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

2023

Índex

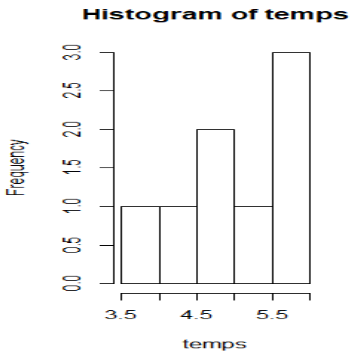
Exemple. Compressor

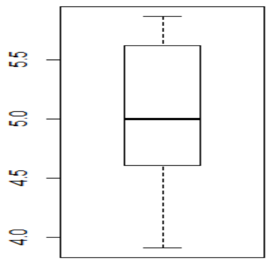
El fabricant d'un determinat compressor d'arxius assegura que arxius de 500Kb són comprimits en 5 segons. Aquest compressor ha estat testejat amb 8 arxius i el temps en segons necessari per a cada compressió han estat : 4.85,4.36,5.12,5.64,5.6,5.87,3.91,4.88

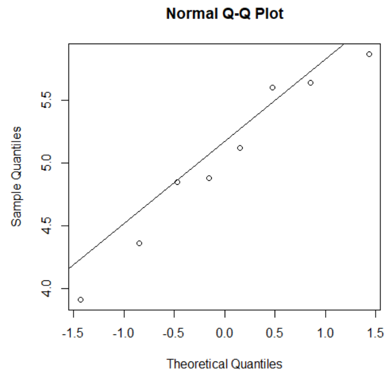
```
R: temps <- c(4.85,4.36,5.12,5.64,5.6,5.87,3.91,4.88)
```

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

PH H_0: μ=5 H_1: $\mu \neq 5$ estadístic: (**5.0287**-5) / 0.2379 = **0.12** (p-value $P(|t_7|>0.12)$= pt(-0.12,7)+(1-pt(0.12,7)) = **0.908)**

5 cau dins del IC (el p-value és superior a un risc del 5%), per tant 5 segons de mitjana és un valor versemblant, la diferència entre la mitjana mostral i l'esperada és deguda a 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)

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

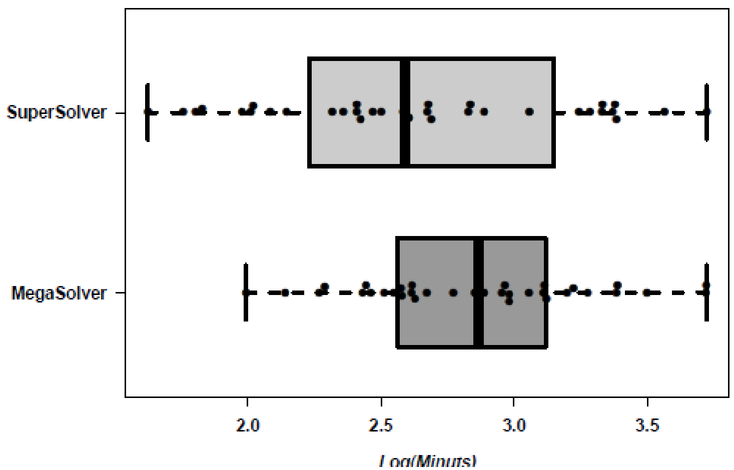
0.22 cau dins del IC, per tant té una qualitat suficient

Exercici. Temps optimitzadors (Parcial 2 Primav 2012)

Per comparar la velocitat amb la qual resolen dos servidors diferents, *SuperSolver* i *MegaSolver*, problemes d'optimització s'envia un total de 70 problemes de maximització diferents als dos servidors, 35 a cadascun. Pel fet que el temps que triguen els servidors per resoldre els problemes, és asimètrica cap a la dreta, treballem a continuació amb els logaritmes dels temps. Siguin X el logaritme del temps que triga el *SuperSolver* i Y el del *MegaSolver* .

