PREVISIÓ

Previsió i disseny d'experiments

Model "quantitativa vs quantitativa":

- model, paràmetres i interpretació

Estimadors dels paràmetres: distribució, inferència

Fases del procés de models estadístics

Anàlisi de les premisses. Anàlisi de residus

Predicció

Model "quantitativa vs categòrica":
- model, paràmetres i interpretació
descomposició de la variabilitat

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

wikipedia.org: "causality" "causality contrasted with conditionals" i "Judea Pearl"

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]

Estudis experimentals: *fem* i podem intervenir, canviar el futur

Observem l'efecte en Y havent assignat X a les unitats

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

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...

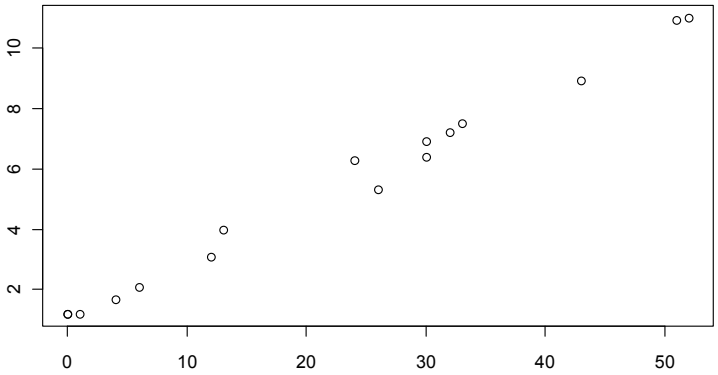
Resum:

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 per predir o per intervenir

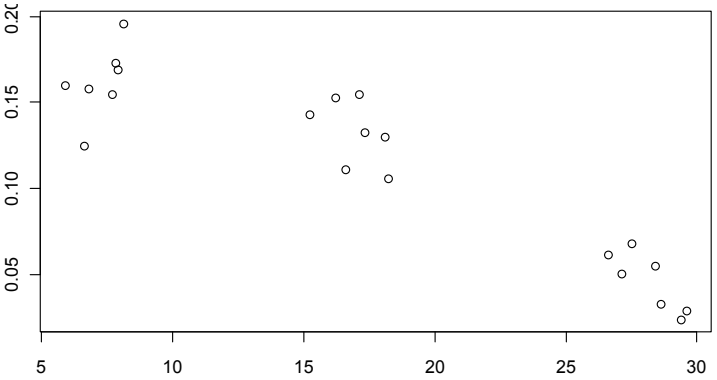
Què diries d'aquestes històries?

Temps dels records mundials de 10000 metres masculí: 1912-2004

Consum de gas per calefacció, segons *una mesura de fred al exterior* (nombre dies temperatura inferior a 65ºF ~ 18.3ºC)



Canvi climàtic: stress del monsó al Mar Aràbic, segons superfície nevada continental.



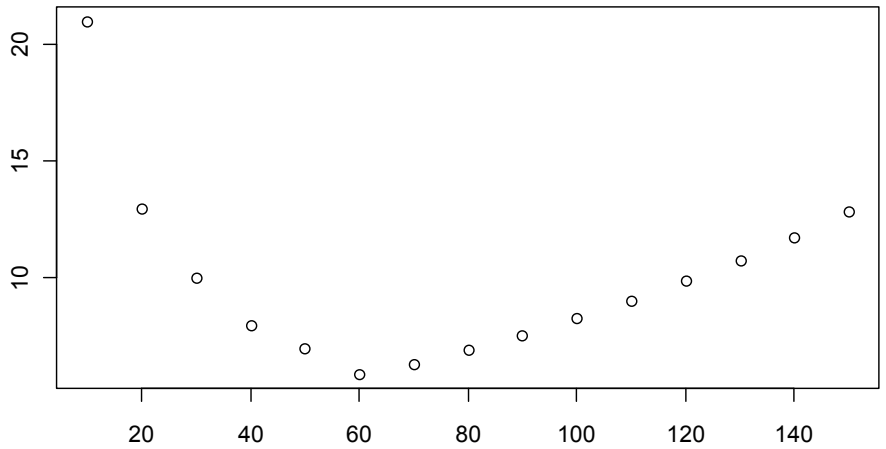
Model "quantitativa vs quantitativa"
REGRESSIÓ (lineal simple)

Una equació com $Y=b_0+b_1X$ pot relacionar-nos dues variables com aquestes:

Així, tenim un model per previsions del **consum** (Y) segons la **velocitat** (X). En aquest cas, Y = 11.058 – 0.01466X

- *Què vol dir el coeficient –0.01466?*
- *Realment podem esperar menys consum amb més velocitat?*

No oblidem que el consum de benzina no depèn només de la velocitat.



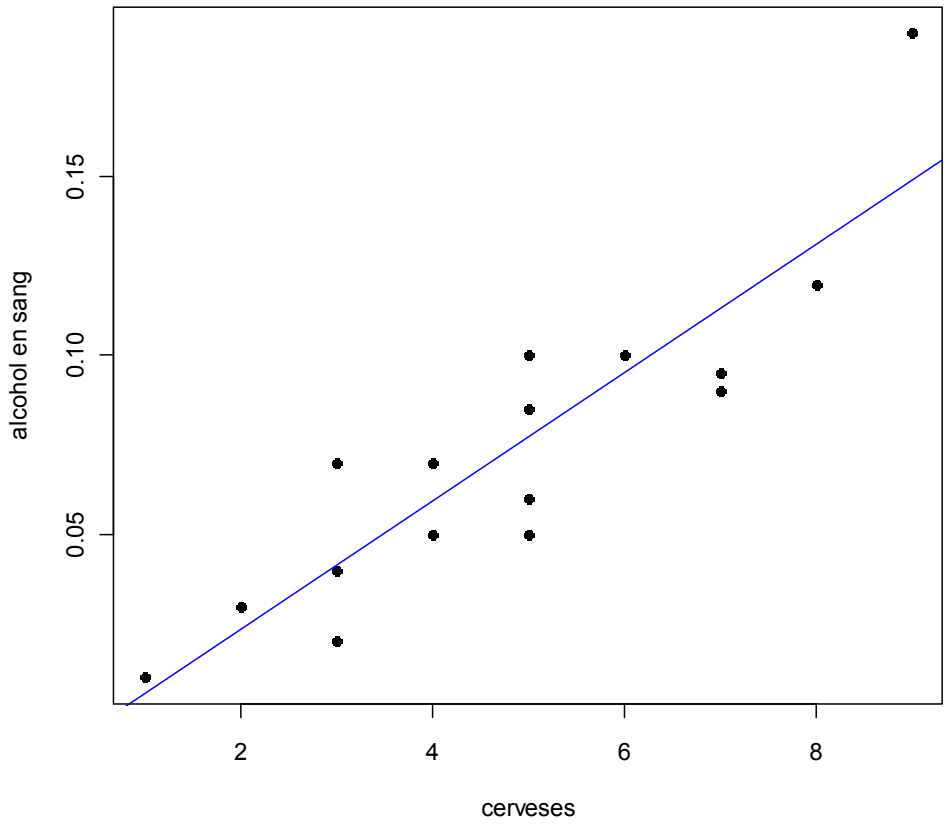
speed	fuel
10	21
20	13
30	10
40	8
50	7
60	5.9
70	6.3
80	6.95
90	7.57
100	8.27
110	9.03
120	9.87
130	10.79
140	11.77
150	12.83
(km/h)	(l/100 km)

EXEMPLE: Cervesa i contingut d'alcohol a la sang

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* i *constant* a l'origen

Al voltant tenim una certa dispersió que requereix un tercer paràmetre: la *variància* σ^2



(The Basic Practice of Statistics. 4th ed. David S. Moore. Example 24.7)

REGRESSIÓ: model i paràmetres

Sigui el model: $\mathbf{Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i}$

Y_i valor de la variable resposta Y en el cas i

X_i valor que pren la condició X en el cas i

ε_i error aleatori o distància a la recta del cas i (centrat al 0)

Els paràmetres seran: β_0 com a constant a l'origen, β_1 com a pendent de la recta,
i σ^2 com la variància de ε_i o variància residual

(distingim: $\beta_0 + \beta_1 X_i$ com a part determinista (lineal) de Y, ε_i com a part aleatòria de Y)

EXEMPLE: (*Estadística per a enginyers informàtics*. Ed UPC pg 141 Ref: *Eei.Ed.UPC* pg141)

Homes adults i sans de Barcelona: Y és Pes en Kg, X és Alçada en cm. Suposem model recta amb paràmetres: β_0 = -100 Kg β_1 = +1 Kg/cm σ = 6Kg

Quin pes correspon a un senyor de 160 cm? *Solució: ...* I a un de 180 cm? *Solució: ...*

Què significa "correspon"? *Solució: ...*

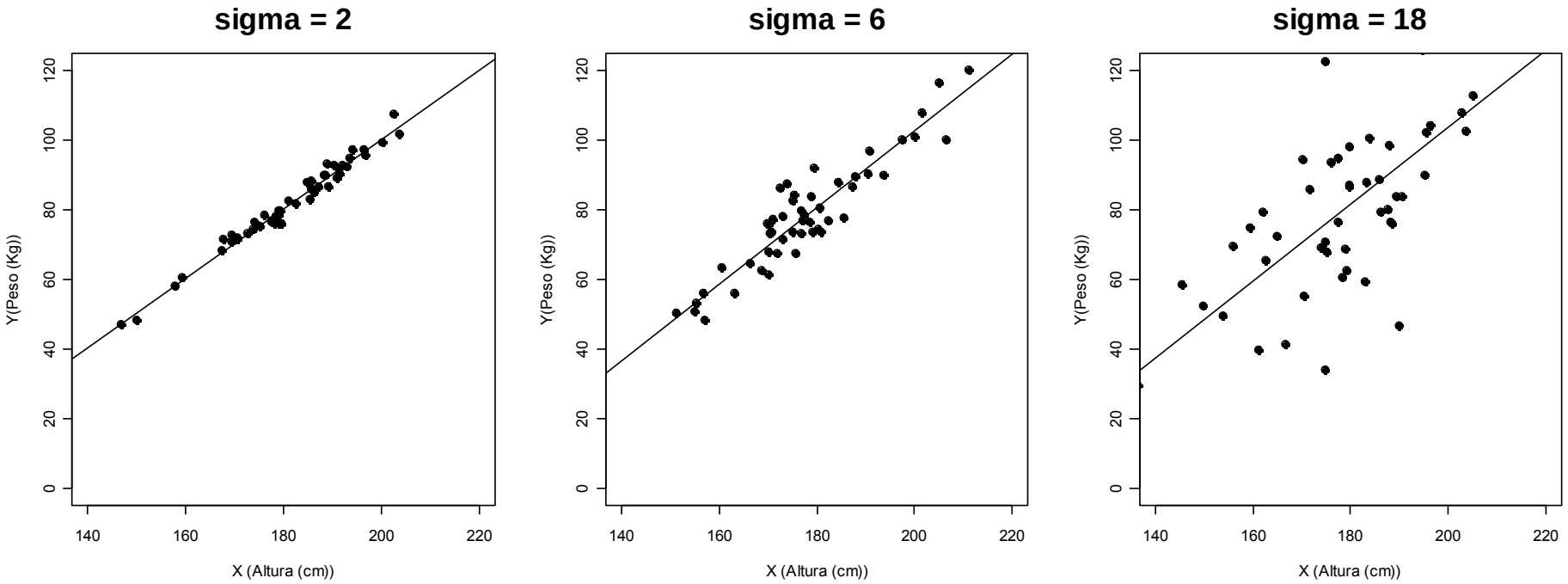
Què significa σ = 6Kg ? *Solució: ...*

Què opina de la etiqueta 'pes ideal' en algunes farmàcies? *Solució: ...*

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

Coeficient R^2

Es va veure que: $r_{XY} = r = \dfrac{S_{XY}}{S_X S_Y}$

De fet, r estudia la relació (lineal) entre dues variables X i Y amb un rol simètric.

