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

Previsió

Bloc 6 – Probabilitat i Estadística

Novembre 2021

Índex

1. Previsió i disseny d'experiments
2. Fases del procés de models estadístics
3. Model "quantitativa vs quantitativa": model, paràmetres i interpretació
4. Estimadors dels paràmetres: distribució, inferència
5. Anàlisi de les premisses. Anàlisi de residus
6. Predicció
7. Model "quantitativa vs categòrica":
   a. model, paràmetres i interpretació
   b. descomposició de la variabilitat

Inferència estadística. Guió

Guió de la part d'Estadística de PE:

- B4: Tècnica general de la inferència [estadística]
  - estimar un paràmetre (*Intervals de Confiança*)
  - refutar un paràmetre (*Proves d'Hipòtesis*)
- B5: Aplicació (I): Avaluació de millores
  - *Disseny d'experiments:* comparació de dues poblacions.
- **B6: Aplicació (II): Predicció**
  - ***Previsió* d'una var. resposta, en funció d'una var. explicativa.**

Previsió i disseny d'experiments

- Al B5 parlem de variables i condicions; i es defineix el disseny d'experiments com: "estimar l'efecte causal de la intervenció X en la resposta Y donades les condicions Z"
  - La resposta Y ha de mesurar el nostre objectiu
  - La intervenció X és el nostre potencial per canviar el futur
  - Les condicions Z 'predeterminen' el futur i permeten anticipar Y

- **Tipus d'estudis: *VEURE*** enfront de ***FER***:
  - Estudis observacionals: ***veiem*** i podem fer previsions, predir, anticipar,...Els individus arriben amb el valor de Z, que relacionem amb la resposta Y [Ex: comparem les notes de PE (Y) en funció del gènere (Z)]
  - Estudis experimentals: ***fem*** i podem intervenir, canviar el futur. Observem l'efecte en Y havent assignat X a les unitats [Ex: comparem les notes de PE (Y) en un experiment on a uns alumnes se'ls ha assignat emprar e-status i els altres no (X)]

Nota 1: La clau per intervenir és ser 'propietaris' de la variable X

Nota 2: El passat (Z) ens esclavitza, el futur (X) ens allibera

Previsió i disseny d'experiments

Per respondre una pregunta '**causal**' sobre una condició Z, hem de pensar un experiment on '**assignar**' aquesta condició Z.

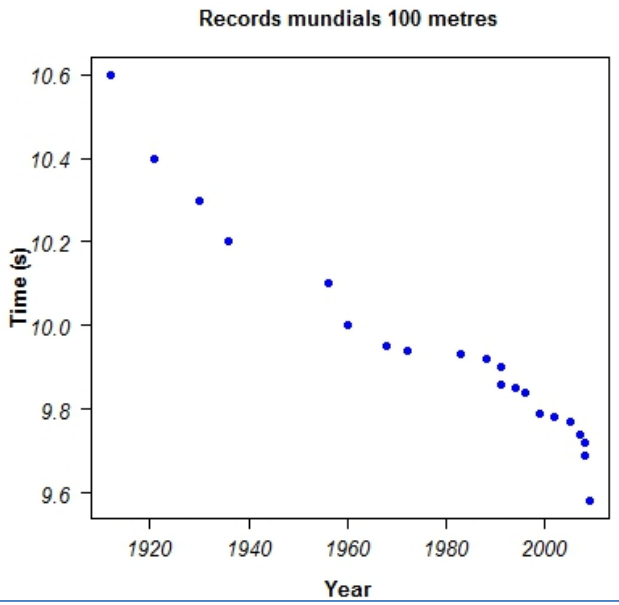
Exemple: per respondre si hi ha discriminació per gènere, podem 'assignar' a l'atzar un nom i una foto de dona/home a uns currículums i preguntar quin salari els hi pagarien. [Això permet deixar fixes o iguals ('controlar') totes les altres variables: experiència, dedicació, formació,...]. Així podríem estimar l'efecte de ser dona/home en el salari. Però, en el futur, no podrem 'assignar' el gènere a un ciutadà.

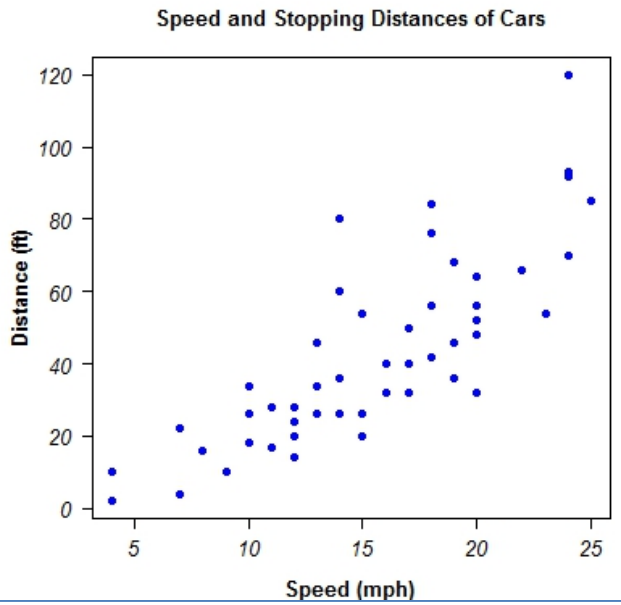
Resum:

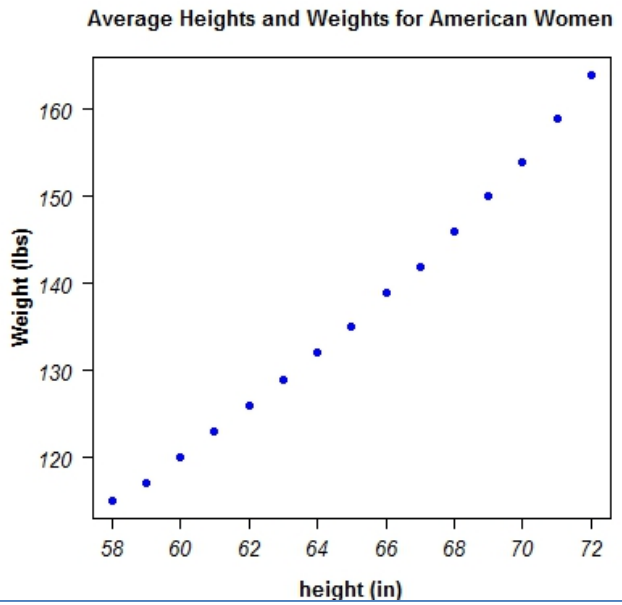
1) Un **experiment** amb **assignació** a l'atzar permet estimar '**efectes**' havent controlat totes les altres variables.

2) Convé valorar la possibilitat **d'assignar** en el futur per saber si podem utilitzar la relació només per **predir** o també per **intervenir**

Previsió i disseny d'experiments







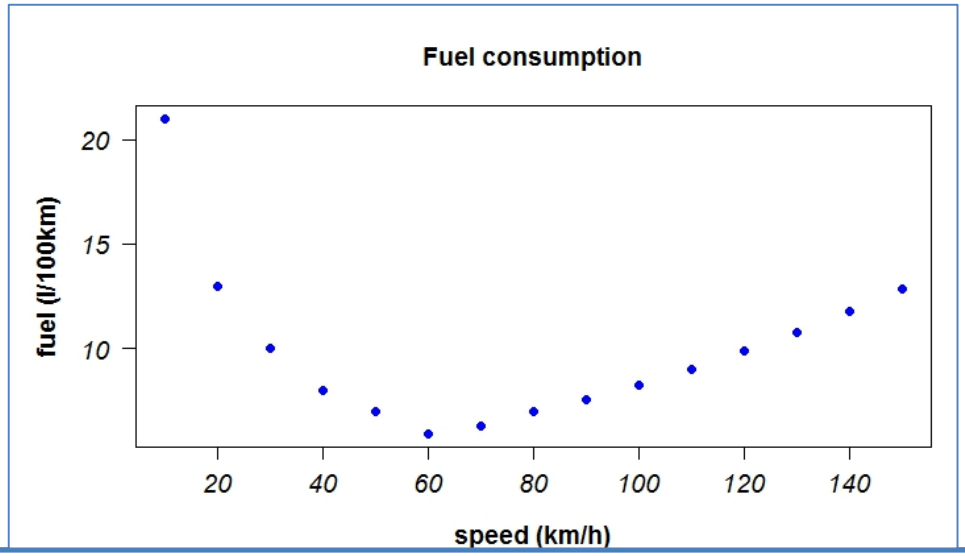
- Temps dels records mundials de 100 metres masculí: 1912-2012. Sempre serà decreixent. Previsió o Intervenció?
- Distància de frenada en funció de la velocitat. Previsió o Intervenció?
- Pes en funció d'alçada en dones. Previsió o Intervenció?

Model quantitativa vs. quantitativa. Exemple I

- 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.

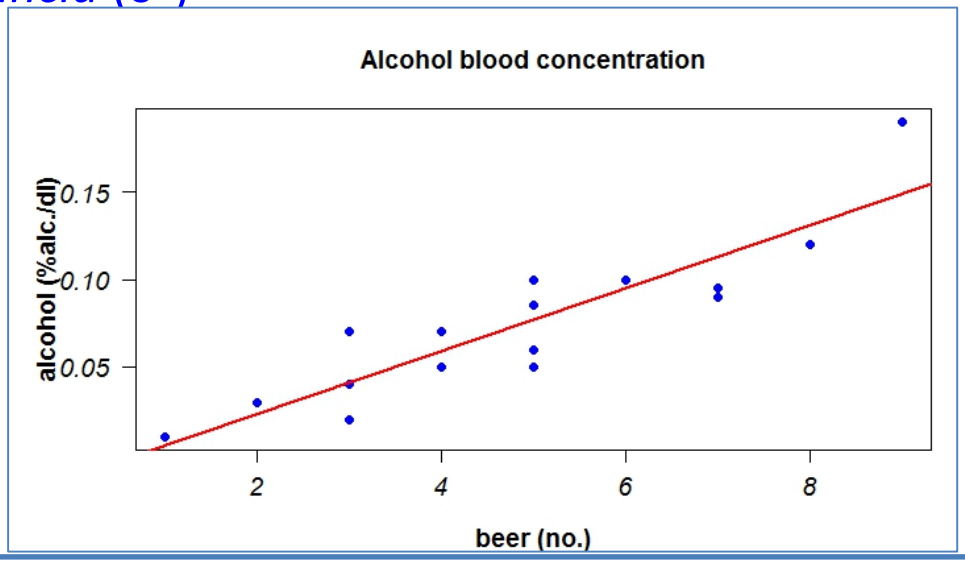


speed (km/h)	fuel (l/100 km)
10	21
20	13
30	10
40	8
50	7
60	5.9
70	6.3
80	6.95
90	7.57
100	8.27
110	9.03
120	9.87
130	10.79
140	11.77
150	12.83

Model quantitativa vs. quantitativa. Exemple II

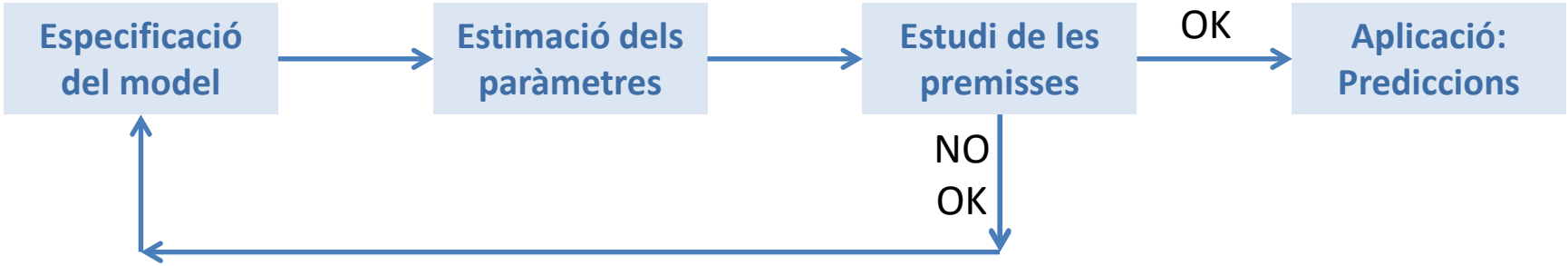
- Un estudi ha sol·licitat a 16 voluntaris que es prengui una quantitat determinada (aleatòriament) de cervesa, mesurada en llaunes, i es mesura l'alcohol a la sang trenta minuts després [%alc. /dl sang].
- Un model simple és ajustar-hi una recta, que implica dos paràmetres: *pendent* (β_1) i *constant* (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