Els valors descriptius a cada mostra són els següents i a més a més es mostra una representació gràfica:

	Mitjana	Mediana	Desv. est.	Mínim	Màxim
SuperSolver	2,63	2,59	0,57	1,63	3,72
MegaSolver	2,85	2,86	0,44	1,99	3,72



Exemple. Temps algoritme Dijkstra

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

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

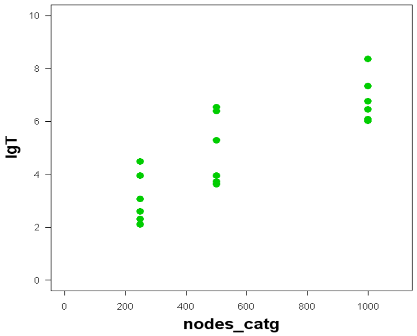
x_i (nodes)	y_i (lgt)
250	2.31
250	4.48
250	2.59
250	3.06
250	2.10
250	3.95
500	3.94
500	6.38
500	6.52
500	5.27
500	3.72
500	3.61
1000	6.45
1000	7.32
1000	6.76
1000	6.08
1000	8.35
1000	6.01

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)             3.082      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

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)

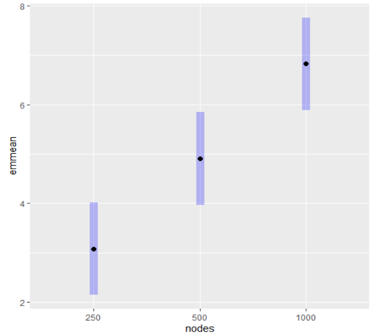
Per tant, el canvi en mitjana en el temps entre grafs de 250 i 500 és exp(1.825)= $e^{1.825}$ = 6.2

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 (preordre,inordre,postordre)

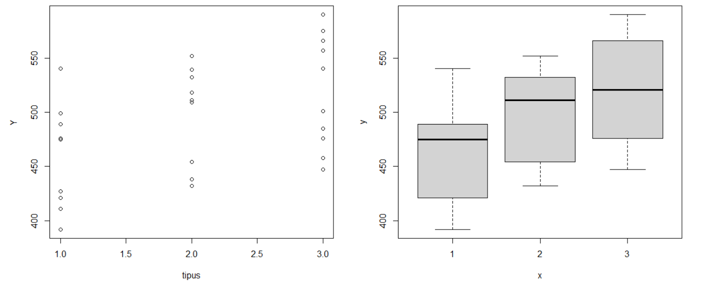
```
Temps <- c(392,421,540,475,427,411,476,489,499,454,509,432,539,552,518,511,438,532,447,590,566,557,540,501,575,458,476,485)
metode <- c(1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,2,3,3,3,3,3,3,3,3,3,3)
nodes <- c(100,140,200,160,120,130,170,180,190,150,160,100,210,200,180,170,140,190,120,190,180,170,160,150,200,110,130,140)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

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
```