Definim el coeficient R^2 (*Coeficient de determinació*, o *R-squared*), com el quadrat de la correlació lineal r. Noteu que **$0 \leq R^2 \leq 1$** ...

Ve a significar quina fracció de la variabilitat de Y s'explica per el factor X (la interpretació torna a ser asimètrica).

- Tingueu present que 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

- r és un indicador d'associació entre dues variables relacionades linealment, però no suposa cap model al darrera (caràcter descriptiu)

REGRESSIÓ: estimació dels paràmetres

β_0 , β_1 , σ^2 i ε_i són valors poblacionals, *autèntics*, desconeguts, a 'estimar'.

L'estimació dels dos primers, dóna lloc a la recta estimada:

$$\hat{y}_i = \mathbf{b_0} + \mathbf{b_1 x_i}$$

que permet fer predicciones per a cada observació amb el seu error de predicció

$$e_i = y_i - \hat{y}_i$$

L' estimació mínim quadràtica (annexe 6.12 d'*Estadística per a enginyers informàtics*. Ed UPC) 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: $\sum(e_i)^2 = \sum(y_i - \hat{y}_i)^2 = \sum(y_i - b_0 - b_1 X_i)^2$

La solució al problema de minimització és el següent: (Ref: *Eei.Ed.UPC* pg144)

$$b_1 = \frac{S_{XY}}{S_X^2} = r_{XY}\frac{S_Y}{S_X} \qquad b_0 = \overline{Y} - b_1\overline{X} \qquad S^2 = \frac{\sum e_i^2}{n-2}$$

(Recordeu que $\mathbf{S_{XY}}$ és la covariança mostral i $\mathbf{r_{XY}}$ la correlació mostral)

(a l'enllaç *Regression by Eye* a Laboratori a http://www-eio.upc.es/teaching/pe/ podeu comprovar (*Show*) el valor de *r* i el de *MSE* de les rectes ajustades amb el valor mínim que s'obté amb la recta de regressió)

EXEMPLE: Cervesa i contingut d'alcohol a la sang

cerveses	alcohol
5	0.1
2	0.03
9	0.19
8	0.12
3	0.04
7	0.095
3	0.07
5	0.06
3	0.02
5	0.05
4	0.07
6	0.1
5	0.085
7	0.09
1	0.01
4	0.05

Usant que:

$$\bar{y} = \frac{\sum y_i}{n}$$

$$s_Y^2 = \frac{\sum y_i^2 - \frac{(\sum y_i)^2}{n}}{n-1}$$

$$s_{XY} = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{n-1}$$

calculem:

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

$$\bar{x} = 4.8125 \quad s_X^2 = 4.829167$$

$$s_{XY} = 0.08675$$

$$r_{XY} = s_{XY} / s_X s_Y = 0.894338$$

$$\sum x_i = 77 \quad \sum x_i^2 = 443$$

$$\sum y_i = 1.18 \quad \sum y_i^2 = 0.11625$$

$$\sum x_i y_i = 6.98$$

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

Call:
lm(formula = alc ~ n.cerv)

Coefficients:
(Intercept)       n.cerv
   -0.01270      0.01796
```

Resultats de la regressió:

$$b_1 = \frac{S_{XY}}{S_X^2} = r_{XY}\frac{S_Y}{S_X} =$$

$$b_0 = \overline{Y} - b_1 \overline{X} = \qquad \hat{y}_i =$$

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

(variància de l'error amb R: sum(lm(alc ~ n.cerv)$resid ^2)/14

REGRESSIÓ: 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** és necessària per construir el model, però secundària en sí mateixa]

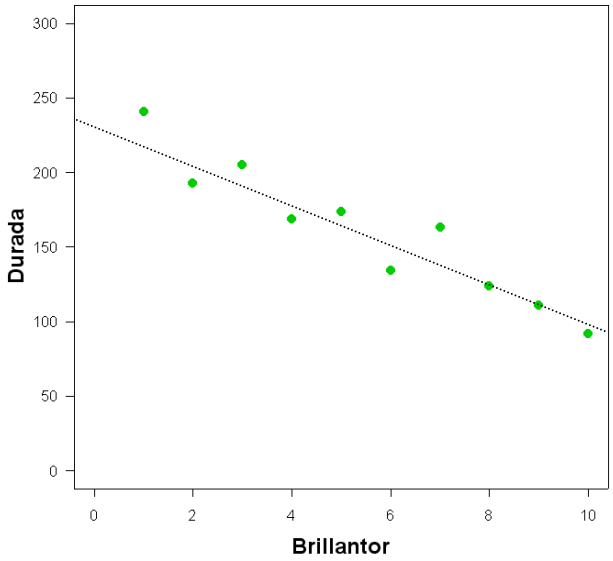
EXERCICI: **Pantalla d'ordinador**

(Ref: *Eei.Ed.UPC* pg 143)

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:

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

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



R: > lm(Durada~Brillantor)

$\bar{y} =$

$s_y^2 =$

$\bar{x} =$

$s_x^2 =$

$s_{xy} =$

$r_{XY} = s_{XY} / s_X s_Y =$

$b_1 = \frac{S_{XY}}{S_X^2} = r_{XY} \frac{S_Y}{S_X} =$

$b_0 = \overline{Y} - b_1 \overline{X} =$

$\hat{y}_i =$

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

REGRESSIÓ: distribució dels estimadors

DISTRIBUCIÓ DELS ESTIMADORS MÍNIMS QUADRÀTICS

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

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

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

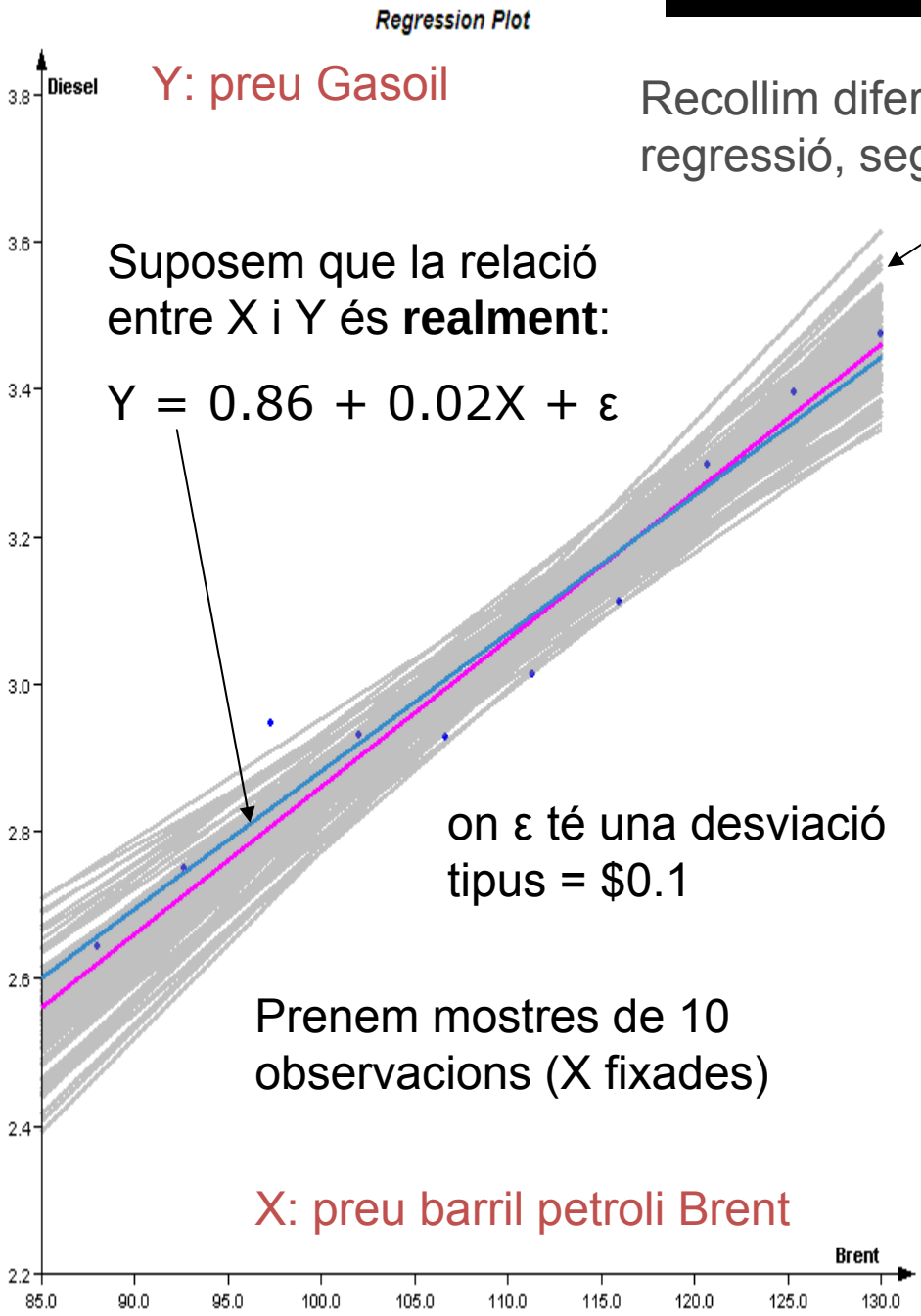
$$b_0 \sim N(\beta_0, \sigma^2(\frac{1}{n} + \frac{\bar{X}^2}{(n-1)S_X^2}))$$

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

EXEMPLE d'aplicació: 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 ($S_{b_1} = \sqrt{\frac{S^2}{(n-1)S_X^2}}$), ¿què es pot fer?

Solucions per disminuir S_{b_1} : intentar 'controlar' les fonts de variació en σ^2 ; augmentar 'n';
i ampliar la 'finestra' de l'estudi S^2_X

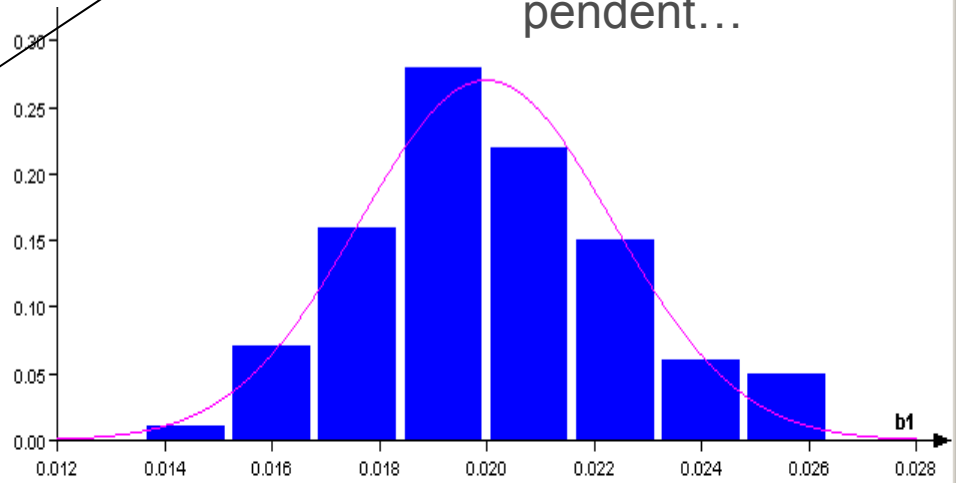
EXEMPLE



Y: preu Gasoil

Recollim diferents rectes de regressió, segons la mostra.

… observem diferents valors del pendent…



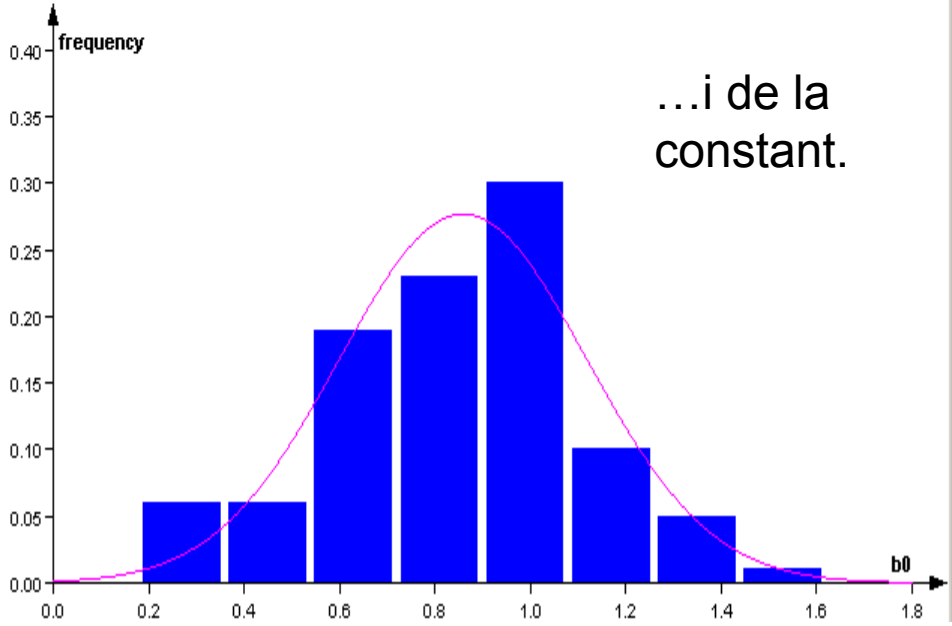
Suposem que la relació entre X i Y és **realment**:

$$Y = 0.86 + 0.02X + \varepsilon$$

on ε té una desviació tipus = $0.1

Prenem mostres de 10 observacions (X fixades)

X: preu barril petroli Brent



…i de la constant.

REGRESSIÓ: inferència

Podem realitzar la inferència habitual (IC, PH) per β_0 , β_1 i σ^2 a partir de:

$$b_1 = \frac{S_{XY}}{S_X^2} \qquad \frac{(b_1 - \beta_1)}{S_{b_1}} = \frac{(b_1 - \beta_1)}{\sqrt{\frac{S^2}{(n-1)S_x^2}}} \sim t_{n-2}$$

$$b_0 = \overline{Y} - b_1 \overline{X} \qquad \frac{(b_0 - \beta_0)}{S_{b_0}} = \frac{(b_0 - \beta_0)}{\sqrt{S^2\left(\frac{1}{n} + \frac{\overline{X}^2}{(n-1)S_x^2}\right)}} \sim t_{n-2}$$

(S^2 és l'estimador de σ^2)

(els graus de llibertat són n-2 per les dues restriccions al necessitar estimar dos paràmetres previs)

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

Estimadors

Estadístics i la seva distribució

RESUM DE LES PROPIETATS DELS ESTIMADORS

Paràmetre	**β_0**	**β_1**	**σ^2**
Estimador	$b_0 = \bar{Y} - b_1 \bar{X}$	$b_1 = S_{XY} / S^2_X$	$S^2 = \Sigma e_i^2/(n-2)$
Esperança	$E(b_0) = \beta_0$	$E(b_1) = \beta_1$	$E(S^2) = \sigma^2$
Variància	$V(b_0) = \sigma^2(\frac{1}{n} + \frac{\bar{X}^2}{(n-1)S_x^2})$ $(S_{b_0} = \sqrt{S^2(\frac{1}{n} + \frac{\bar{X}^2}{(n-1)S_x^2})})$	$V(b_1) = \sigma^2 / (n-1)S_x^2$ $(S_{b_1} = \sqrt{S^2 / (n-1)S_x^2})$	$V(S^2) = 2\sigma^4/(n-2)$
Distribució	$b_0 \sim N$ $(b_0 - \beta_0) / S_{b0} \sim t_{n-2}$.	$b_1 \sim N$ $(b_1 - \beta_1) / S_{b1} \sim t_{n-2}$.	$(n-2)S^2/\sigma^2 \sim \chi^2_{n-2}$
Interval de Confiança	$IC(95\%,\beta_0) =$ $= b_0 \pm t_{n-2,0.975} \cdot S_{b0}$	$IC(95\%,\beta_1) =$ $= b_1 \pm t_{n-2,0.975} \cdot S_{b1}$	$IC(95\%,\sigma^2) \rightarrow$ $(n-2)S^2/\chi^2_{n-2,0.975} \leq \sigma^2$ $\leq (n-2)S^2/\chi^2_{n-2,0.025}$
H_0 usual	$\beta_0 = 0$	$\beta_1 = 0$	
Rebutgem H_0 si	$b_0 / S_{b0} > t_{n-2,0.975}$	$b_1 / S_{b1} > t_{n-2,0.975}$	

EXEMPLE: Cervesa i contingut d'alcohol a la sang

Teníem:

$$\hat{y}_i = \mathbf{-0.0127 + 0.0180\ X_i}$$

amb $b_1 = 0.0180$

$s_X^2 = 4.8292$

$S^2 = 0.000418$

Per tant,

$$S_{b_1} = \sqrt{\frac{S^2}{(n-1)S_x^2}} = \mathbf{0.0024}$$