Fases en el disseny d'un model



Font: Capítol 7 d'*Estadística per a enginyers informàtics*. Ed UPC

- Un cop especificat el model i estimats els paràmetres, perquè sigui útil (aplicar-lo i fer prediccions), cal estudiar les premisses assumides. Serà suficient una anàlisi exploratòria per confirmar que són "raonables"
- Si durant el procés de modelar, no s'aconsegueix trobar els resultats desitjats, pot ser que el model sigui millorable. En aquest cas, podem procedir a realitzar **transformacions** (ln(X), ln(Y), 1/Y, Y/X, arrels, potències,...) o buscar **altres variables predictores**

Model quantitativa vs. quantitativa. Paràmetres

- Model: $Y_i = \beta_0 + \beta_1 \cdot X_i + \varepsilon_i$ **sent** $\varepsilon_i \sim N(0, \sigma)$

  Y_i valor de la variable resposta Y en el cas i-èsim

  X_i valor que pren la condició X en el cas i-èsim

  ε_i error aleatori o distància a la recta del cas i-èsim

- Els paràmetres seran: β_0 com a **constant** a l'origen, β_1 com a **pendent** de la recta i i σ^2 com la **variància dels** ε_i o variància residual ($\beta_0+\beta_1 \cdot X_i$ serà la part determinista de Y; i ε_i serà la part aleatòria de Y)

EXEMPLE: (*Estadística per a enginyers informàtics*. Ed UPC pàg 141)**.** Homes adults i sans de Barcelona: *Y* és Pes en Kg; *X* és Alçada en cm. Suposem com a model una recta amb paràmetres:

β_0 = -100 Kg β_1 = +1 Kg/cm σ = 6Kg

Quin pes correspon a un senyor de 160 cm? 60Kg

I a un de 180 cm? 80 kg

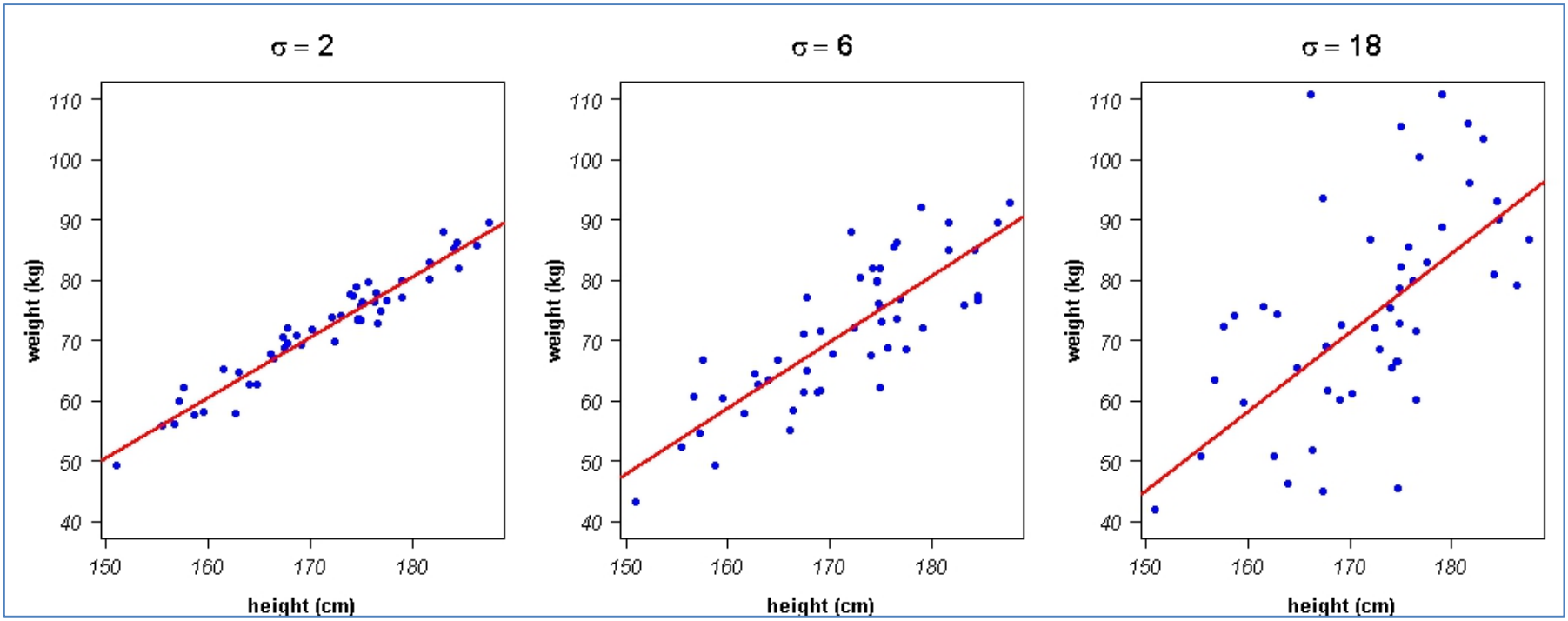
Què significa "correspon"? 'Esperat, en mitjana'

Què significa σ = 6Kg ? *Separar-se uns 6 kg del pes esperat és habitual. (Rarament és més del doble)*

Què opines de l'etiqueta 'pes ideal' en algunes farmàcies? Que ignoren la variabilitat natural

Model i Paràmetres

- El **paràmetre** més important per un estadístic és la variància σ^2 (encara que σ és més fàcil d'interpretar).
- Diferents valors de **σ** condicionaran la forma del núvol de punts



- Noms possibles per **ε**: *negatius* → error, residu, pertorbació

  positius → idiosincràsia

Estimació dels paràmetres

- β_0 , β_1 i σ^2 són valors poblacionals, *autèntics*, desconeguts, a 'estimar'. L'estimació dels dos primers, dóna lloc a la recta estimada:

$$\hat{y}_i = b_0 + b_1 \cdot X_i$$

- Aquesta permet fer prediccions per a cada observació amb el seu error de predicció:

$$e_i = y_i - \hat{y}_i \quad \text{[els } e_i \text{ són els residus del model]}$$

- L'estimació mínim quadràtica consisteix en calcular els estimadors $\mathbf{b_0}$ i $\mathbf{b_1}$ de β_0 i β_1, minimitzant la suma dels errors de predicció al quadrat: $\sum(e_i)^2 = \sum(y_i - \hat{y}_i)^2 = \sum(y_i - b_0 - b_1 x_i)^2$ [*annex 6.12 d'Estadística per a enginyers informàtics. Ed UPC*]
- La solució al problema de minimització és el següent: [*Ref: Eei.Ed.UPC pg144*]

$$\hat{\beta}_1 = b_1 = \frac{S_{XY}}{S_X^2} = r \cdot \frac{S_Y}{S_X}$$

$$\hat{\beta}_0 = b_0 = \bar{Y} - b_1 \bar{X}$$

$$\hat{\sigma}^2 = S^2 = \frac{\sum(e_i^2)}{n-2} = \frac{(n-1)S_Y^2(1-r^2)}{n-2} = \frac{(n-1)(S_Y^2 - b_1 S_{XY})}{n-2}$$

[Recordeu que $\mathbf{S_{XY}}$ és la covariància mostral, i $\mathbf{r}=\mathbf{r_{XY}}$ la correlació mostral]

Estimació dels paràmetres. Exemple

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

Recordatori:

$$\bar{y} = \frac{\sum y_i}{n} \qquad s_Y^2 = \frac{\sum y_i^2 - \frac{(\sum y_i)^2}{n}}{n-1} \qquad s_{XY} = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{n-1}$$

Càlculs dels estadístics convencionals:

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

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

Resultats de la regressió:

$$b_1 = \frac{s_{XY}}{s_X^2} = r_{XY} \cdot \frac{s_Y}{s_X} = 0.01796$$

$$b_0 = \bar{Y} - b_1 \bar{X} = -0.0127$$

$$S = \sqrt{\frac{\sum(e_i^2)}{n-2}} = 0.0204$$

Model amb R:

```
> lm(alc ~ n.cerv)
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
```

Interpretació dels paràmetres

- Els **paràmetres** de la recta han de ser interpretats d'acord amb les seves unitats.
- El **pendent** s'interpreta directament com a tal:
  - Experiments: La resposta Y tindrà un canvi esperat de β_1 (unitats de Y) per cada increment de 1 unitat fet en la causa X.
  - Previsió: Una variació de 1 unitat en la variable X s'associa amb una variació de β_1 unitats en la variable Y.
- La **variància residual** s'interpreta:
  - Experiments: Variabilitat de la variable Y.
  - Previsió: Error de predicció de la variable Y, conegut el valor de X.
- La **constant**, en certs casos, es pot interpretar com el valor que pren la resposta en absència de la variable predictora. [La constant és necessària per construir el model, però secundària en sí mateixa]

Interpretació dels paràmetres. Exemple

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?

Interpretació dels paràmetres. Exemple (cont)

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

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

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

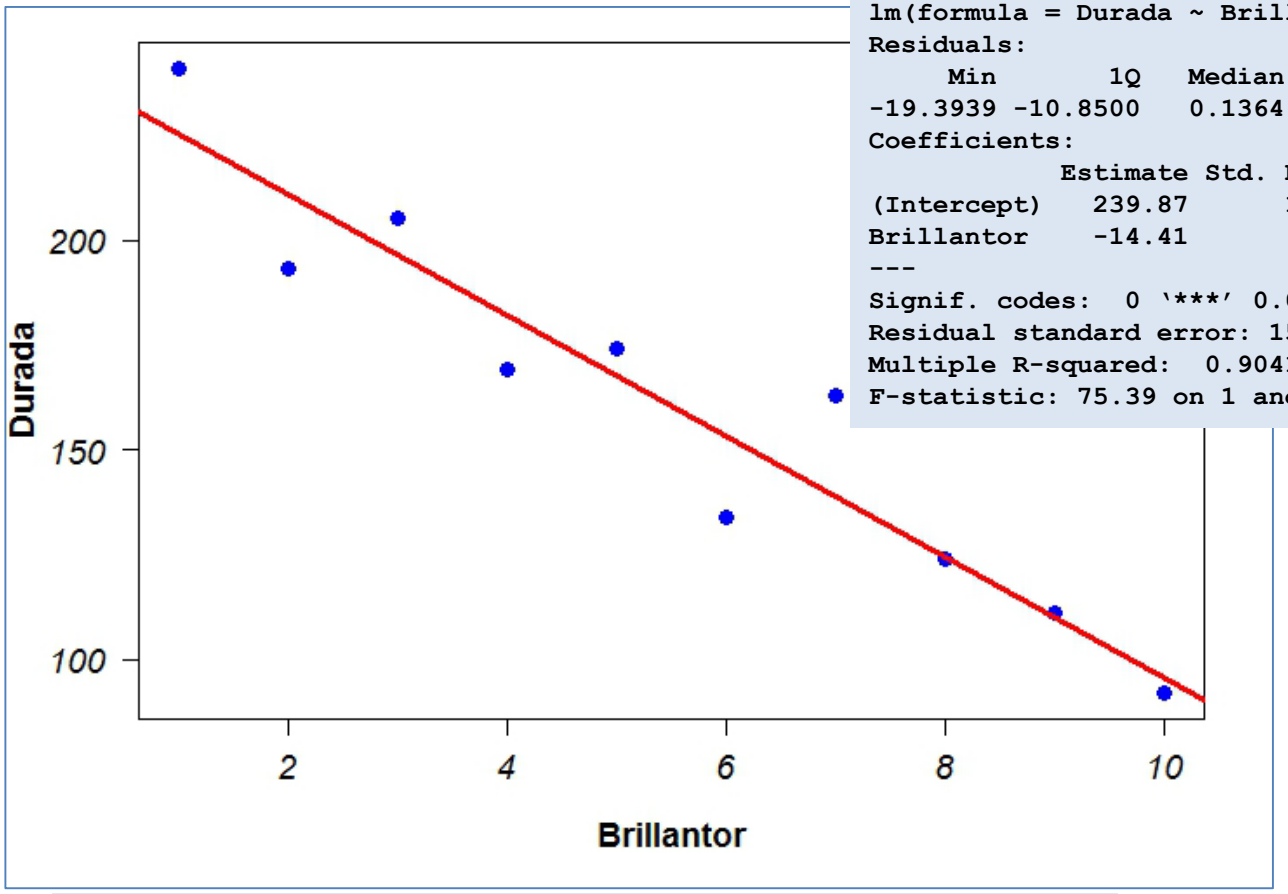
Interpretació de b_1: Per cada grau de brillantor que augmentem, la bateria dura uns 14.4 minuts menys.