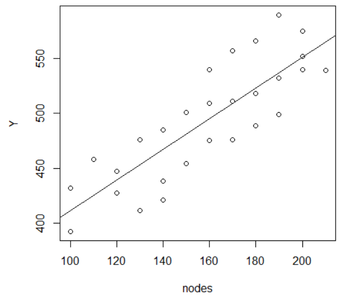


Aquest primer model només explica un 22.92 % de la variabilitat

```
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. Recorregut arbre (preordre,inordre,postordre)

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 ±qt(0.975,24)*15.88211 -> [207.001, 272,56]

L'estimació del pendent és 1.42 amb IC 1.42±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±qt(0.975,24)*7.56 -> [6.5, 37.71]

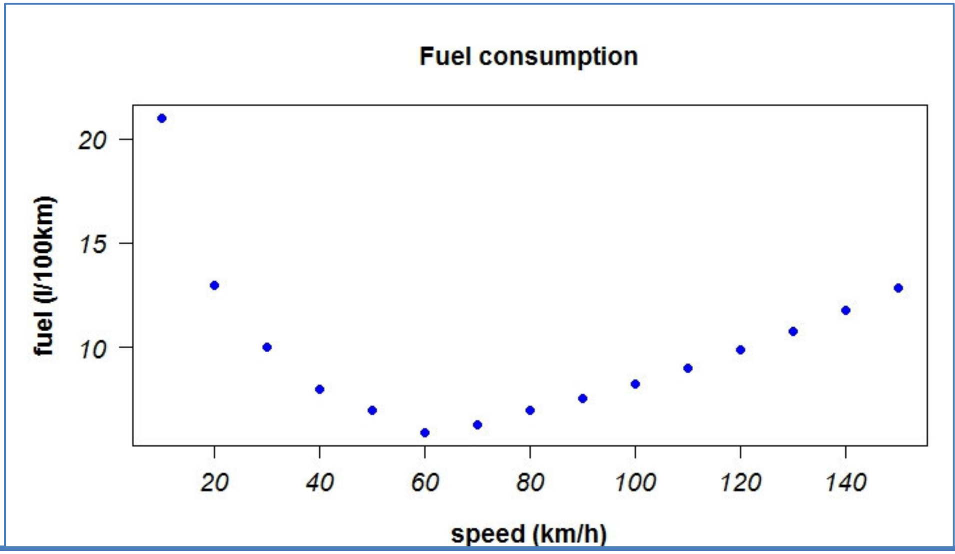
L'estimació del canvi de la mitjana del 1r al 3r mètode és 59.823 amb IC 59.823±qt(0.975,24)*7.27785 -> [44.8, 74.84]

Exemple. Benzina i velocitat (model lineal)

- 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.



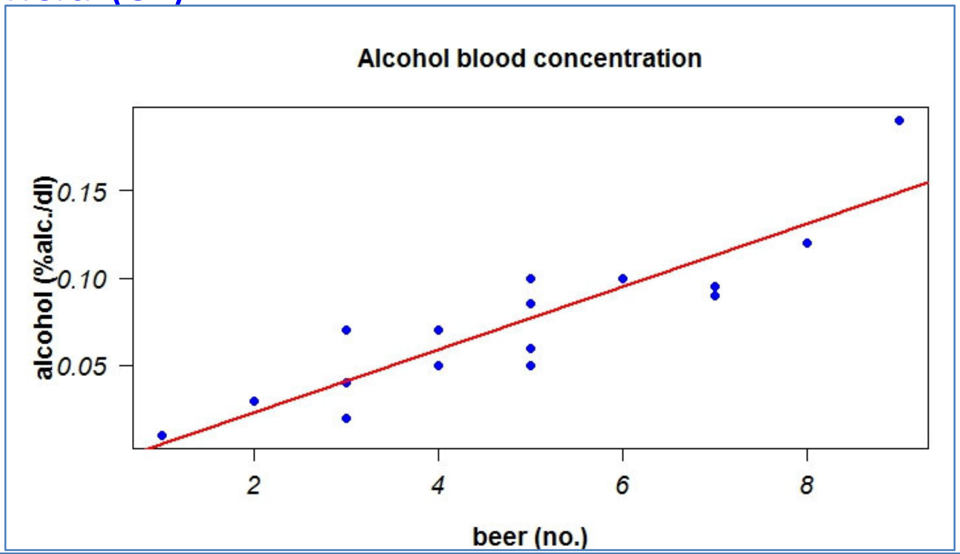
speed (km/h)	fuel (l/100 km)
10	21
20	13
30	10
40	8
50	7
60	5.9
70	6.3
80	6.95
90	7.57
100	8.27
110	9.03
120	9.87
130	10.79
140	11.77
150	12.83

Exemple. Cervesa i alcohol (model lineal)

- 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)*

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

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



Exemple. Cervesa i alcohol (model lineal)

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:

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

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. Cervesa i alcohol (model lineal)