```
R: > summary(lm(alc ~n.cerv))

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
```

Objectiu : **'saber' si β_1 és diferent de zero**

Hipòtesi : H : $\beta_1 = 0$ (bilateral, per simplicitat)

Estadístic: $\hat{t} = \dfrac{(b_1 - \beta_1)}{\sqrt{\dfrac{S^2}{(n-1)S_X^2}}} = \dfrac{(b_1 - \beta_1)}{S_{b_1}}$ Distribució sota H: t_{n-2} *(premisses a indicar i analitzar)*

Càlculs: $\hat{t} =$ **7.48** P-valor= **3.0e-06 (2*(1-pt(7.48,14))), per tant << 0.05**

Punt crític = **2.145 (qt(0.975,14))), per tant 7.48 > 2.145**

Decisió : No és versemblant que el coeficient del pendent sigui 0

IC(95%,β_1) = $b_1 \pm t_{n-2,0.975} \cdot S_{b1}$ = [0.0128 ; 0.0231]

Conclusió pràctica: *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 95% de confiança.*

EXEMPLE: Cervesa i contingut d'alcohol a la sang

Teníem:

$$\hat{y}_i = \mathbf{-0.0127 + 0.0180\ X_i}$$

amb $b_1 = 0.0180$

$s_X^2 = 4.8292$ $\quad \overline{X} = 4.8125$

$S^2 = 0.000418$

Per tant,

$$S_{b_0} = \sqrt{S^2\left(\frac{1}{n} + \frac{\overline{X}^2}{(n-1)S_x^2}\right)} = \quad \mathbf{0.0126}$$

```
R: > summary(lm(alc ~n.cerv))