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

Interpretació dels paràmetres. Exemple (cont)

```
> datos <- read.table("clipboard",header=TRUE)
> mod.lm <- lm(Durada~Brillantor,datos)
> summary(mod.lm)
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
```




```
> par(cex.lab=1.2,cex.axis=1.2,las=1,font.lab=2,font.axis=3)
> plot(Durada~Brillantor,datos,pch=19,col=4,cex=1.2)
> abline(mod.lm,col=2,lwd=3)
```

Distribució dels estimadors (mínims quadrats)

- b_1 és una combinació lineal de normals i, per tant, continuarà seguint una distribució Normal. Així, la distribució de l'estimador b_1 és:

$$b_1 \sim N\left(\beta_1, \frac{\sigma^2}{(n-1)S_X^2}\right)$$

- b_0 també és una combinació lineal de normals. Així, la distribució de l'estimador b_0 és:

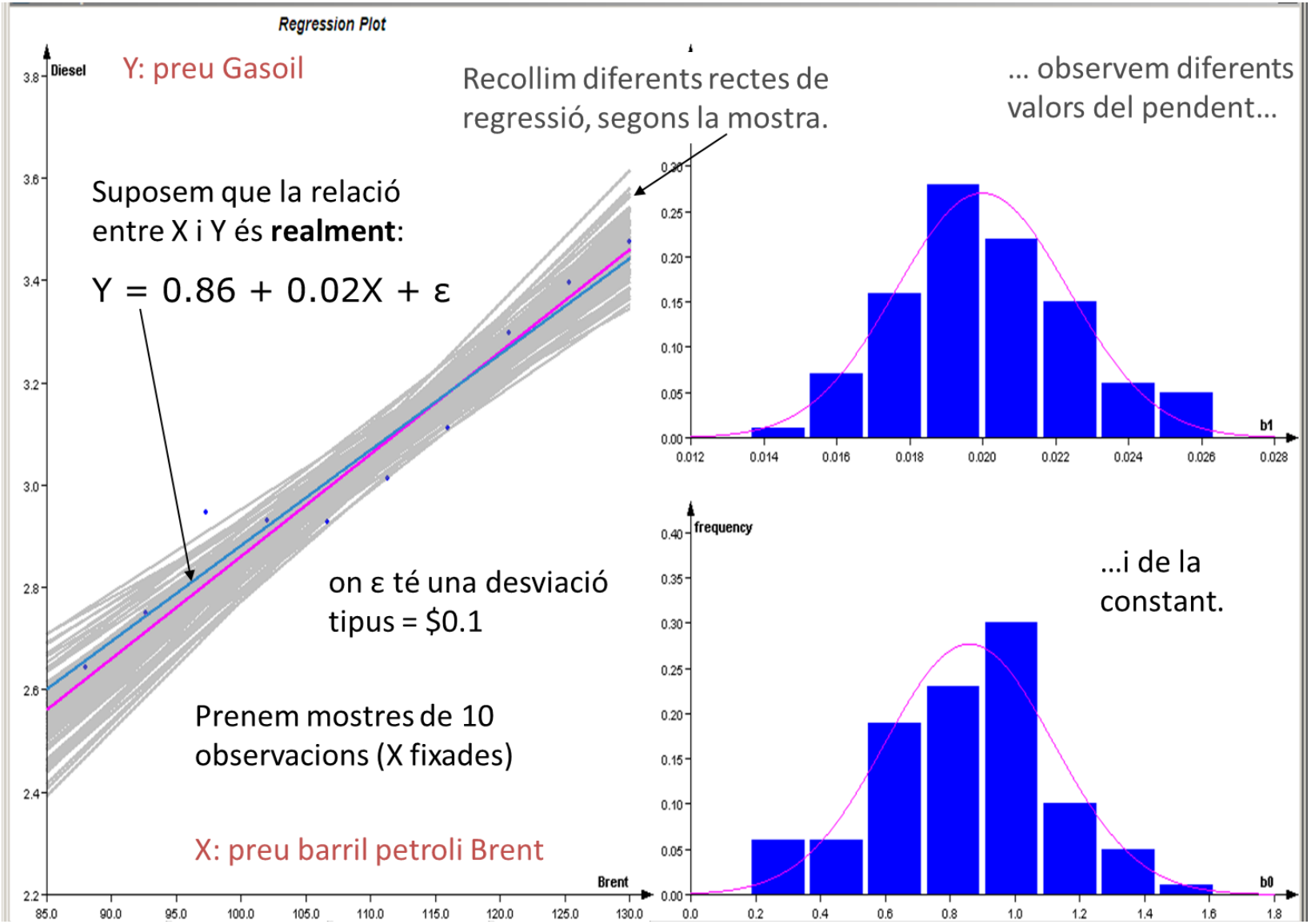
$$b_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{X}^2}{(n-1)S_X^2}\right)\right)$$

- $S^2 = \frac{\sum(e_i^2)}{n-2}$ és estimador no esbiaixat de σ^2, i coneixem que $\frac{\sum(e_i^2)}{\sigma^2} \sim \chi^2_{n-2}$. Així, la distribució de referència de la variància residual és:

$$\frac{(n-2)S^2}{\sigma^2} \sim \chi^2_{n-2}$$

Feu clic aquí per veure apps per entendre distribució dels estimadors

Distribució dels estimadors. Simulació



Distribució dels estimadors. Disminuir S_{b1}

Què es pot fer si es vol millorar una recollida de dades on s'ha observat un error estàndard massa gran en l'estimació b_1 de β_1 ?

$$S_{b_1} = \sqrt{\frac{S^2}{(n-1)S_X^2}}$$

Solucions per disminuir S_{b_1} :

- Intentar "controlar" les fonts de variació en S^2
- Augmentar la "n"
- Ampliar la "finestra" de l'estudi per augmentar $\sum(x_i - \bar{x})^2$

Distribució dels estimadors. Inferència

Es pot realitzar la inferència habitual amb PH per β_0, β_1 i σ^2:

- **Prova d'hipòtesi sobre el pendent:**

$$\begin{cases} H_0: \beta_1 = \beta_1' \\ H_1: \beta_1 \neq \beta_1' \end{cases} \rightarrow \frac{b_1 - \beta_1'}{S_{b_1}} = \frac{b_1 - \beta_1'}{\sqrt{\frac{S^2}{(n-1)S_x^2}}} \sim t_{n-2} \quad \text{sent} \quad \hat{\beta}_1 = b_1 = \frac{S_{XY}}{S_X^2}$$

No ens perdem amb la notació!!

$\boldsymbol{\beta_1}$: Paràmetre que volem contrastar

$\boldsymbol{\beta_1'}$: Valor a contrastar

$\boldsymbol{\hat{\beta}_1}$ o $\boldsymbol{b_1}$: estimador del paràmetre

- **Prova d'hipòtesi sobre el terme independent:**

$$\begin{cases} H_0: \beta_0 = \beta_0' \\ H_1: \beta_0 \neq \beta_0' \end{cases} \rightarrow \frac{b_0 - \beta_0'}{S_{b_0}} = \frac{b_0 - \beta_0'}{\sqrt{s^2\left(\frac{1}{n} + \frac{\bar{X}^2}{(n-1)S_x^2}\right)}} \sim t_{n-2} \quad sent \quad \hat{\beta}_0 = b_0 = \bar{Y} - b_1\bar{X}$$

- **Prova d'hipòtesi sobre la desviació tipus residual:**

$$\begin{cases} H_0: \sigma = \sigma_0 \\ H_1: \sigma \neq \sigma_0 \end{cases} \rightarrow \frac{(n-2)S^2}{\sigma_0^2} \sim \chi_{n-2}^2 \quad sent \quad \hat{\sigma} = S^2 = \frac{\sum(e_i^2)}{n-2}$$

Els graus de llibertat són **n-2** per les dues restriccions necessàries per estimar dos paràmetres previs

Formulari

Paràmetre	β_0	β_1	σ^2
Estimador	$b_0 = \bar{Y} - b_1\bar{X}$	$b_1 = s_{XY}/(s_X^2)$	$S^2 = \left(\sum(e_i^2)\right)/(n-2)$
Esperança	$E(b_0) = \beta_0$	$E(b_1) = \beta_1$	$E(S^2) = \sigma^2$
Variància	$S_{b_0}^2 = \sigma^2\left(\frac{1}{n} + \frac{\bar{X}^2}{(n-1)S_X^2}\right)$	$S_{b_1}^2 = \frac{\sigma^2}{(n-1)S_X^2}$	$V(S^2) = \frac{2\sigma^4}{n-2}$
Distribució	$b_0 \sim N$ $\frac{b_0 - \beta_0'}{S_{b_0}} \sim t_{n-2}$	$b_1 \sim N$ $\frac{b_1 - \beta_1'}{S_{b_1}} \sim t_{n-2}$	$\frac{(n-2)\cdot S^2}{\sigma^2} \sim \chi_{n-2}^2$
Interval de Confiança	$IC(95\%, \beta_0)$ $= b_1 \mp t_{n-2,0.975} \cdot S_{b_1}$	$IC(95\%, \beta_1)$ $= b_1 \mp t_{n-2,0.975} \cdot S_{b_1}$	$IC(95\%, \sigma^2)$ $= \left[\frac{(n-2)S^2}{\chi_{n-2,0.975}^2}, \frac{(n-2)S^2}{\chi_{n-2,0.025}^2}\right]$
H_0 usual	$\beta_0 = 0$	β_1=0	
Rebutgem H_0 si	$\frac{b_0}{S_{b_0}} > t_{n-2,0.975}$	$\frac{b_1}{S_{b_1}} > t_{n-2,0.975}$	

Prova d'hipòtesi sobre el pendent (β_1). Exemple

1. **Variables**: Cervesa i contingut d'alcohol a la sang

R: `lm`

```
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
```

2. **Estadístic**: $\hat{t} = \frac{b_1 - \beta_1'}{s_{b_1}} = \frac{b_1 - \beta_1'}{\sqrt{s^2/[(n-1)\cdot S_x^2]}}$
3. **Hipòtesis**: $H_0: \beta_1 = 0 \quad vs \quad H_1: \beta_1 \neq 0$
4. **Distr. estadístic sota H_0**: $t_{n-2} = t_{14}$
5. **Càlculs**: $\hat{t} = 0.018/0.024 = 7.48$
6. ***P*-valor:** $P(|t_{14}| > 7.48) \approx 3\cdot10^{-6}$ (punt crític = $t_{14,0.975} = 2.145$)
7. **Conclusió**: rebutgem H_0 (*P*-valor < 0.05 o que 7.48 > 2.145)

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

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

[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]

Prova d'hipòtesi sobre el terme independent (β_0). Ex.

