els individus (les mesures individuals) i dos factors: el tipus de navegador, i el SO (usuari, hardware i SO es confonen).

No és un anàlisi inferencial, però podem apreciar: 1) poques diferències entre Chrome i Edge, 2) Linux (Iván) millor rendiment que la resta, 3) Windows (Javi), molta variabilitat.