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
```

Objectiu: **'saber' si β_0 és diferent de zero**

Hipòtesi: $H_0 : \beta_0 = 0$ (bilateral, per simplicitat)

Estadístic: $\hat{t} = \dfrac{(b_0 - \beta_0)}{S_{b_0}}$ Distribució sota H: t_{n-2} *(premisses a indicar i analitzar)*

Càlculs: $\hat{t} =$ **-1.0** P-valor= **0.33 (2*(1-pt(1.0,14))), per tant 0.33 > 0.05**

Punt crític = **2.145 (qt(0.975,14))), per tant |-1.0| < 2.145**

Decisió : SI que és versemblant que el coeficient de la constant a l'origen sigui 0

IC(95%,β_0) = $b_0 \pm t_{n-2,0.975} \cdot S_{b0}$ = [-0.0397; 0.0143]

Conclusió pràctica: No es pot rebutjar que la recta passi per l'origen, pel punt (0,0):

a 0 llaunes de cervesa li correspon alcohol 0.0%.

EXERCICI: **Pantalla d'ordinador** (n=10)

Teníem:

R: > summary(lm(Durada~Brillantor))

$$\hat{y}_i = \mathbf{239.9 - 14.41\ X_i}$$

amb $b_1 = -14.41$

$s_x^2 = 9.167$

$S^2 = 227.3$

Per tant,

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

Objectiu: 'saber' si β_1 és diferent de zero

Hipòtesi: $H_0 : \beta_1 = 0$ (bilateral, per simplicitat)

Estadístic: Distribució sota H: t_{n-2} *(premisses a indicar i analitzar)*

Càlculs: $\hat{t} =$ P-valor=

Punt crític =

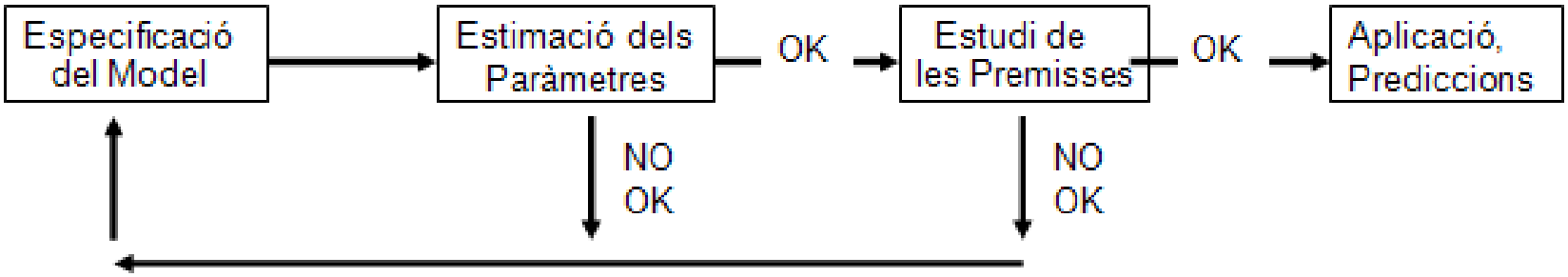
Decisió :

IC(95%,β_1) = $b_1 \pm t_{n-2,0.975} \cdot S_{b1}$ =

Conclusió pràctica:

Fases del procés de models estadístics

(Capítol 7 d'*Estadística per a enginyers informàtics*. Ed UPC)



Un cop especificat el model i estimats els paràmetres, perquè sigui útil (aplicar-lo, fer prediccions), cal estudiar les premisses assumides (una anàlisis 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**.

Validació model lineal: anàlisi dels residus

L'anàlisi de les premisses en la variable de resposta en regressió es poden traspassar a la part determinista (recta) i la aleatòria (residual).

En la part determinista:

Linealitat entre X i Y en el rang considerat

En la part aleatòria:

com que X_i no és v.a, és constant, no està mesurada amb error,

llavors $V(y_i)=V(\beta_0 + \beta_1 X_i + e_i)=V(e_i)=\sigma^2$, i així

les premisses sobre la part aleatòria de y_i les analitzem sobre els residus e_i

e_i són v.a. i.i.d. amb una D. Normal $N(0,\sigma^2)$ (e_i soroll blanc)

Homoscedasticitat: mateixa σ^2 per qualsevol i

Independència: un error no porta informació sobre el valor de l'altre

Normalitat: resultat de molts fenòmens aleatoris amb pesos petits

Els models en estudi són una representació simplificada de la realitat, per això més que l'acompliment exacte de les premisses interessa un grau raonable de compliment perquè el model sigui útil.

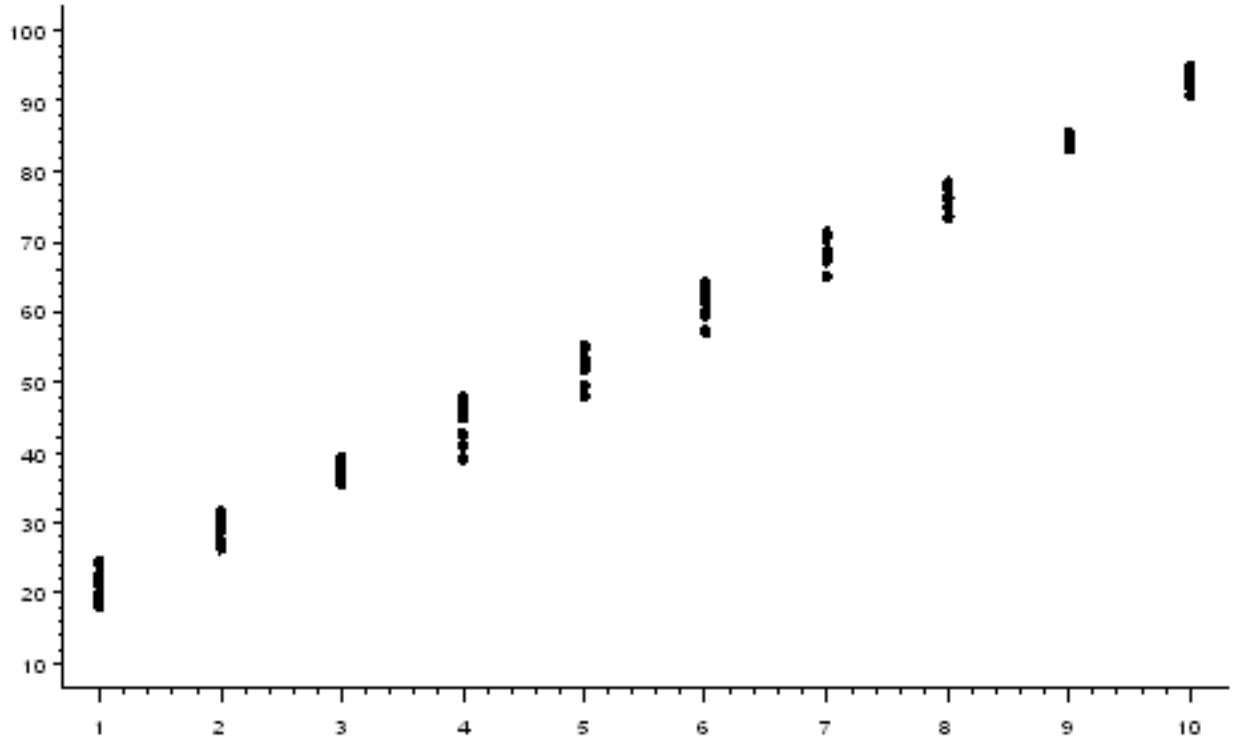
El compliment de les premisses anteriors permet poder recórrer a les distribucions de referència (per fer IC, PH) i garanteix que el model és el millor possible (si alguna no es compleix pot disminuir-ne l'eficiència).

L'anàlisi de les premisses estudia si són raonables o, en cas contrari, per trobar un model alternatiu que encaixi millor.

La validació es basa en el coneixement previ teòric i en anàlisis gràfiques:

- e_i vs X_i (també e_i vs "FittedValues")
- e_i vs ordre observacions
- gràfic de probabilitat normal (qqnorm, millor que l'histograma de residus e_i)

Gràfic Y_i versus X_i



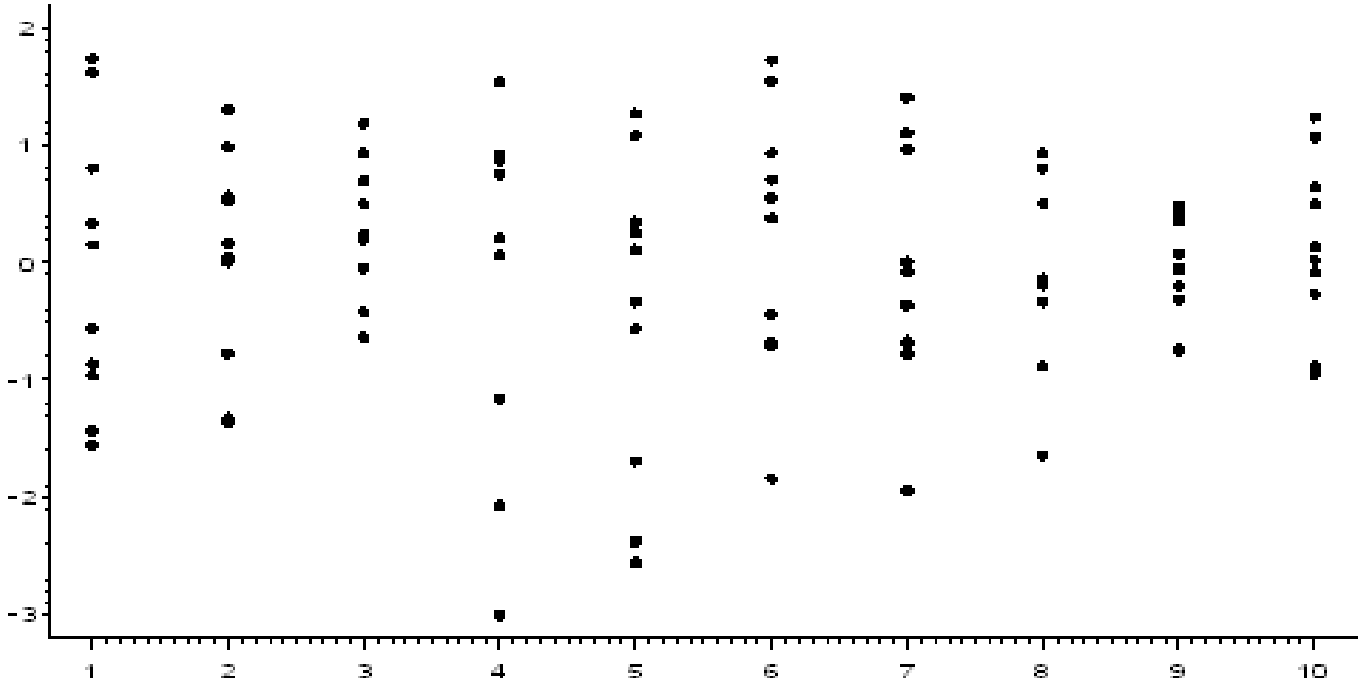
Permet estudiar la linealitat i la homoscedasticitat.

(En aquest exemple: OK tots dos)

És molt fàcil i intuïtiu, però ineficient: molts espais en blanc

Es pot millorar, substituint Y pels residus.

Gràfic e_i versus x_i (mateixes dades)

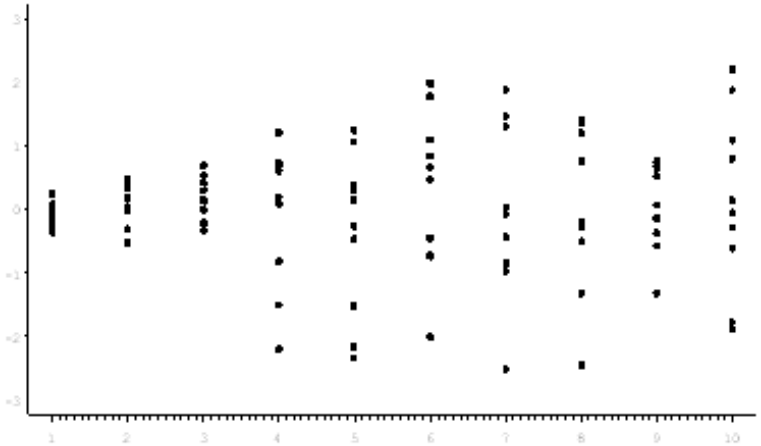


Linealitat: totes les mitjanes a la mateixa alçada (aprox.).

Homoscedasticitat: Encara que semblen variar les S, cal recordar que S té molta oscil·lació mostral: de fet han estat generades totes amb la mateixa σ

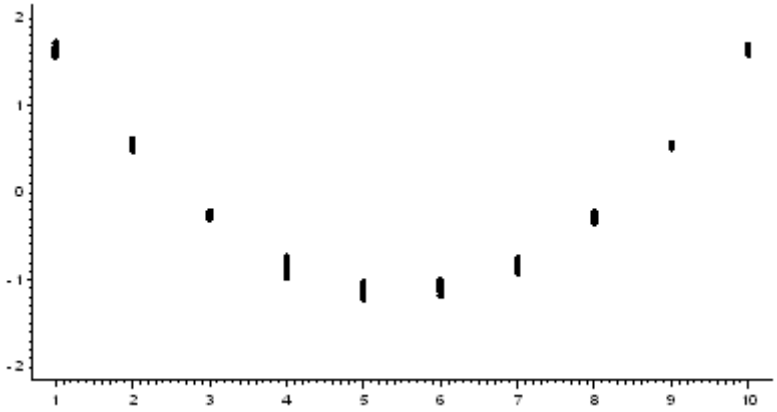
Gràfic e_i versus x_i

- Exemple de **heteroscedasticitat**:



La variància de les pertorbacions augmenta amb X.