1. **Variables**: Cervesa i contingut d'alcohol a la sang

R: `lm`

```
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
```

2. **Estadístic**: $\hat{t} = \frac{b_0 - \beta_0'}{s_{b_0}} = \frac{b_0 - \beta_0'}{\sqrt{s^2\left(\frac{1}{n} + \frac{\bar{X}^2}{(n-1)S_x^2}\right)}}$

3. **Hipòtesis**: $H_0: \beta_0 = 0 \quad vs \quad H_1: \beta_0 \neq 0$

4. **Distr. estadístic sota H_0**: $t_{n-2} = t_{14}$

5. **Càlculs**: $\hat{t} = -0.0127/0.0126 = -1.00$

6. ***P*-valor:** $P(|t_{14}| > 1.00) \approx 0.33$ (punt crític = $t_{14,0.975} = 2.145$)

7. **Conclusió**: NO rebutgem H_0 (*P*-valor > 0.05 o que -1.00 > -2.145)

   Conclusió pràctica: 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%

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

Prova d'hipòtesi sobre el pendent (β_1). Exercici

1. **Variables**:

2. **Estadístic**:

3. **Hipòtesis**:

4. **Distr. estadístic sota H_0**:

5. **Càlculs**:

6. ***P*-valor:**

7. **Conclusió**:

   Conclusió pràctica:

8. **$IC_{95\%}$**:

R: `lm`

```
Call: lm(formula = Durada ~ Brillantor)
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
```

Validació del model lineal

- L'anàlisi de les premisses en la variable de resposta en regressió pot fer referència a la part determinista (recta) o a la part aleatòria (residual).
- En la part determinista (1 premissa):
  - **Linealitat** entre X i Y en el rang considerat

- En la part aleatòria (3 premisses). Com que X_i no és v.a., és constant, no està mesurada amb error, llavors $V(y_i) = V(\beta_0 + \beta_1 X_i + e_i) = V(e_i) = \sigma^2$. Així les premisses sobre la part aleatòria de y_i les analitzem sobre els residus e_i. Els e_i són v.a. i.i.d. amb una distribució Normal $N(0,\sigma^2)$ [es diu que e_i és soroll blanc]:
  - **Homoscedasticitat**: mateixa σ^2 per qualsevol *i*
  - **Independència**: un error no aporta informació sobre el valor de l'altre
  - **Normalitat:** resultat de molts fenòmens aleatoris amb pesos petits

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \rightarrow \epsilon_i \sim N(0,\sigma) \qquad \epsilon_i, \epsilon_j \ \ ind. \ \forall i,j$$

Linealitat | Normalitat | Homoscedasticitat | Independència

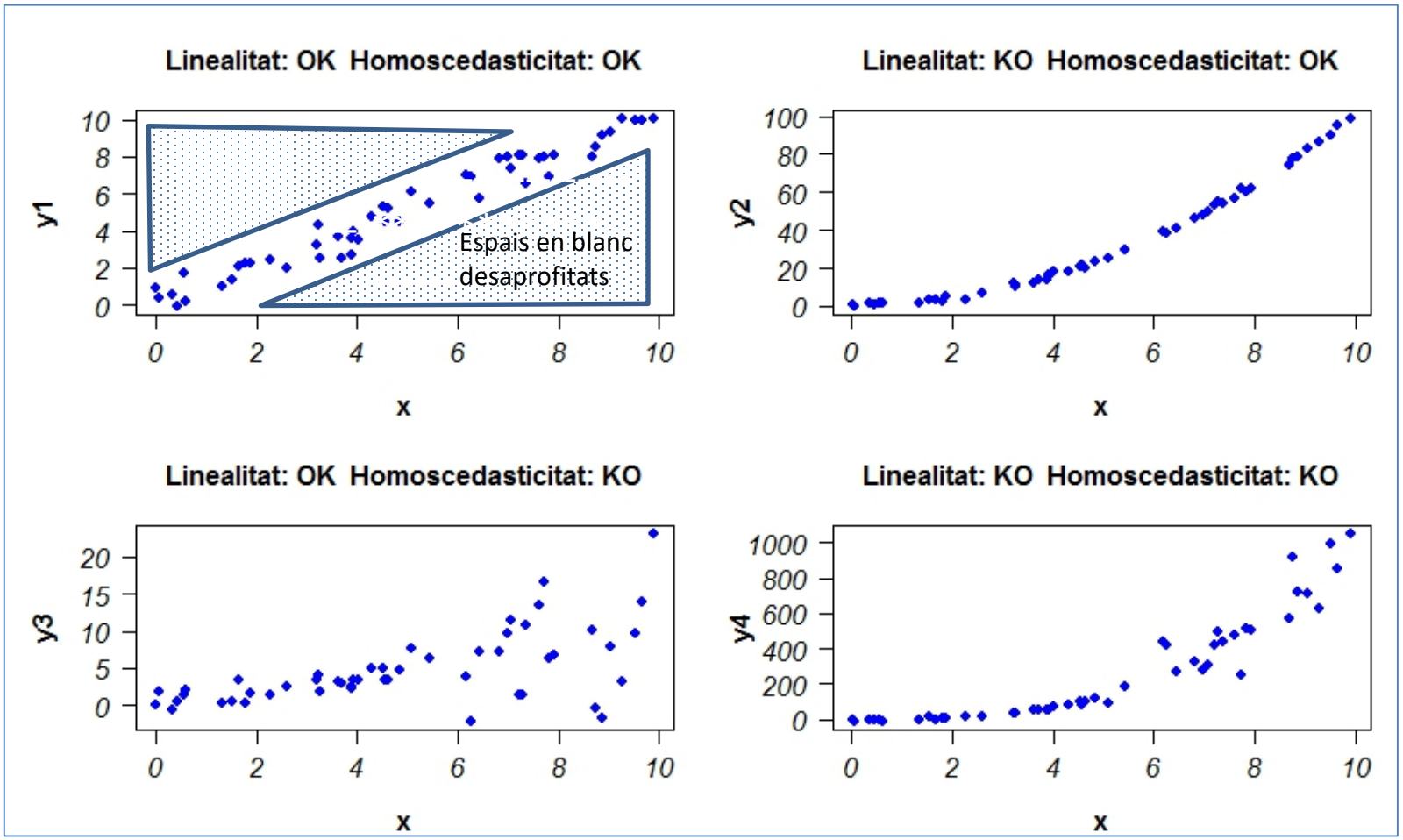
Validació model lineal. Anàlisi dels residus

- El compliment de les premisses anteriors permet:
  - poder recórrer a les distribucions de referència (per fer IC, PH)
  - garantir que el model és el millor possible
- L'anàlisi de les premisses:
  - Estudia si són raonables
  - O, en cas contrari, com trobar un model alternatiu per a que es compleixin
- La validació es realitza mitjançant gràfics dels residus. Usarem els següents gràfics per validar (o no) les premisses:

- Y_i versus X_i → Linealitat i homoscedasticitat
- **e_i versus "Fitted Values"** → Linealitat i homoscedasticitat
- **e_i versus ordre observacions** → Independència
- **Qqnorm dels residus (e_i)** → Premissa de Normalitat
- **Histograma dels residus (e_i)** → Premissa de Normalitat

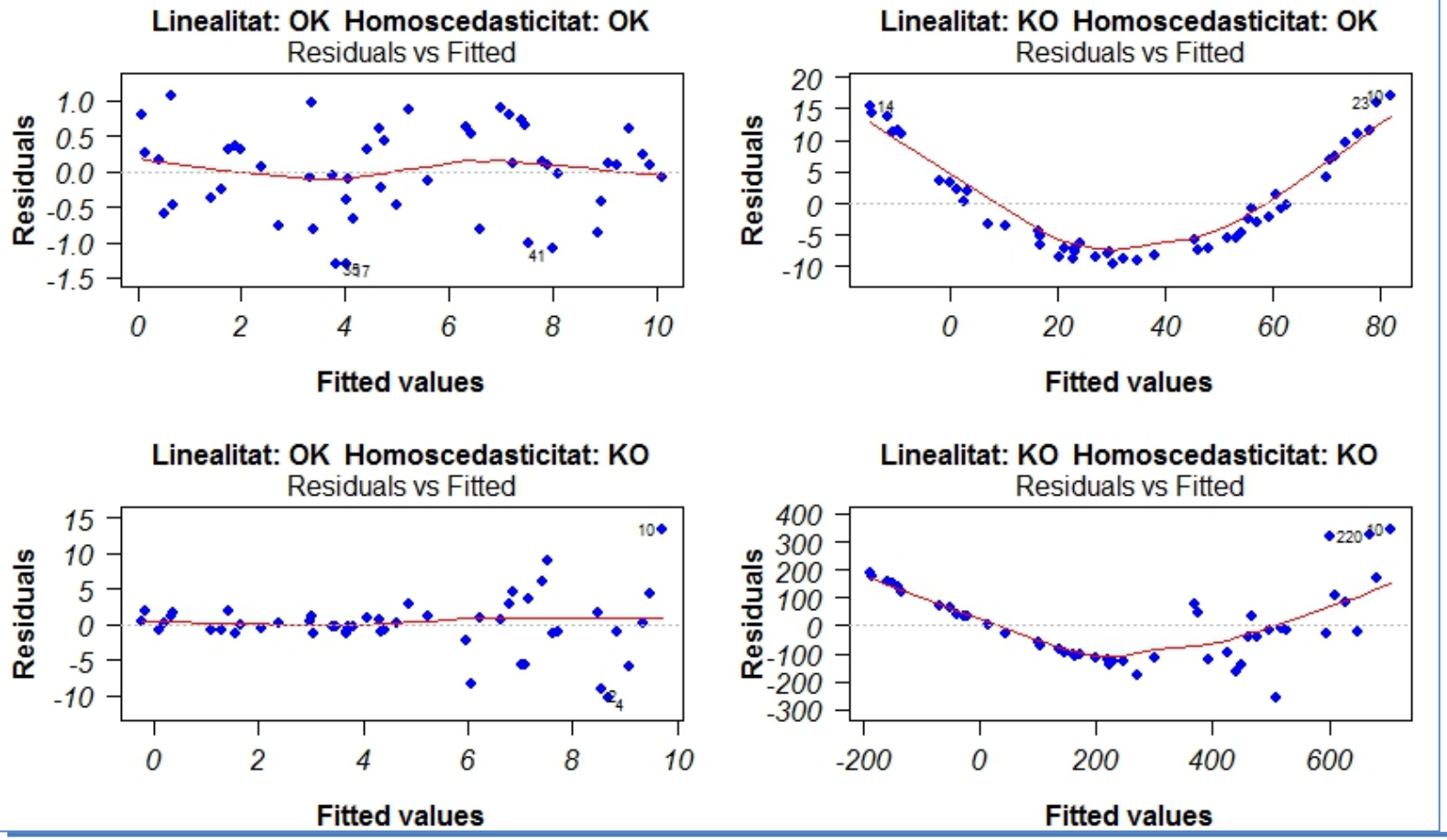
Validació model lineal. Gràfic Y_i versus X_i

- Permet estudiar la linealitat i la homoscedasticitat. És molt fàcil i intuïtiu, però ineficient: molts espais en blanc. Es pot millorar, substituint Y pels residus (e_i).



Validació model lineal. e_i versus fitted values.