```
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)

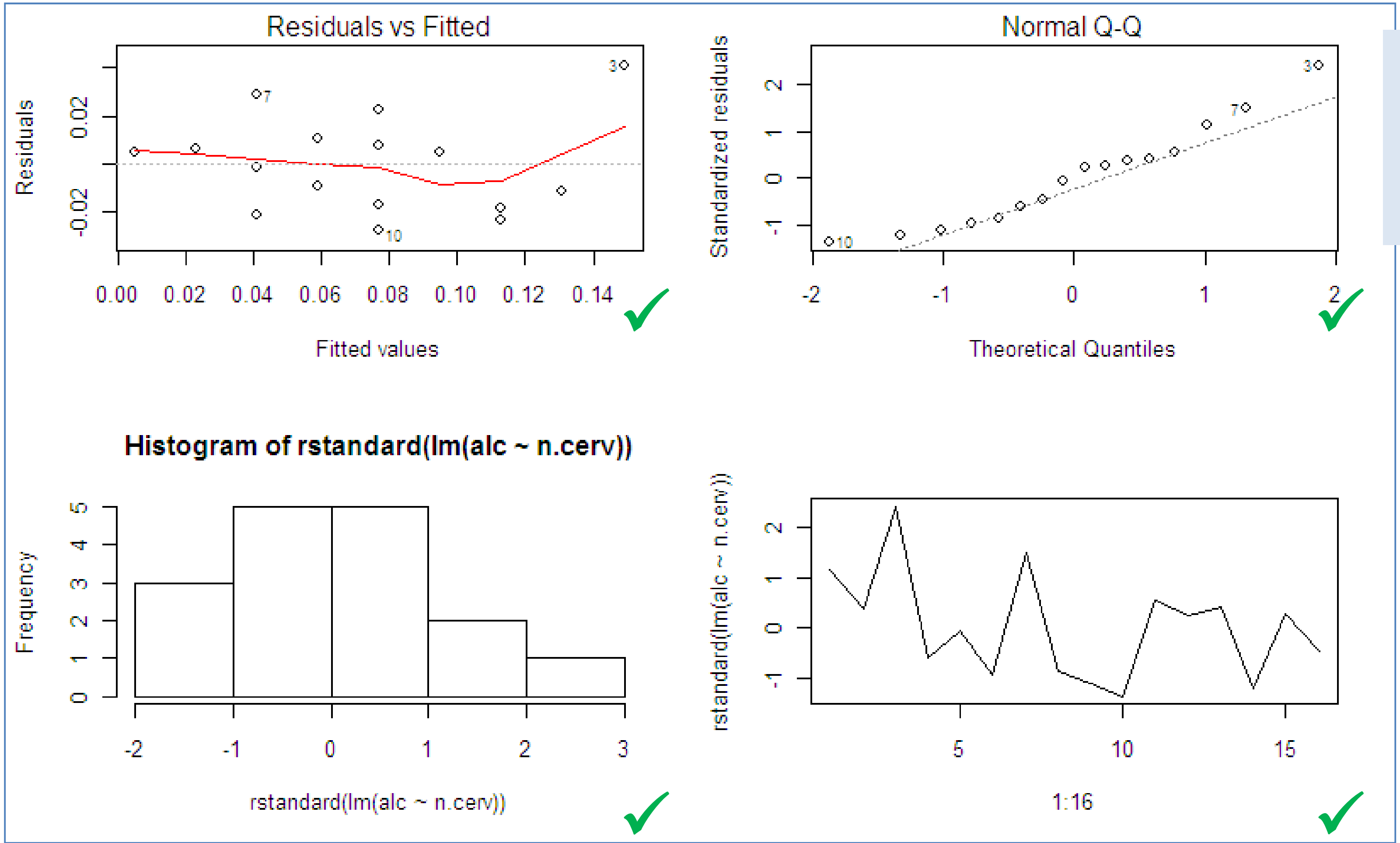
Conclusió pràctica: No és versemblant que el coeficient del pendent sigui 0