- Exemple de **no linealitat**:

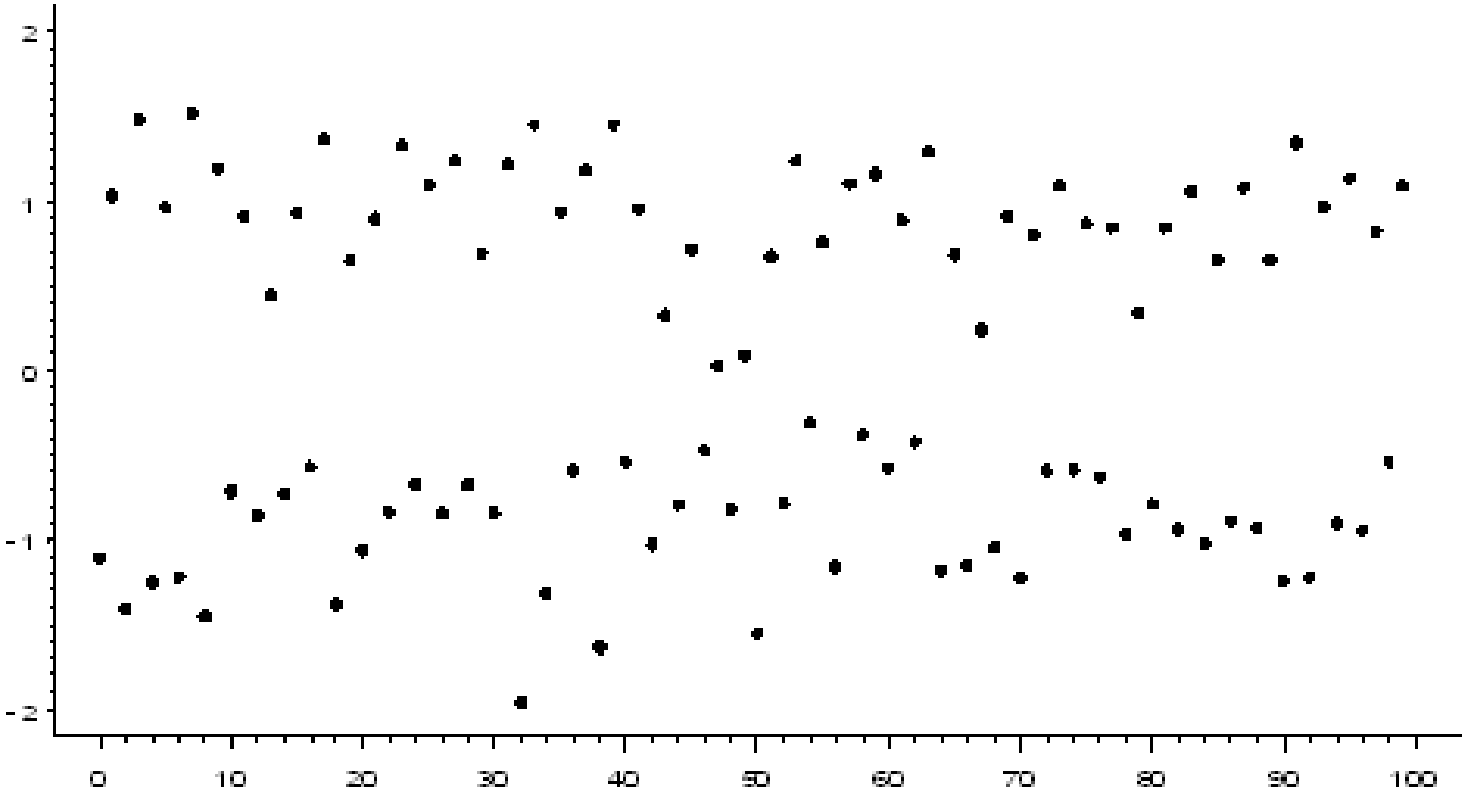


Suggeriment:

proveu amb les dades del consum de benzina i velocitat

Una variable interessant és l'ordre de les observacions.

Gràfic e_i versus ordre observacions



Aquestes dades mostren un patró: residus + i – s'alternen.
No hi ha independència entre observacions consecutives.
És típica de *series temporals*: variables recollides al llarg del temps.
p.e.: les hores dormides en dies consecutius solen tenir una correlació -

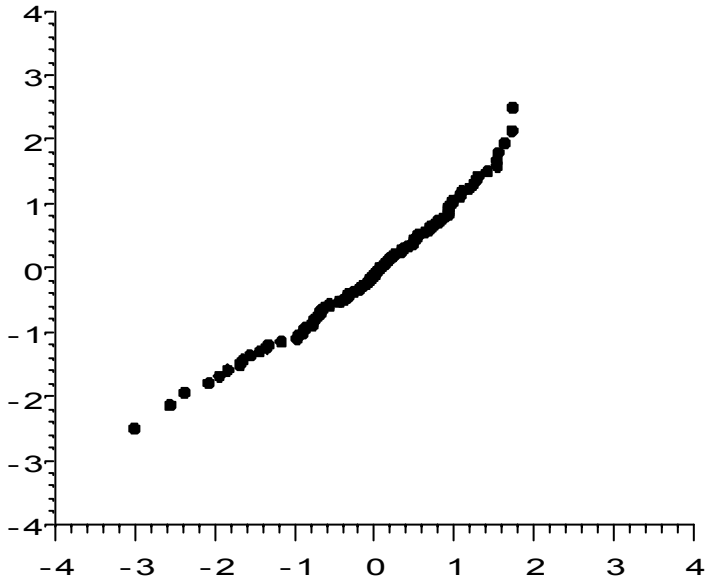
Gràfic de probabilitat normal (qqnorm(), Recta de Henry,...)

Representa la mateixa variable:

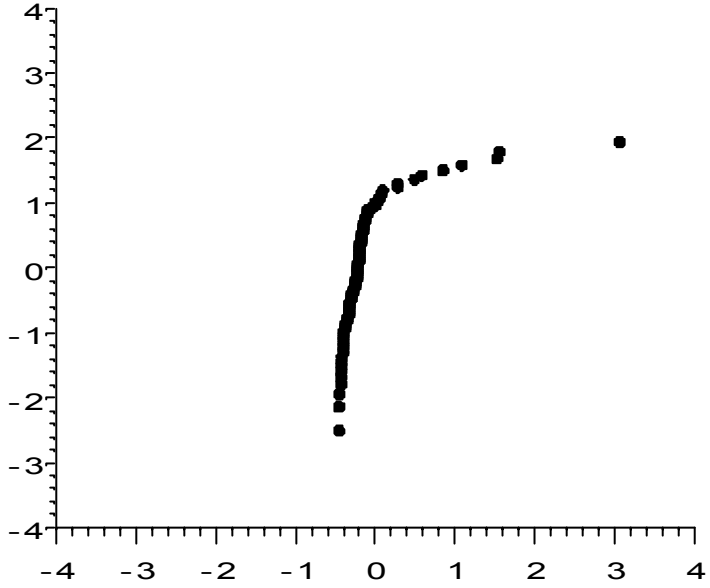
A les abscisses, els valors observats: Residus tipificats E(e)=0, V(e)=1

A les ordenades, con si fossin normals: Residus normalitzats N(0,1)

Si coincideixen en un recta: observats=normals



Normalitat

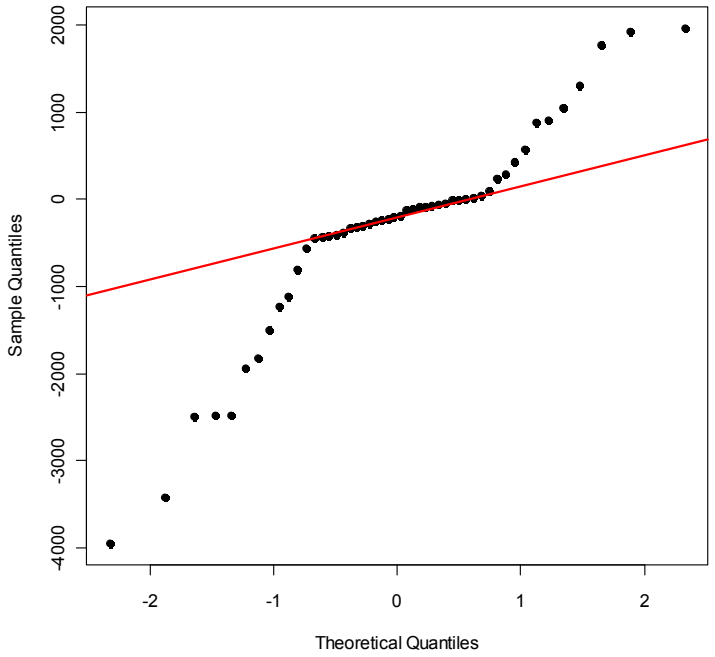


No normalitat

Hi ha moltes variants: les variables poden estar tipificades o no.

EXEMPLE: NORMALITAT (?)

Aquesta variable ha estat simulada a partir d'un model NO Normal, i s'interpreta correctament la seva NO normalitat.
Si qualifiquem les següents expressions, de més correcta (4) a més incorrecta (1)



Aquesta variable NO pot ser representada amb la D. Normal.

Aquesta variable pot ser modelada amb la D. Normal.

Aquesta variable segueix la D. Normal

Aquesta variable NO segueix la D. Normal

La seva aproximació a la Normal es suficient (o raonable)

La seva aproximació a la Normal NO és suficient (o raonable)

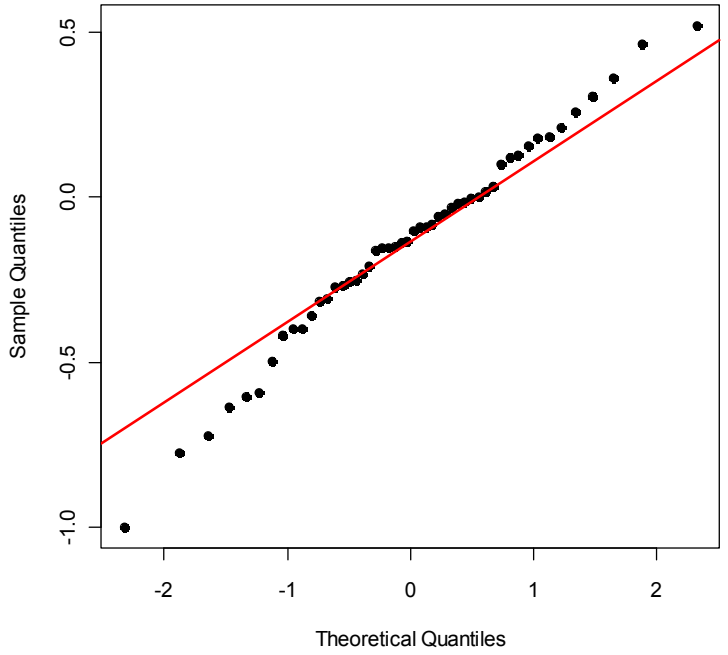
Hem demostrat que la variable és Normal

Hem demostrat que la variable NO és Normal

EXEMPLE: NORMALITAT (?)

Aquesta variable ha estat simulada a partir d'un model Normal, i s'interpreta correctament la seva normalitat.

Si qualifiquem les següents expressions, de més correcta (4) a més incorrecta (1):



Aquesta variable NO pot ser representada amb la D. Normal.

Aquesta variable pot ser modelada amb la D. Normal.

Aquesta variable segueix la D. Normal

Aquesta variable NO segueix la D. Normal

La seva aproximació a la Normal es suficient (o raonable)

La seva aproximació a la Normal NO és suficient (o raonable)

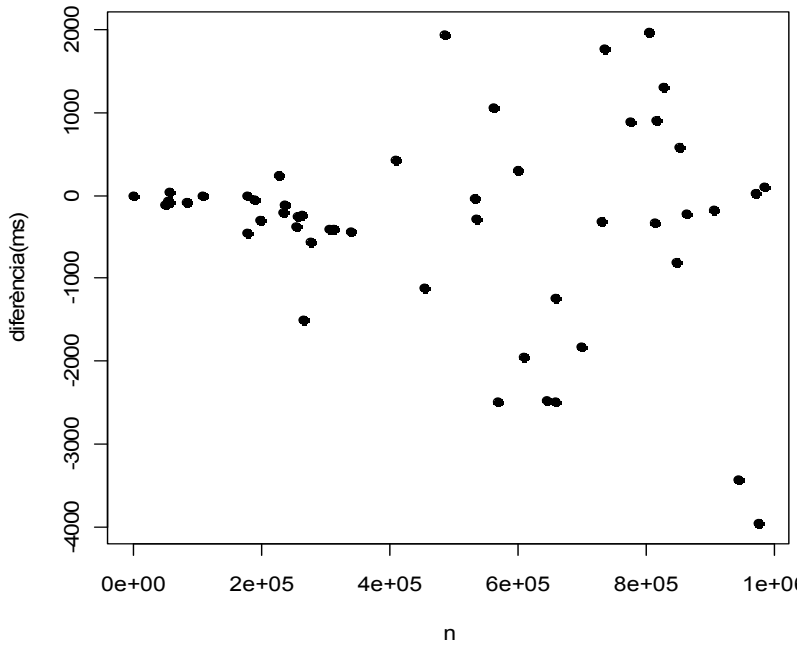
Hem demostrat que la variable és Normal

Hem demostrat que la variable NO és Normal

EXEMPLE: HOMOSCEDASTICITAT (?)

Aquesta variable ha estat simulada a partir d'un model amb variàncies creixents, i s'interpreta correctament la seva heteroscedasticitat.

Si qualifiquem les següents expressions, de més correcta (4) a més incorrecta (1)



Aquesta variable NO pot ser representada amb un model homoscedàstic

Aquesta variable pot ser modelada assumint homoscedasticitat

Aquesta variable és homoscedàstica

Aquesta variable NO és homoscedàstica

La seva aproximació a la homoscedasticitat es suficient (o raonable)

La seva aproximació a la homoscedasticitat NO és suficient (o raonable)

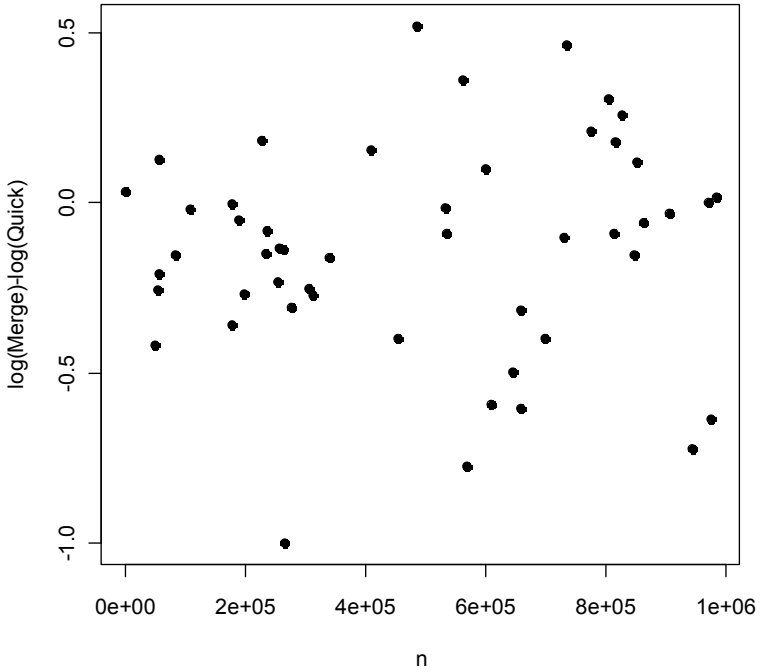
Hem demostrat que la variable és homoscedàstica

Hem demostrat que la variable NO és homoscedàstica

EXEMPLE: HOMOSCEDASTICITAT (?)

Aquesta variable ha estat simulada a partir d'un model amb igualtat de variàncies, i s'interpreta correctament la seva homoscedasticitat.

Si qualifiquem les següents expressions, de més correcta (4) a més incorrecta (1)



Aquesta variable NO pot ser representada amb un model homoscedàstic

Aquesta variable pot ser modelada assumint homoscedasticitat

Aquesta variable és homoscedàstica

Aquesta variable NO és homoscedàstica

La seva aproximació a la homoscedasticitat es suficient (o raonable)

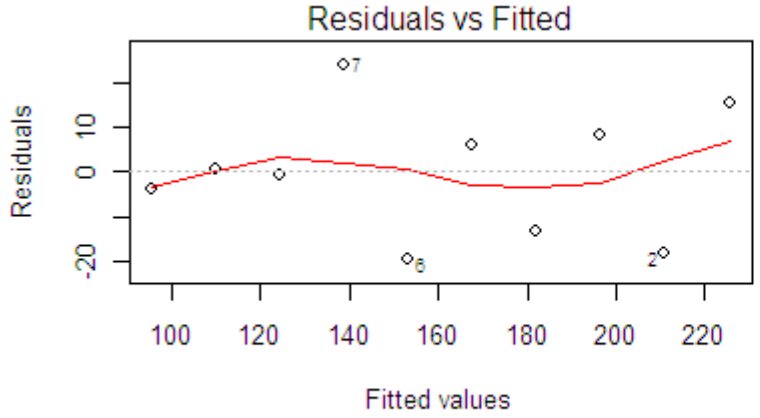
La seva aproximació a la homoscedasticitat NO és suficient (o raonable)

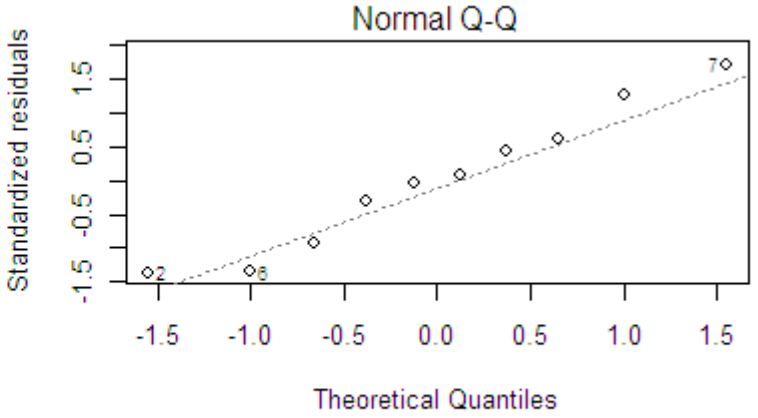
Hem demostrat que la variable és homoscedàstica

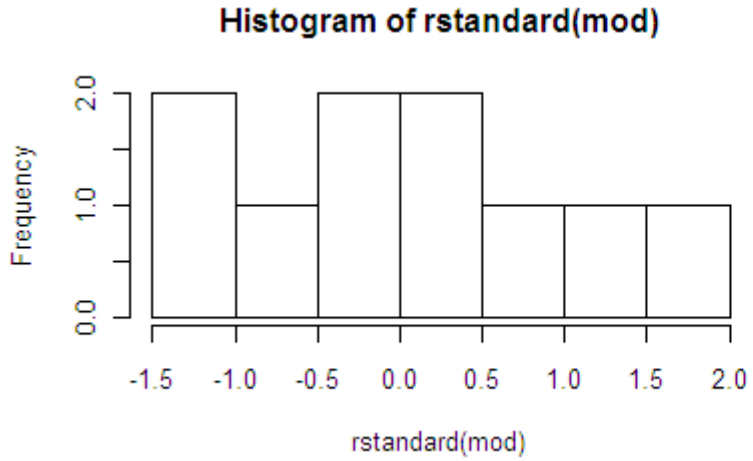
Hem demostrat que la variable NO és homoscedàstica

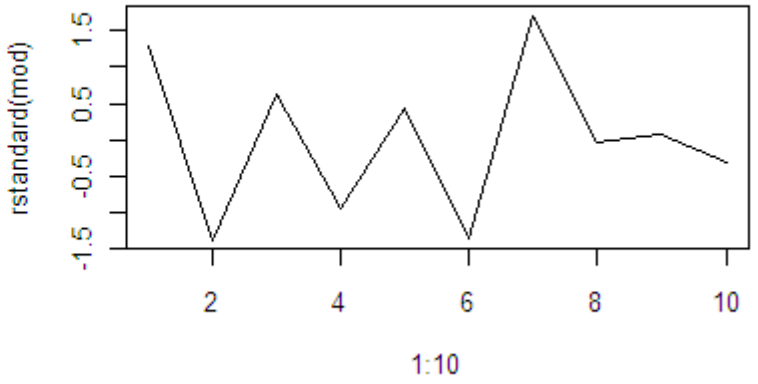
EXEMPLE: Pantalla d'ordinador