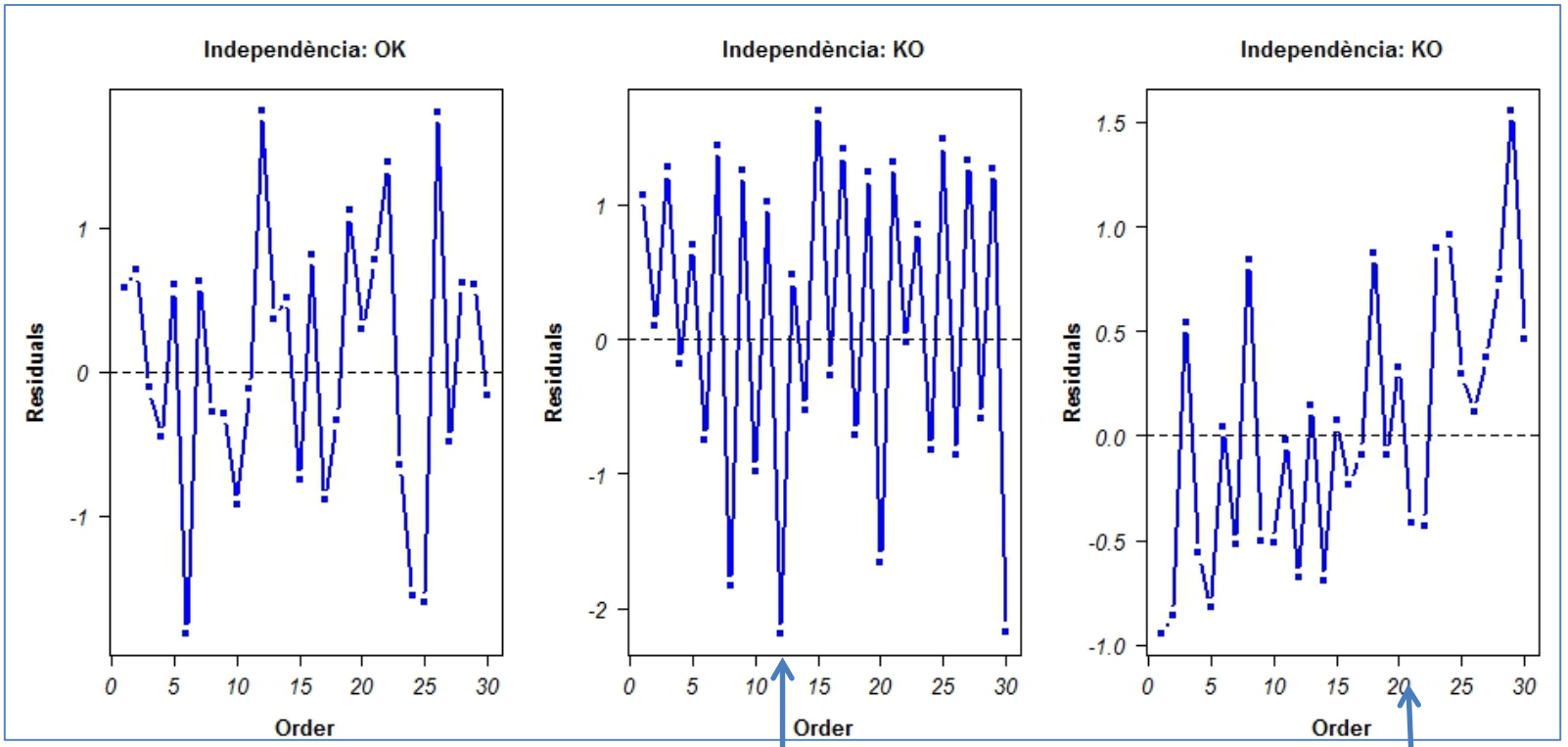
- **Linealitat**: El núvol de punts ha de mantenir sempre la mateixa alçada (aprox.).
- **Homoscedasticitat**: La variabilitat dels residus ha de mantenir-se constant independentment dels valors predits (*fitted values*).



Suggeriment: proveu amb les dades del consum de benzina i velocitat

Validació model lineal. e_i versus ordre de les observac.

- **Independència**: Els residus no han de mostrar cap patró enfront l'ordre.

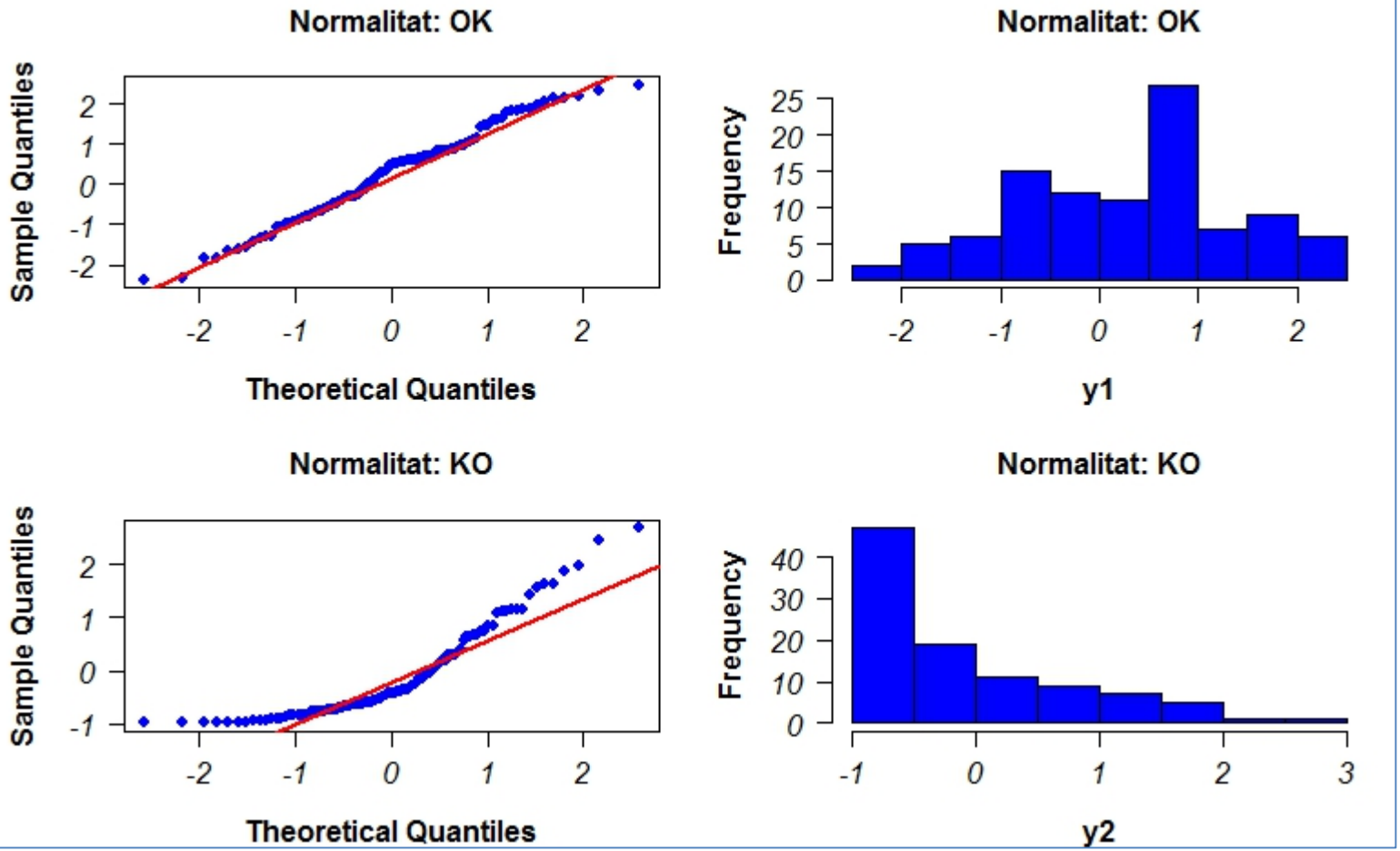


Hi ha patró: s'alternen pujades i baixades sistemàticament: no hi ha independència entre observacions consecutives. És típica de variables recollides al llarg del temps. [Ex: hores dormides en dies consecutius]

Hi ha un patró: tendència creixent dels residus. Segurament, s'ha anat canviant el criteri de recollida amb el temps.

Validació model lineal. qqnorm i histograma de residus

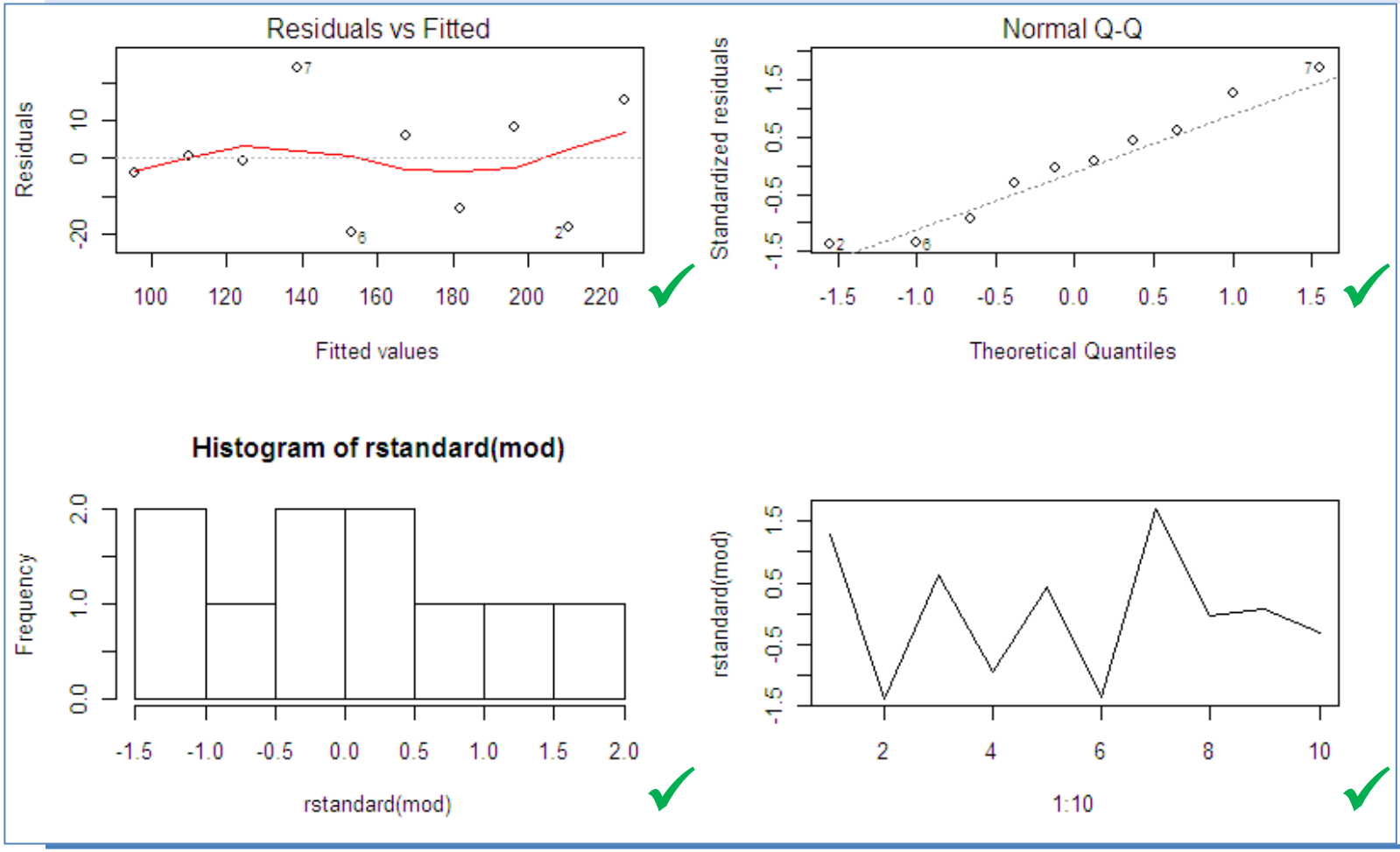
- **Normalitat**: Els residus han de ser normals: situar-se sobre la recta en el qqnorm i forma de campana a l'histograma.