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. Cervesa i alcohol (model lineal, 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, res s'oposa a validar cap de les 4 premisses

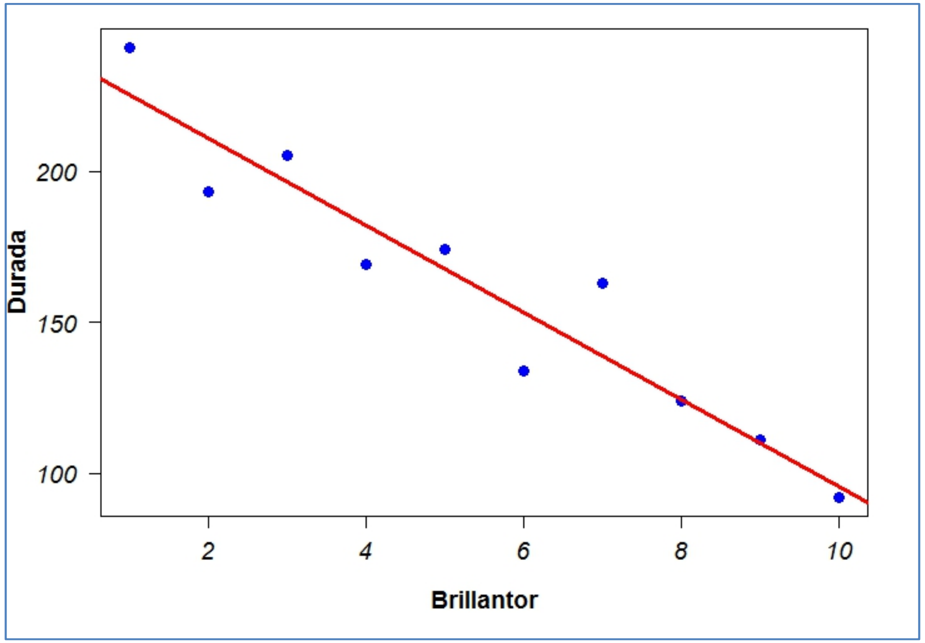
Exemple. Brillantor i durada (model lineal)

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. Brillantor i durada (model lineal)