```
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
                estandaritzats
```



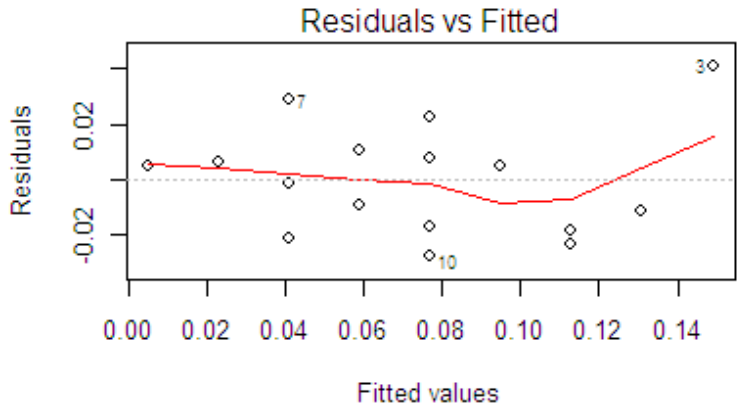


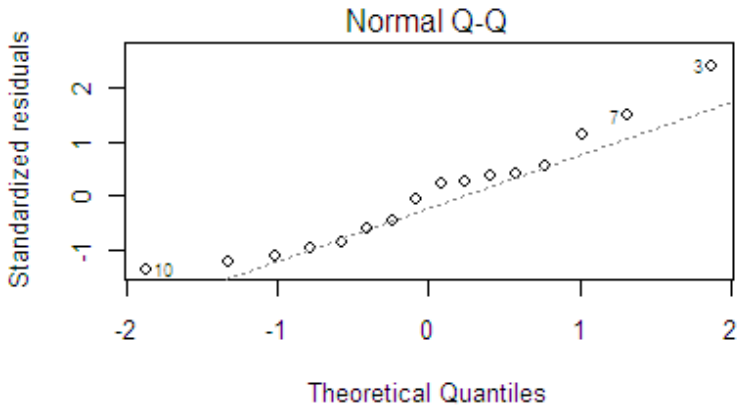


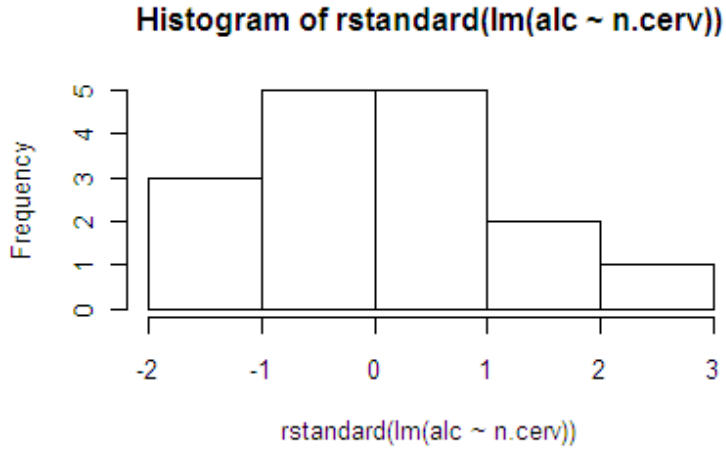


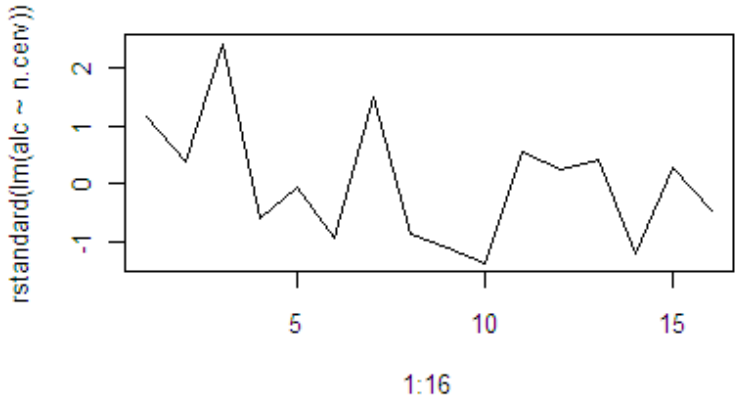
EXEMPLE: Cervesa i contingut d'alcohol a la sang