Nota: El qqnorm és molt més fiable a l'hora d'avaluar la Normalitat

Validació model lineal. Codi R

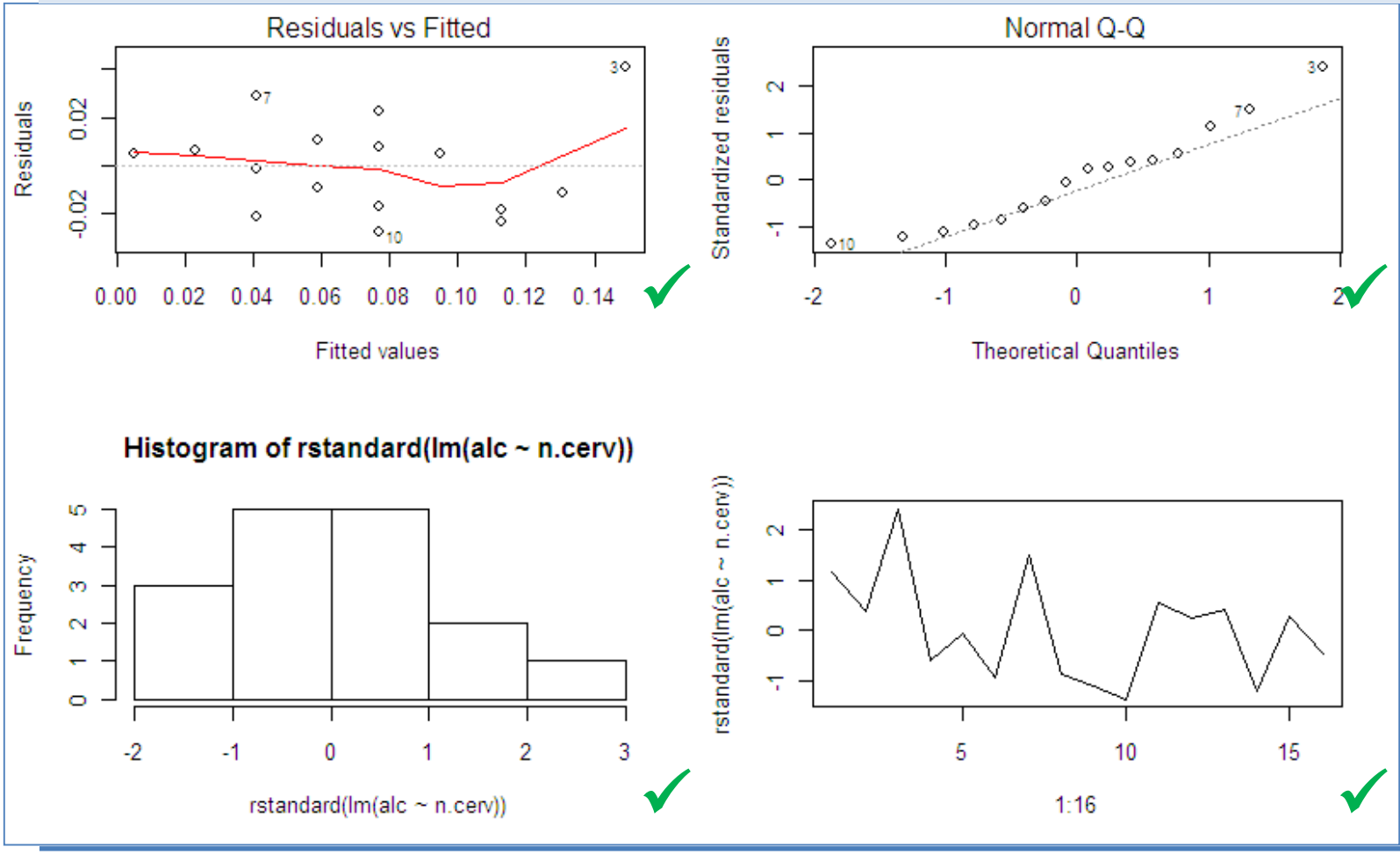
```
##-- Exemple de la pantalla d'ordinador
par(mfrow=c(2,2))
plot(lm(Durada ~ Brill),c(2,1))                    # QQ-Norm i Standard Residuals vs. Fitted
hist(rstandard(lm(Durada ~ Brill)))                # Histograma dels residus estandaritzats
plot (1:10,rstandard(lm(Durada ~ Brill)),type="l") # Ordre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



Encara que són poques dades, res s'oposa a validar cap de les 4 premisses

Validació model lineal. Codi R

```
##-- 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

Validació model lineal. Consideracions generals

- **Totes les premisses**: Generalment, amb poques dades, és difícil avaluar les premisses i l'opció més prudent és acceptar-les a no ser que ho veiem molt clar que alguna d'elles s'infringeix.
- **Totes les premisses**: No s'ha de ser categòric i s'han d'interpretar els resultats amb cautela. S'ha d'evitar "Aquesta variable és homoscedàstica" o "Hem demostrat que la variable és Normal". Millor dir "Aquesta variable pot ser modelada assumint homoscedasticitat" o "Aquesta variable pot ser modelada amb la distribució Normal"
- **Linealitat i Homoscedasticitat**: Encara que es pot fer servir directament el gràfic de les Y_i vs X_i, es recomana emprar els residus vs "valors predits" per estudiar-les.
- **Homoscedasticitat**: cal recordar que S té molta oscil·lació mostral i, a vegades és poden observar fluctuacions al llarg dels valors predits que poden ser degudes al atzar.
- **Normalitat**: És més fiable el qqnorm que no l'histograma per avaluar-la
- **Independència**: El fet de ser independents respecte a l'ordre no garanteix del tot la indepèndencia. Aquesta ha de ser garantida amb un bon disseny de recollida.
- **Totes les premisses**: En motes ocasions, traient logaritmes (o arrel o fent la inversa) de alguna/es de les variable/s, podem solucionar l'incompliment de les premisses

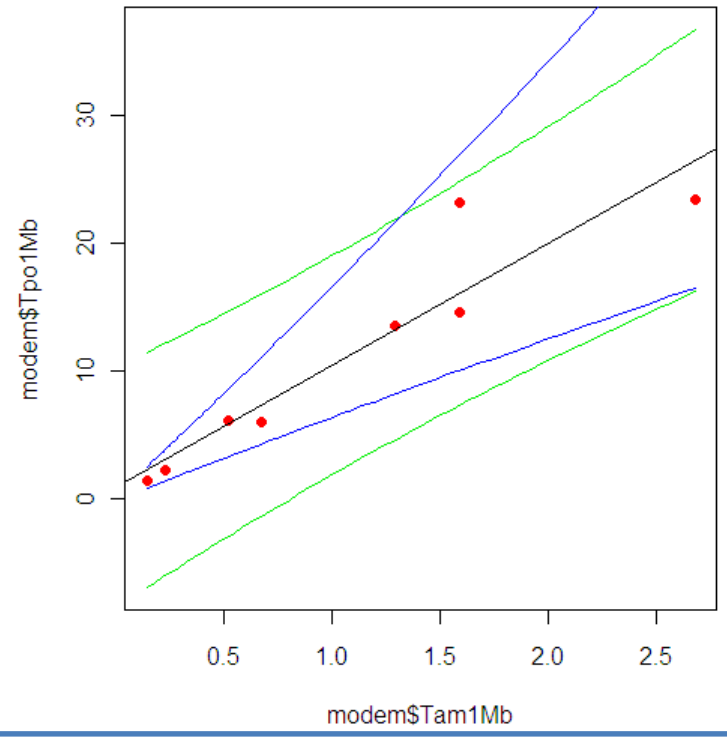
Validació. Transformació logarítmica

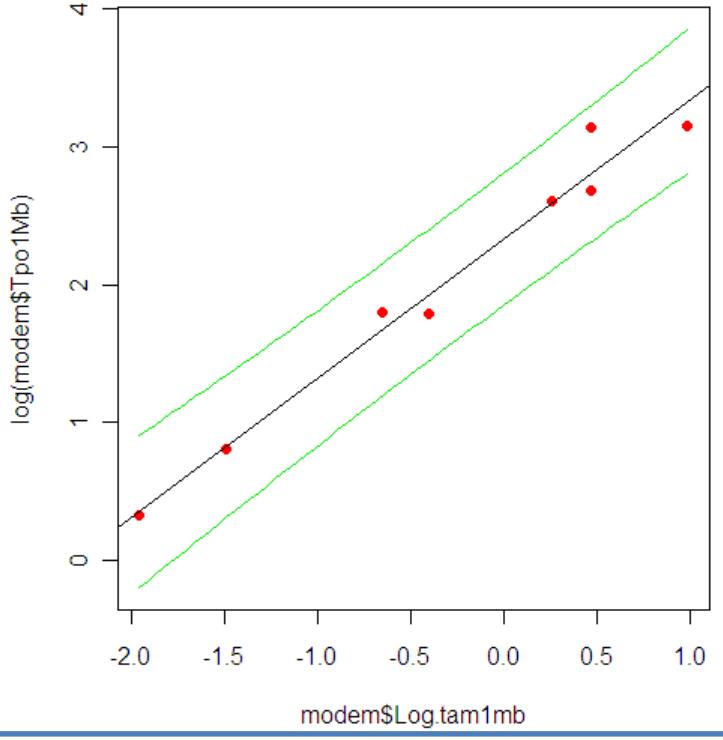
En ocasions, fer la transformació logarítmica pot solucionar el NO compliment de les premisses. Ex: Velocitat de descarrega de fitxers amb un mòdem de 1Mbps (**Resposta**: temps [s] ; **Var. explicativa**: mida fitxer [MB])

Model #1: temps vs mida. Problema: heteroscedasticitat. Tenim prediccions negatives

Model #2: log(temps) vs log(mida). Desfem canvi amb **exp(predicció)**; ara són satisfactòries i tenen en compte que fitxers petits tenen fluctuacions petites en temps

Sense logaritmes





Amb logaritmes

Predicció

- En primer lloc la predicció puntual de Y per a valors concrets de X (X_h) usa la part determinista: $\hat{y}_h = b_0 + b_1 \cdot X_h$
- Però, com tenir en compte la part aleatòria? Dues situacions ben diferenciades:
  1. Estimar un interval de confiança pel **valor esperat** de les observacions X = X_h
  2. Estimar un interval de confiança pel **valor individual** corresponent a X = X_h

1. La estimació puntual per *y* donat un valor X_h és:

$$\hat{y}_h = b_0 + b_1 \cdot X_h = \bar{Y} + b_1 \cdot (X_h - \bar{X})$$

Podem estimar l'esperança i la variància d'aquest estimador:

$$E(\hat{y}_h) = E(b_0 + b_1 \cdot X_h) = \beta_0 + \beta_1 \cdot X_h = \mu_h$$ [És no esbiaxiat!!!]

$$V(\hat{y}_h) = V\left(\bar{Y} + b_1 \cdot (X_h - \bar{X})\right) = V(\bar{Y}) + (X_h - \bar{X})^2 \cdot V(b_1) = \frac{\sigma^2}{n} + \frac{(X_h - \bar{X})^2 \sigma^2}{(n-1) \cdot S_x^2} = \sigma^2 \left(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{(n-1) \cdot S_x^2}\right)$$

Nota 1: Noteu que major variància a major distància entre X_h i $\bar{X}$ (més precissió al mesurar punts propers a la mitjana)

Nota 2: Substituint s per S podem fer regions de confiança per μ_h amb una t_{n-2}