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

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

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

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

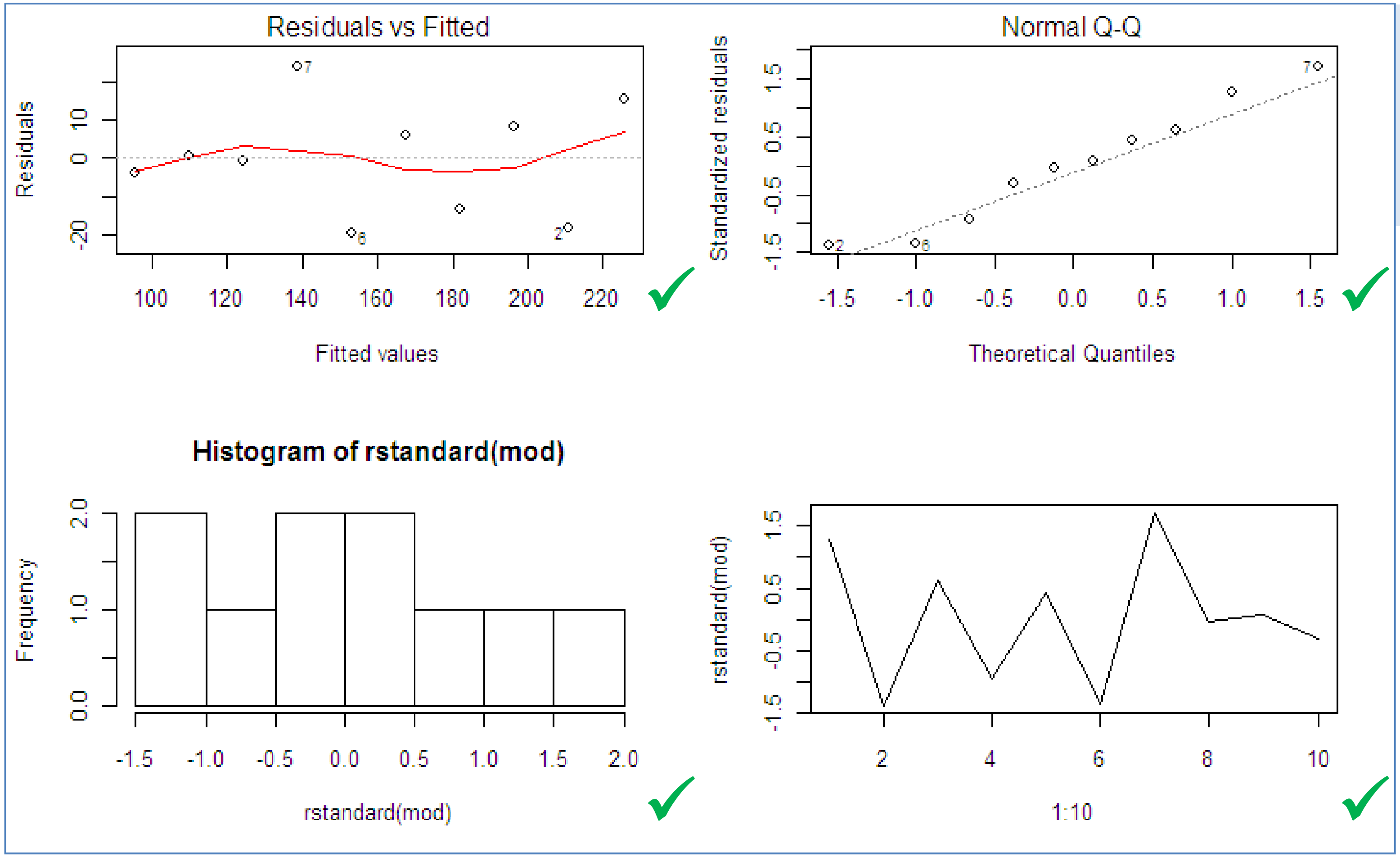
(Cada grau que pugem la brillantor de la pantalla significa entre uns 10 i 18 minuts menys de durada de la bateria)

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. Brillantor i durada (model lineal, 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, res s'oposa a validar cap de les 4 premisses

Exemple. Brillantor i durada (model lineal, 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:

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

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

Quina durada podem esperar per a pantalles de brillantor 7.5?

	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. Modem (model lineal, predicció)

```
> modem$Tam1Mb
[1] 1.59129 1.59129 0.51858 1.29297 0.14062 0.22461 0.66895 2.68000
> modem$Tpo1Mb
[1] 23.22 14.56  6.07 13.50  1.38  2.24  5.95 23.45
> mod1 = lm(Tpo1Mb ~ Tam1Mb, data=modem)
> summary(mod1)
...
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.908      1.962    0.46  0.65995
Tam1Mb         9.544      1.447    6.59  0.00058 ***
> modem$Log.tam1mb = log(modem$Tam1Mb)
> mod2 = lm(log(Tpo1Mb) ~ Log.tam1mb, data=modem)
> summary(mod2)
...
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.3322     0.0679    34.3  4.1e-08 ***
Log.tam1mb    1.0083     0.0673    15.0  5.6e-06 ***
> predict(mod2, int="prediction")
  fit       lwr      upr
1 2.80061  2.30739 3.29384
2 2.80061  2.30739 3.29384
3 1.67006  1.18913 2.15100
...
```

$\bar{x} = -0.293$
$s_x^2 = 1.065$
$s^2 = 0.0338$