```
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
```









Predicció
(REGRESSIÓ)

En primer lloc la predicció puntual de Y per a valors concrets de X (X_h) usa la part determinista: $\hat{y}_h = b_0 + b_1 X_h$

Però, com tenir en compte la part aleatòria? Dues situacions ben diferenciades:

1) estimar la variabilitat del valor esperat per a les observacions X=x_h

2) predir un interval pel valor individual corresponent a X=x_h

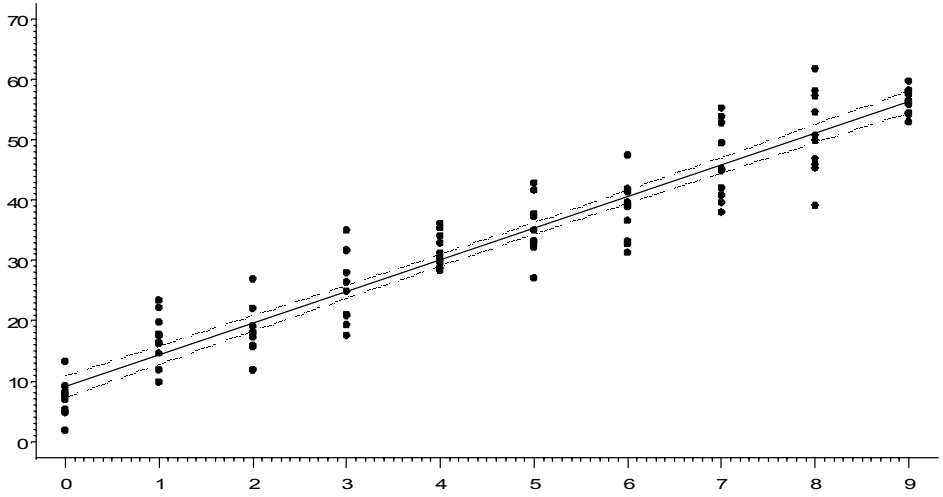
1) per **estimar** esperança i variança de la predicció, utilitzarem

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

i llavors,

$$E(\hat{y}_h) = E(b_0 + b_1 X_h) = \beta_0 + \beta_1 x_h = \mu_h$$

(és no esbiaixat!)



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

(major variància a major distància de $\overline{X}$)

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

2) Per **predir** l'interval dels valors individuals y_h de Y per $X=x_h$

utilitzarem també $y_h = \hat{y}_h = b_0 + b_1 X_h$

I llavors

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

amb Error Quadràtic Mitjà de Predicció

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

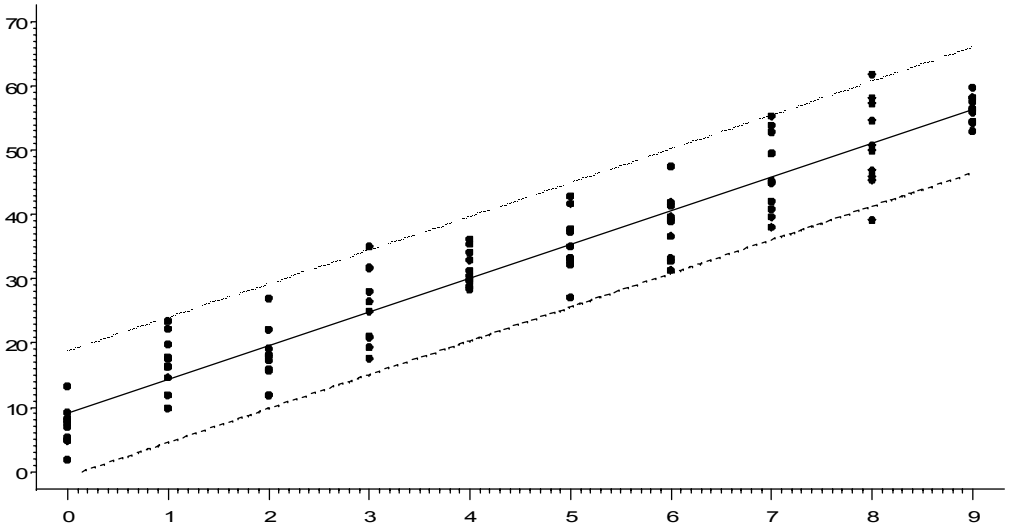
que es pot descomposar

de forma semblant

a la descomposició

de sumes de quadrats:

$$E(y_h - \hat{y}_h)^2 = E(\hat{y}_h - m_h)^2 + E(y_h - m_h)^2$$



donant lloc a

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

que permet identificar 3 fonts de variabilitat en la predicció dels valors individuals:

Natural (σ^2) **+ Per estimació mitjana** (σ^2/n) **+ Per estimació pendent**

Resum de previsions de la resposta

Estimació puntual	$\hat{y}_h = b_0 + b_1 X_h$	$y_h = \hat{y}_h = b_0 + b_1 X_h$
Estimació per interval	Per al valor esperat $\hat{y}_h \pm t_{n-2,0.975} S \sqrt{(1/n + \frac{(X_h - \bar{X})^2}{\sum (X_i - \bar{X})^2})}$	Per a valors individuals $\hat{y}_h \pm t_{n-2,0.975} S \sqrt{(1 + 1/n + \frac{(X_h - \bar{X})^2}{\sum (X_i - \bar{X})^2})}$

EXEMPLE: **Pantalla d'ordinador** Sabíem: $\hat{y}_i = 239.9 - 14.41\ X_i$ i $\bar{X} = 5.5$

$S_x^2 = 9.167$

$S^2 = 227.3$

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

Quina durada podem esperar per a pantalles de brillantor 7.5?

Estimació puntual	$\hat{y}_{7.5} = b_0 + b_1 X_{7.5} =$	$y_{7.5} = \hat{y}_{7.5} = b_0 + b_1 X_{7.5} =$
Estimació per interval	Per al valor esperat $\hat{y}_{7.5} \pm t_{8,0.975} S \sqrt{1/10 + \frac{(X_{7.5} - \bar{X})^2}{\sum (X_i - \bar{X})^2}} =$	Per a valors individuals $\hat{y}_{7.5} \pm t_{8,0.975} S \sqrt{1 + 1/10 + \frac{(X_{7.5} - \bar{X})^2}{\sum (X_i - \bar{X})^2}} =$
	Per a pantalles de brillantor esperada de 7.5 podem esperar una durada entre ...	Per a una pantalla de brillantor 7.5 podem esperar una durada entre ...

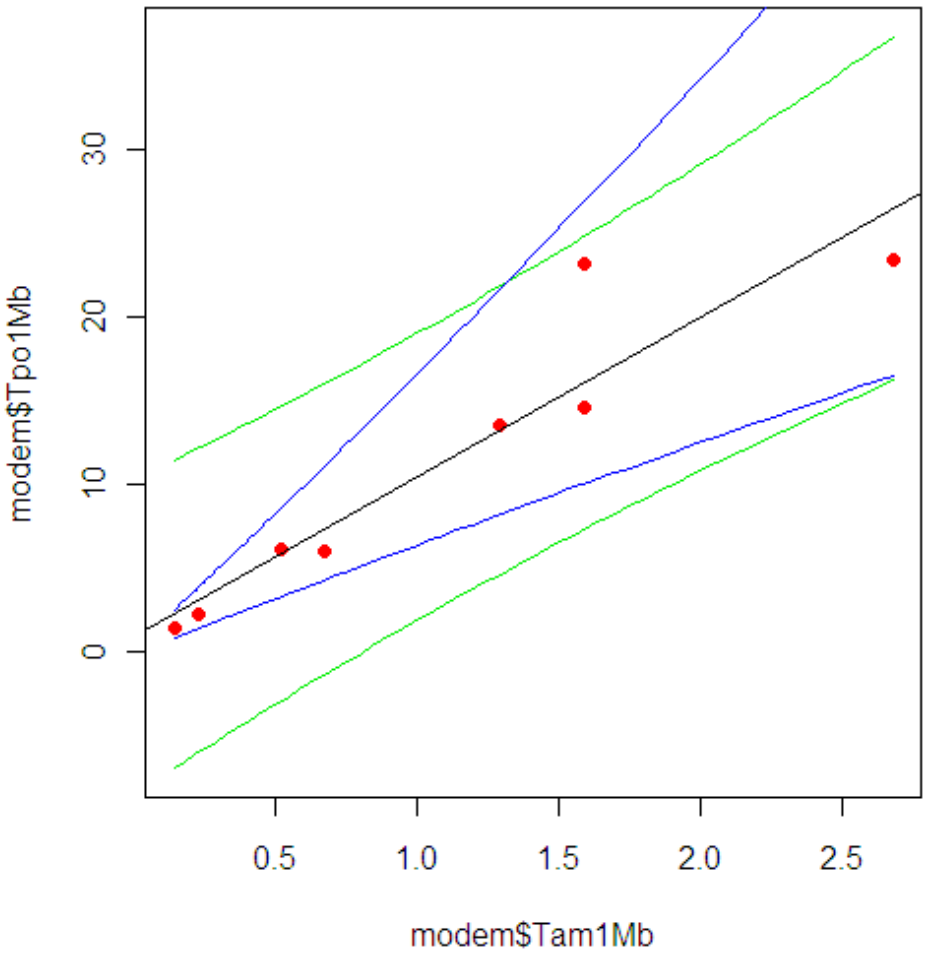
(Veure gràfics de pags. 191-192 a *Estadística per a enginyers informàtics*. Ed UPC)

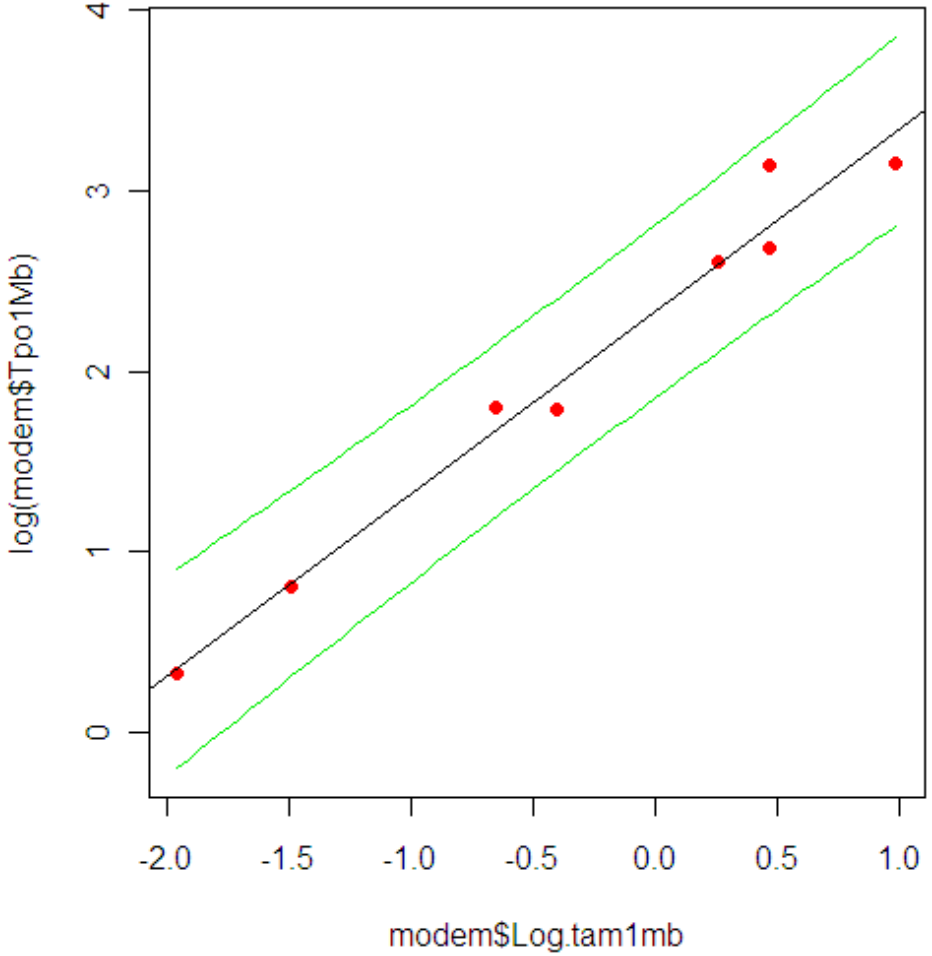
EXEMPLE: **FTP** *Baixant fitxers amb modem 1Mbps*

Resposta: temps [s]. Var. explicativa: mida fitxer [MB]

Model #1: temps vs mida. Problema, heterocedasticitat. 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

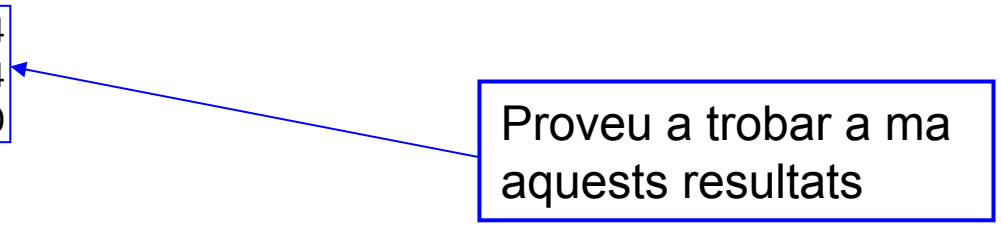




EXEMPLE: **FTP** *Baixant fitxers amb modem 1Mbps*

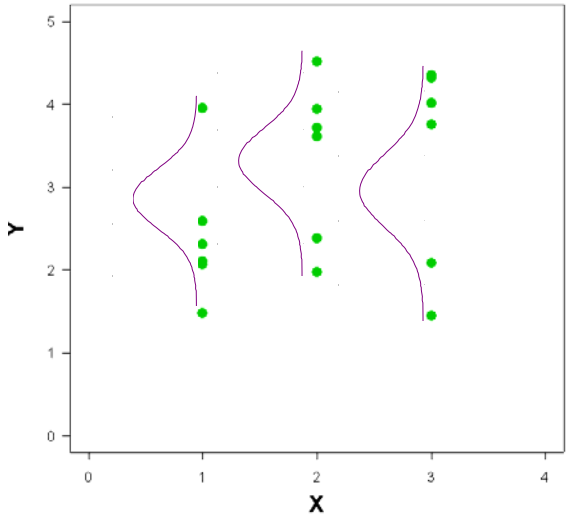
Detalls