Predicció

2. Per **predir** l'interval dels valors individuals y_h de Y per X=X_h utilitzarem també:

$$\hat{y}_h = b_0 + b_1 \cdot X_h$$

Calcularem esperança i variança:

$$E(y_h) = E(\hat{y}_h) = \mu_h$$

Té Error Quadràtic Mitjà de Predicció (EQMP) que es pot descomposar de forma semblant a la descomposició de sumes de quadrats:

$$EQMP = E[(\hat{y}_h - y_h)^2] = E(\hat{y}_h - \mu_h)^2 + E(y_h - \mu_h)^2$$

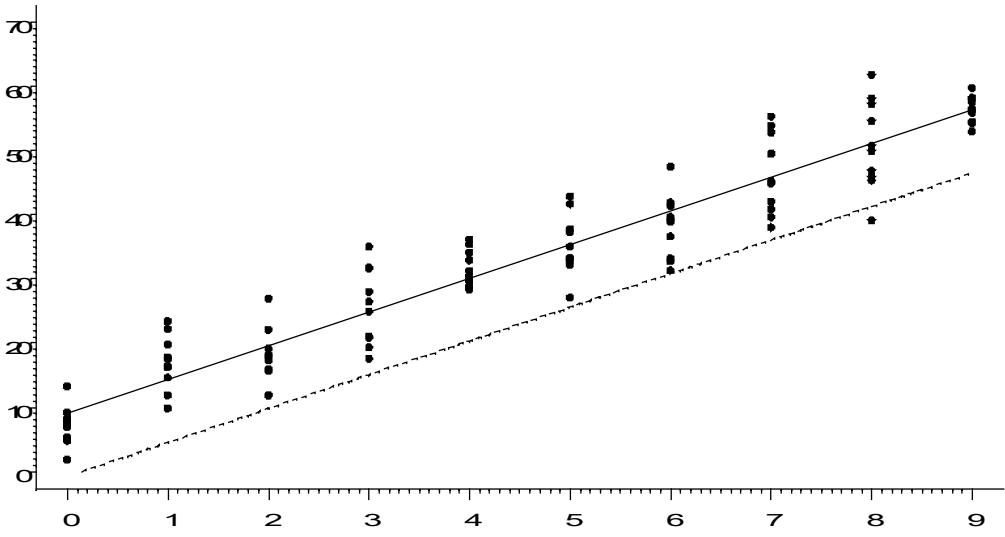
És a dir:

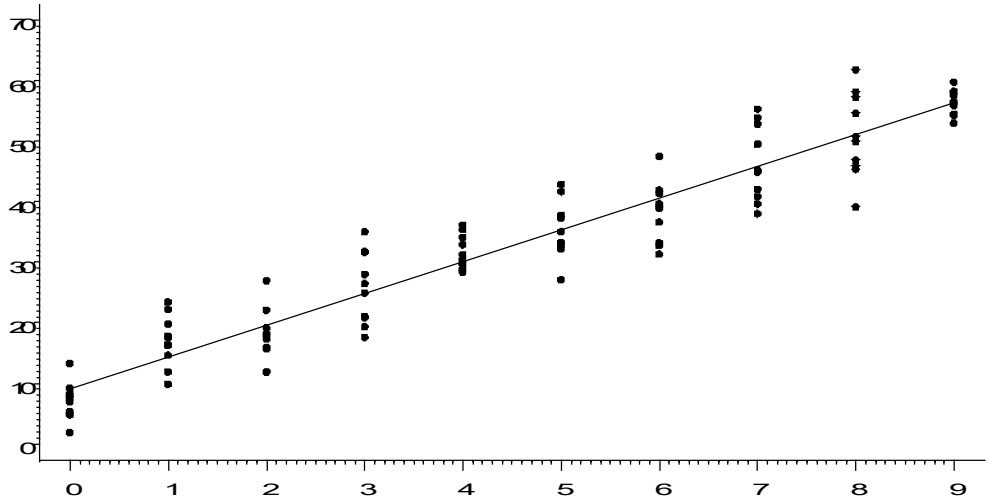
$$V(y_h) = \sigma^2 \left(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{(n-1)S_x^2} \right) + \sigma^2 = \sigma^2 \left(1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{(n-1)S_x^2} \right)$$

Permet identificar 3 fonts de variabilitat en la predicció dels valors individuals: **Natural** (σ^2) **+ Per estimació mitjana** (σ^2/n) **+ Per estimació pendent**

Predicció. Gràfics

Predicció sobre valors individuals





Predicció sobre valors esperats

Formulari. Resum de previsions de la resposta

	Valor esperat	Valors individuals
Estimació puntual	$\hat{y}_h = b_0 + b_1 X_h$	$\hat{y}_h = b_0 + b_1 X_h$
Estimació per interval	$\hat{y}_h \mp \mathrm{t}_{\mathrm{n}-2,0.975} \cdot \mathrm{S} \sqrt{\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum (X_i - \bar{X})^2}}$	$\hat{y}_h \mp \mathrm{t}_{\mathrm{n}-2,0.975} \cdot \mathrm{S} \sqrt{1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum (X_i - \bar{X})^2}}$
R	`predict(…,interval='confidence')`	`predict(…,interval='prediction')`

Predicció. Exemple

Recuperem 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:

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

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

Quina durada podem esperar per a pantalles de brillantor 7.5?

	Valor esperat	Valors individuals
Estimació puntual	$\hat{y}_h = 239.9 - 14.41 \cdot 7.5 = \mathbf{131.83}$	$\hat{y}_h = 239.9 - 14.41 \cdot 7.5 = \mathbf{131.83}$
Estimació per interval	$131.83 \mp 2.31 \cdot \sqrt{227.3} \cdot \sqrt{\frac{1}{10} + \frac{(7.5-5.5)^2}{9 \cdot 9.167}}$ $= [\mathbf{118.41, 145.25}]$	$131.83 \mp 2.31 \cdot \sqrt{227.3} \cdot \sqrt{1 + \frac{1}{10} + \frac{(7.5-5.5)^2}{9 \cdot 9.167}}$ $= [\mathbf{94.50, 169.16}]$
Conclusió	**Per a les pantalles** de brillantor de 7.5 podem esperar una **durada mitjana** entre 118.41 i 145.25 min. amb una confiança del 95%	**Per a una pantalla** de brillantor 7.5 podem esperar una **durada** entre 94.50 i 169.16 min. amb una confiança del 95%

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

Predicció. Exemple (Amb R)

```
> 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
...
```

Trobeu aquests valors a mà

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

Capacitat predictiva. Coeficient R^2

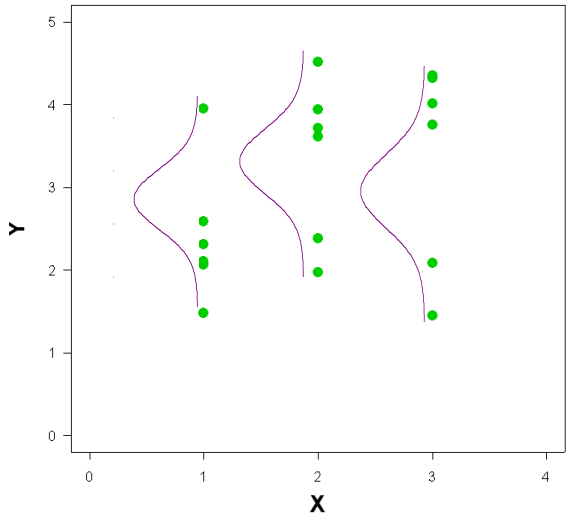
- Es va veure que la correlació entre X i Y (r_{XY}) estudia la relació (lineal) entre dues variables X i Y amb un rol simètric:

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

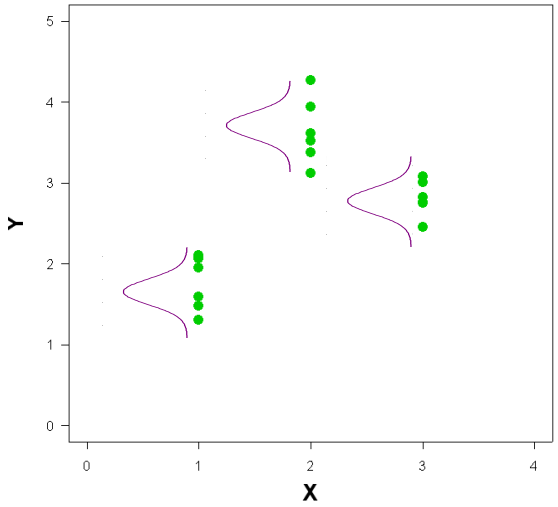
- Definim el coeficient ***R^2*** (***Coeficient de determinació, o R-squared***), com el quadrat de la correlació lineal *r*. Noteu que:
  - $R^2 = r_{XY}^2$ → $0 \leq R^2 \leq 1$
  - Ve a significar quina fracció de la variabilitat de Y s'explica per el factor X (la interpretació és asimètrica).
  - Un R^2 alt ens diu que el model lineal fa un bon ajustament de les dades :: els punts s'allunyen poc de la recta :: poca variabilitat d'origen aleatori
  - Recíprocament, amb R^2 baix, les dades no s'ajusten be :: els punts es poden allunyar molt :: gran variabilitat d'origen aleatori (no explicada per X, volem dir).
  - R^2 és un indicador de qualitat de l'ajustament, partint de que tenim un model lineal mentre que *r* és un indicador d'associació entre dues variables relacionades linealment, però no suposa cap model al darrera (caràcter descriptiu)

Annexe: quantitativa vs. categòrica.

- Quin model usar quan la intervenció **X** (o condició Z) és categòrica?



Petita variabilitat entre grups
Gran variabilitat intra grups



Gran variabilidad entre grups
Petita variabilitat intra grups

- Com a la regressió, podrem descomposar la variabilitat total en dues fonts de variació: entre-grupos (between) i intra-grupos (intra)
- Com a la regressió, podem tenir 2 objectius ben difereciats:
  - predir la resposta Y a partir (observació) dels valors Z (interesarà conneixer R^2)
  - canviar la resposta Y escollint (disseny d'experiments) els valor de X (interesaran les μ_i)

Model quantitativa vs. categòrica. Exemple

Exemple: Temps i nombre de nodes de graf en Dijkstra

Volem estudiar: - el temps de CPU empleat per l'algorisme Dijkstra

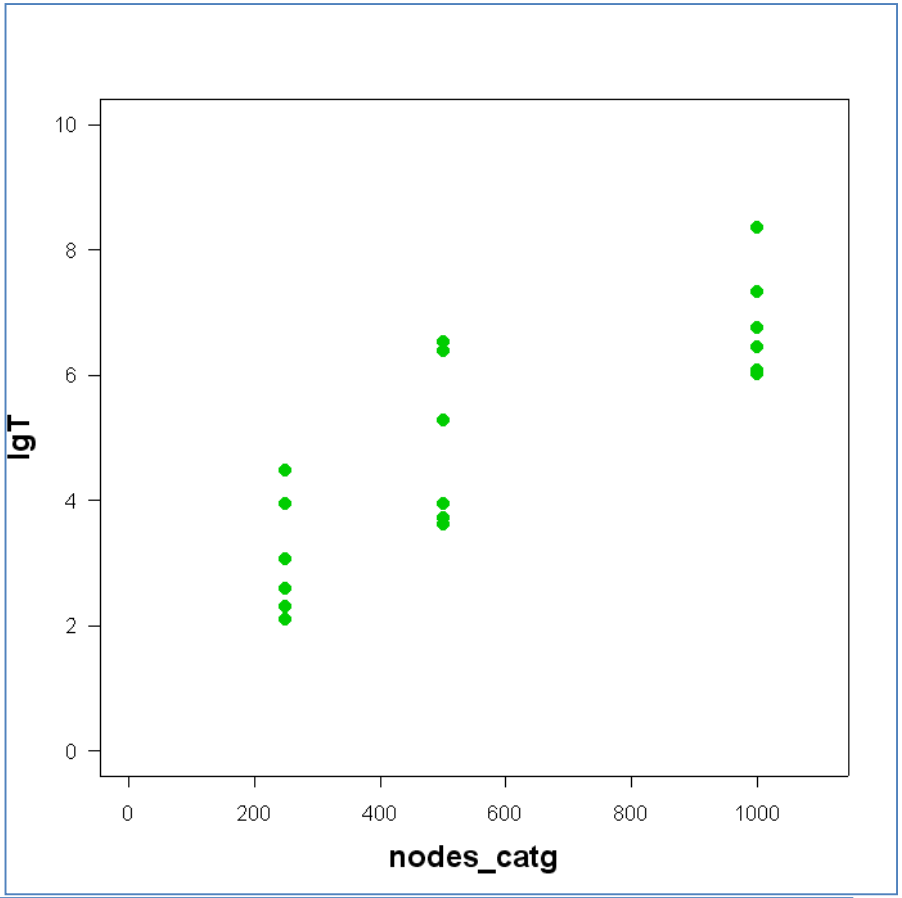
- segons el número de nodes del graf

Ho fem estudiant les característiques en grafs de 250, 500 i de 1000 nodes.

Si dubtem de la linealitat de les 3 mitjanes podem optar per un model que oblidi el número de nodes i els tracti com 3 categories, mirant únicament quin és el valor mitjà de Y per cada categoria de Z.

En aquest cas, té sentit estudiar la descomposició de la variabilitat en:

- Deguda al factor X
- Aleatòria o residual



Model quantitativa vs. categòrica. Paràmetres

Sigui el model: $\mathbf{Y_{ij} = \mu_j + e_{ij}}$

Y_{ij} valor de la Y en el cas i del grup j

μ_j esperança del grup j (de $\mathbf{n_j}$ observacions de les **N** totals)
(el paràmetre μ_j s'estima per la mitjana $\bar{y}_j$) (desviació $\mathbf{S_j}$)

e_{ij} error aleatori o diferència del cas i a la mitjana del seu grup j
(el paràmetre σ^2 és la variància de e_i o variància residual)

Aquest model es pot veure com l'extensió de la comparació de 2 μ al cas de k μ. Però també, com l'estudi de la descomposició de la variabilitat:

- **Entre els grups**: quant expliquen de la variabilitat global
- **Dins els grups**: variabilitat 'dins' que no es pot relacionar amb el grup

Model quantitativa vs. categòrica. Paràmetres

EXEMPLE: (*Estadística per a enginyers informàtics. Ed UPC pg 154 Ref: Eei.Ed.UPC pg154*). Notes en 3 grups d'una assignatura. Els paràmetres μ_1, μ_2 i μ_3 són les esperances de la nota en cadascun dels grups de grandàries n_1=32, n_2=28 i n_3=25 casos respectivament (en total N=85)

Les mitjanes i desviacions mostrals en cada grup són:

$\bar{y}_1 = 6.15; \bar{y}_2 = 5.73; \bar{y}_3 = 5.48; s_1 = 1.8; s_2 = 1.5; s_3 = 2.0$

Quant val la mitjana global? $\bar{Y} = 5.81$

Quant val la variabilitat combinada dins els grups? $S^2 = 3.136$

(com en la regressió, S^2 és l'estimació de σ^2 i es comprovarà amb resultat a taula Anova)

Model quantitativa vs. categòrica. Anàlisi: ANOVA

TAULA D'ANÀLISIS DE LA VARIANÇA (ANOVA, ANalysis Of VAriance)

R: `aov`

Posarem els termes de SQ en forma de taula (Ref: *Eei.Ed.UPC* pg 154):

Font Var.	SQ (Sum Squares)	Df*	QM$	Ratio	P-value
Between (Explicada pel model)	$SQ_E = \sum_{j=1}^{k} n_j \cdot (\bar{y}_j - \bar{Y})^2$	k-1	$QM_E = \frac{SQ_E}{k-1}$	$\hat{F} = \frac{QM_E}{QM_R}$	$P(F_{k-1,N-k} > \hat{F})$
Intra (Residual)	$SQ_R = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2$	N-k	$QM_R = \frac{SQ_R}{N-k}$		
Total	$SQ_T = SQ_E + SQ_R = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{ij} - \bar{Y})^2$	N-1			

Between, Intra: Variabilitat explicada pel model (between) i aleatòria (intra)

***Df**: Degree of freedom (graus de llibertat)

$QM: Quadratic Mean

k: número de grups a comparar

N: Nombre d'observacions totals

$\hat{F}$: Aquest estadístic sota la hiòtesi (H_0) de que les mitjanes no són diferents segueix una F amb k-1 i N-k graus de llibertat

Model quantitativa vs. categòrica. Anàlisi: ANOVA

PH GLOBAL: [Ref: *Eei.Ed.UPC* pg 155]

- La hipòtesi de que X no aporta informació sobre Y (igualtat de totes les m_j) es tradueix en que tota la variabilitat entre les m_j és deguda a la fluctuació del mostreig
- PH: H_0 : Variabilitat(μ_j) = 0

  H_1: Variabilitat(μ_j) > 0 **unilateral!**
- Es resol amb la ràtio F dels quadrats mitjos de la taula de descomposició de la variabilitat:

$$\hat{F} = \frac{QM_E}{QM_R}$$

COEFICIENT DE DETERMINACIÓ:

- Com en el cas anterior, és un rati que ens permet identificar de tota la variabilitat de les Y quina part ve associada a (explicada per) X

$$R^2 = \frac{SQ_E}{SQ_T}$$

ANOVA. Exemple

Notes en 3 grups (Ref: Eei.Ed.UPC pàg. 154)

Font Variabilitat	SQ	GdL DF	QM	Raó
Explicada (entre)	6.60	2	3.3	1.052
Residual (intra)	257.19	82	3.136	
Total	263.79	84		

$$SQ_E = \sum_{j=1}^{k} n_j(\bar{y}_j - \bar{y}) = 6.60 \qquad SQ_R = \sum_{j=1}^{k} (n_j - 1) \cdot s_j^2 = 257.19 \qquad SQ_T = \mathrm{SQ_E} + \mathrm{SQ_R} = 263.79$$

1. **Objectiu:** saber si el grup afecta a l'esperança de la nota
2. **Hipòtesi:** H_0: Variabilitat(μ_j) = 0 (unilateral)
3. **Estadístic**: $\hat{F} = QM_E / QM_R$
4. **Distribució sota H_0**: $\hat{F} \rightarrow F_{2,82}$ (les premisses caldrà indicar-les i analitzar-les)
5. **Càlculs**: $\hat{F} = 1.052$
6. **P-valor:** 0.354 `[(1-pf(1.052,2,82))]`
7. **Conclusió**: SÍ és versemblant que l'esperança de la nota és igual en tots els grups

   Pràctica: El rendiment mitjà no és diferent en els tres grups estudiats

ANOVA. Exemple (cont)

8. Intervals de confiança

Font Variabilitat	SQ	GdL DF	QM	Raó
Explicada (entre)	6.60	2	3.3	1.052
Residual (intra)	257.19	82	3.136	
Total	263.79	84		

Sabent que:

$\bar{y}_1 = 6.15; \bar{y}_2 = 5.73; \bar{y}_3 = 5.48; s_1 = 1.8; s_2 = 1.5; s_3 = 2.0$

$\bar{Y} = 5.81\ ; S^2 = 3.136$

Es pot calcular un IC per a la μ global: $IC(\mu, 1-\alpha) = \bar{Y} \mp t_{N-k,1-\frac{\alpha}{2}} \cdot \sqrt{\frac{QM_R}{N}}$

$IC(\mu, 0.95) = [5.432, 6.197]$

També es pot calcular un IC per a cada μ_j amb la desviació pooled (més robust que calculat amb les dades de cada grup): $IC(\mu_j, 1-\alpha) = \bar{y}_j \mp t_{N-k,1-\frac{\alpha}{2}} \cdot \sqrt{\frac{QM_R}{n_j}}$

$IC(\mu_1, 0.95) = [5.527, 6.773]$ $IC(\mu_2, 0.95) = [5.064, 6.396]$ $IC(\mu_3, 0.95) = [4.775, 6.185]$

ANOVA. Exercici

- X:nombre de nodes, Y: temps amb transformació logarítmica

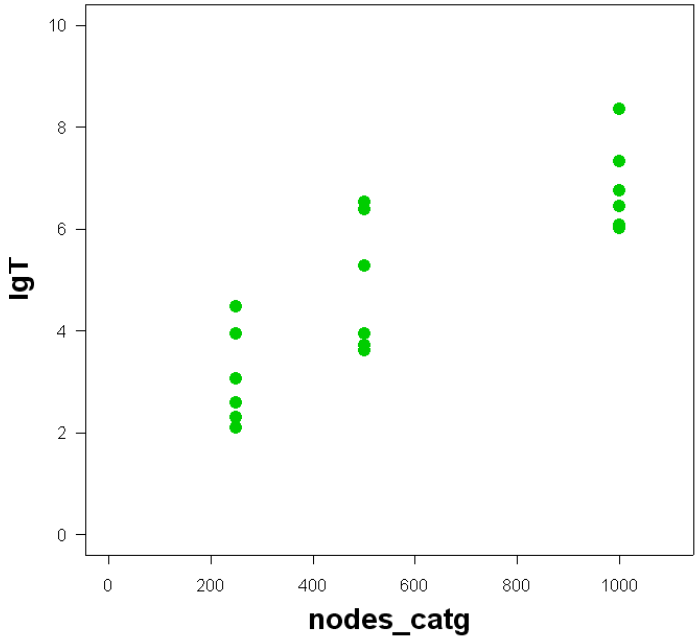
x_i (nodes)	y_i (lgt)
250	2.31
250	4.48
250	2.59
250	3.06
250	2.10
250	3.95
500	3.94
500	6.38
500	6.52
500	5.27
500	3.72
500	3.61
1000	6.45
1000	7.32
1000	6.76
1000	6.08
1000	8.35
1000	6.01

$\bar{y}_1 = 3.082 \quad s_1^2 = 0.90$

$\bar{y}_2 = 4.910 \quad s_2^2 = 1.79$

$\bar{y}_3 = 6.83 \quad s_3^2 = 0.79$

$\bar{y} = 4.94 \;\; s^2 = 1.16 \rightarrow S = 1.08$



Model amb R:

```
> aov(lgT~as.factor(nodes_catg))
Call:aov(formula=lgT~as.factor(
nodes_catg))
Terms:
                  nodes_catg Residuals
Sum of Squares     42.12188  17.37490
Deg. of Freedom           2        15
Residual standard error: 1.076256
Estimated effects may be unbalanced
```

ANOVA. Exercici

Temps i nombre de nodes de graf en Dijkstra

Font Variabilitat	SQ	GdL DF	QM	Raó
Explicada (entre)	42.12	2	21.06	18.15
Residual (intra)	17.38	15	1.16	
Total	59.5	17		

Model amb R:

```
> anova(aov(lgT~nodes_catg))
Call:aov(formula=lgT~nodes_catg)
Analysis of Variance Table
Response:  lgT
            Df Sum Sq Mean Sq F value    Pr(>F)
nodes_catg   2 42.122  21.061  18.182 9.789e-05
Residuals   15 17.375   1.158
```

$$SQ_E = \sum_{j=1}^{k} n_j(\bar{y}_j - \bar{y}) = 42.12 \qquad SQ_R = \sum_{j=1}^{k} (n_j - 1) \cdot s_j^2 = 17.38 \qquad SQ_T = \mathrm{SQ_E} + \mathrm{SQ_R} = 59.50$$

1. **Objectiu:** saber si nombre de nodes afecta al temps
2. **Hipòtesi:** H_0: Variabilitat(μ_j) = 0 (unilateral)
3. **Estadístic**: $\hat{F} = QM_E / QM_R$
4. **Distribució sota H_0**: $\hat{F} \to F_{2,15}$ (les premisses caldrà indicar-les i analitzar-les)
5. **Càlculs**: $\hat{F} = 18.15$
6. **P-valor:** 0.000098 `[(1-pf(18.15,2,15))]`
7. **Conclusió**: No és versemblant que el nombre nodes no aporti informació sobre el temps

Pràctica: El logaritme del temps mitjà és diferent en els tres nivells de nombre de nodes.