```
> 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
...
```

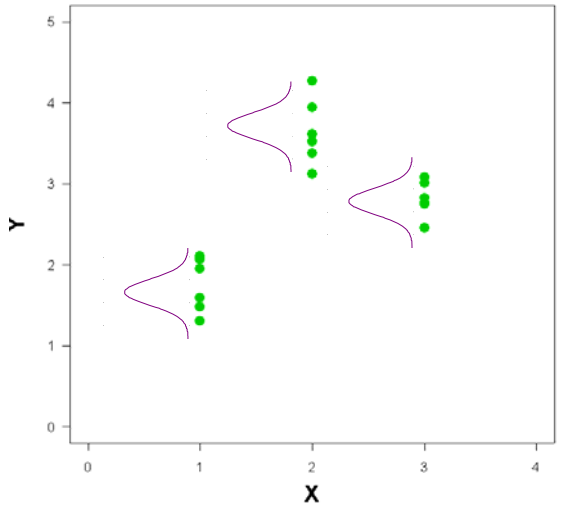


ANNEX: Model "quantitativa vs categòrica"

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 i intra-grupos

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)

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

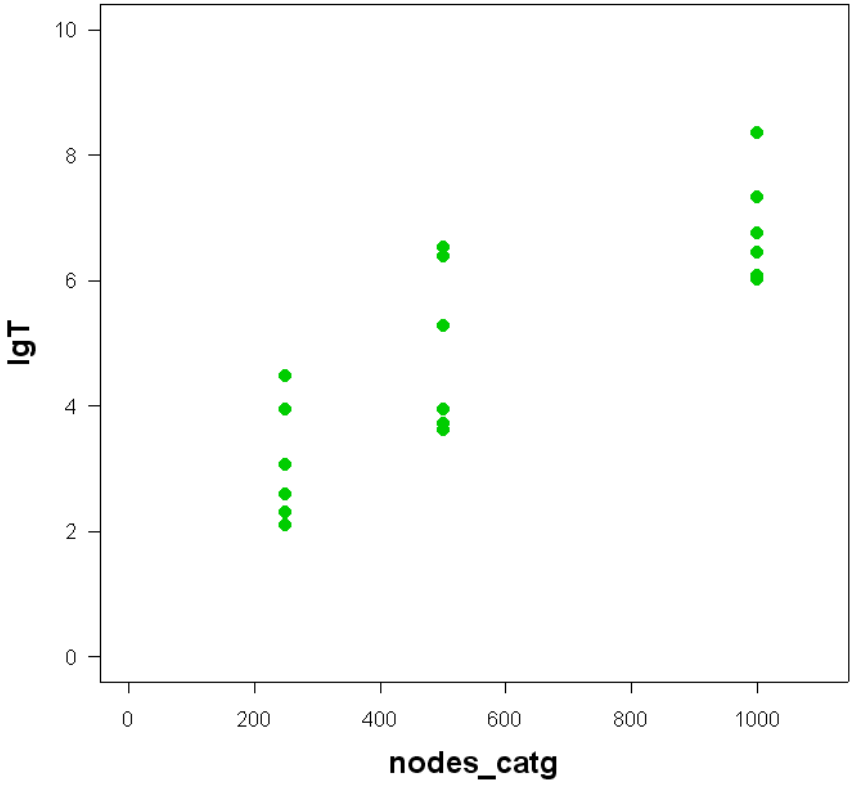
Si ho fem estudiant les característiques en grafs de 250, 500 i de 1000 nodes:

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

També té sentit estudiar la descomposició de la variabilitat en:

-deguda al factor X i

-aleatòria o residual.



Model "quantitativa vs categòrica": model, paràmetres

Ara el model és: $\mathbf{Y_{ij} = \mu_j + \varepsilon_{ij}}$

amb
- 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 $\bar{y}$) (desviació $\mathbf{s_j}$)
- ε_{ij} error aleatori o diferència del cas i a la mitjana del seu grup j
  (el paràmetre σ^2 és la variància de ε_i o variància residual

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 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:

$\bar{y}_1 = 6.15, \bar{y}_2 = 5.73, \bar{y}_3 = 5.48, S_1 = 1.8, S_2 = 1.5, S_3 = 2.0$

Què val la mitjana global? *Solució:* $\bar{Y} = \mathbf{5.81}$

Què val la variabilitat combinada dins els grups? *Solució:* $S^2 = \mathbf{3.136}$

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

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
- i dins els grups: variabilitat 'dins' que no es pot relacionar amb el grup

Model per variable quantitativa vs categòrica: descomposició de la variabilitat

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

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

	SQ (R: Sum Sq)	**Graus llib. GdL** (R: Df)	**QM =SQ/GdL** (R: Mean Sq)	**Rati** (R: **F** value)	**P-valor** (R: Pr(>F))
Explicada pel model (R: *X*) **(ENTRE grups o Between)**	$SQ_E = \Sigma(\hat{y}_i - \bar{Y})^2$ $\sum_{j=1}^{j=k} n_j(\bar{y}_j - \bar{Y})^2$	k–1	$QM_E = SQ_E / (k-1)$	$\hat{F} = \frac{QM_E}{QM_R}$	1-pf(F_value, k-1,N-k)
Residual (R:Residual) **(INTRA grups o Within)**	$SQ_R = \sum_{j=1}^{j=k}\sum_{i=1}^{i=n_j}(y_{ji} - \bar{y}_j)^2$	N–k	$QM_R = SQ_R / (N-k)$		
Total	$SQ_T = \sum_{j=1}^{j=k}\sum_{i=1}^{i=n_j}(y_{ji} - \bar{Y})^2$	N–1			

En aquest cas amb les dades individuals podem calcular els SQ, i partint de mitjanes i desviacions tenim també que $SQ_R = \sum_{j=1}^{j=k}(n_j - 1)s_j^{\,2}$

I llavors $SQ_T = SQ_E + SQ_R$

Model per variable quantitativa vs categòrica: descomposició de la variabilitat

PH GLOBAL

La hipòtesi de que X no aporta informació sobre Y (igualtat de totes les μ_j) es tradueix en que tota la variabilitat entre les μ_j és deguda a la fluctuació del mostreig

PH: H_0 : Variabilitat(μ_j) = 0

H_1: Variabilitat(μ_j) > 0 **unilateral!**

que es resol amb la ràtio F dels quadrats mitjos de la taula de descomposició de la variabilitat:

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

(Ref: *Eei.Ed.UPC* pg 155)

COEFICIENT DE DETERMINACIÓ: $R^2 = \frac{SQ_E}{SQ_T}$

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

EXEMPLE: Notes en 3 grups

(Ref: *Eei.Ed.UPC* pg 154)

$SQ_E = \sum_{j=1}^{j=k} n_j (\bar{y}_j - \bar{Y})^2 =$ **6.60** $\quad SQ_R = \sum_{j=1}^{j=k} (n_j - 1) s_j^{\,2} =$ **257.19** $\quad SQ_T = SQ_E + SQ_R =$ **263.79**

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**		

Objectiu ‘saber’ si el grup afecta a l’esperança de la nota

Hipòtesi H : Variabilitat(μ_j) = 0 (unilateral)

Estadístic $\hat{F} =$ QM_E / QM_R

Distribució sota H : $\hat{F} \rightarrow F_{2,82}$ (les premisses caldrà indicar-les i analitzar-les)

Càlculs $\hat{F} =$ **1.052** P-valor= **0.354** **(1-pf(1.052,2,82))**

Decisió SI és versemblant la hipòtesis que l’esperança de la nota és igual en tots els grups

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

EXEMPLE: Notes en 3 grups. IC de les mitjanes (Ref: *Eei.Ed.UPC* pg 157)

Sabent que: $\bar{y}_1 = 6.15, \bar{y}_2 = 5.73, \bar{y}_3 = 5.48, S_1 = 1.8, S_2 = 1.5, S_3 = 2.0$

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

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**		

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

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

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

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

EXERCICI:

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

nodes catg lgT

x_i	y_i
250	2.31
250	4.48
250	2.59
250	3.06
250	2.1
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

$\bar{y}_1 = 3.082$

$s_1^2 = 0.90$

$\bar{y}_2 = 4.91$

$s_2^2 = 1.79$

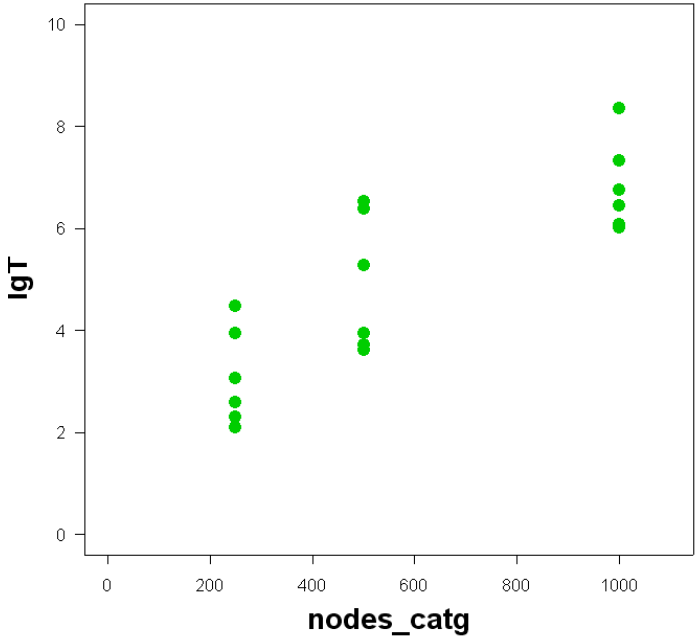
$\bar{y}_3 = 6.83$

$s_3^2 = 0.79$

$\bar{Y} = 4.94$

$S^2 = 1.16$

$S = 1.08$




```
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
```

EXERCICI: "Temps i nombre de nodes de graf en Dijkstra"

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

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**		

```
R: > anova(aov(lgT~as.factor(nodes_catg)))

Analysis of Variance Table
Response:  lgT
                Df Sum Sq Mean Sq F value    Pr(>F)
as.factor(
   nodes_catg)  2 42.122  21.061  18.182 9.789e-05
Residuals      15 17.375   1.158
```

Objectiu 'saber' si el grup afecta a l'esperança del la nota

Hipòtesi H : Variabilitat(μ_j) = 0 (unilateral)

Estadístic $\hat{F} = QM_E / QM_R$

Distribució sota H : $\hat{F} \rightarrow F_{2,15}$ (les premisses caldrà indicar-les i analitzar-les)

Càlculs $\hat{F} =$ **18.15** P-valor= **0.000098** **(1-pf(18.15,2,15))**

Decisió No és versemblant la hipòtesis que el nombre nodes no aporta info sobre el temps

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