4. Estudi de mostres multivariants: el cas bivariant

Part II: Estadística Descriptiva

Professor: Pablo Gregori
gregori@mat.uji.es
Dept. de Matemàtiques

Universitat Jaume I de Castelló

Estadística d'ETIG. Curs 2008-09

Guió

1. Què són i com es representen
   - Taula de freqüències
   - Representació gràfica
2. Estadístics de posició i dispersió
3. Independència estadística entre variables
4. Anàlisi de regressió: cas lineal
   - Càlcul de la funció
   - Bondat d'ajustament
   - Prediccions

Normalment les mostres són multivariants perque interessen molts aspectes, i l'Estadística pot ajudar a trobar relacions sospitoses entre ells.
Exemple de 6 variables:

id	nivell	actitud	estudi	calc	ordin	oci
01	7	3	3	2	2	esports
02	1	0	1	1	2	internet
⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮

En aquest curs només estudiarem les bivariants (2 variables).

Una mostra bivariant és una llista de dades dobles.
Si anomenem X i Y a les variables, la mostra es pot escriure com:

ID	1	2	3	$\cdots$	n
X	x_1	x_2	x_3	$\cdots$	x_n
Y	y_1	y_2	y_3	$\cdots$	y_n

Cada dada és una parella (x_i, y_i). Aquesta mostra té n dades.

Representació de mostres bivariants:

- Numèricament: Es fa recompte de les repeticions, si hi ha, de les dades trobades de cada tipus **(Taula de freqüències)**.
- Gràficament:
  - **Diagrama de barres** (si les dues variables són qualit. o quant. discretes),
  - **diagrames de punts** o **boxplots comparatius**, (si una és qualit. i l'altra quantit.) i
  - **diagrama de dispersió** o **núvol de punts**, si les dues variables són quantit. contínues).

Amb les categories de les dues variables es fa la taula

$X/$ Y	y_1	y_2	$\cdots$	y_l	Fr.abs.(rel.)
x_1	$n_{11}(f_{11})$	$n_{12}(f_{12})$	$\cdots$	$n_{1l}(f_{1l})$	$n_{1.}(f_{1.})$
x_2	$n_{21}(f_{12})$	$n_{22}(f_{22})$	$\cdots$	$n_{2l}(f_{21})$	$n_{2.}(f_{2.})$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_k	$n_{k1}(f_{k1})$	$n_{k2}(f_{k2})$	$\cdots$	$n_{kl}(f_{kl})$	$n_{k.}(f_{k.})$
Fr.abs.(rel.)	$n_{.1}(f_{.1})$	$n_{.2}(f_{.2})$	$\cdots$	$n_{.l}(f_{.l})$	$n(1.000)$

- Quants individus de la mostra tenen la dada (x_i, y_j)? ... $\boxed{n_{ij}}$
- Quin % d'individus de la mostra té la dada (x_i, y_j)? $\boxed{f_{ij}} = \frac{n_{ij}}{n}$

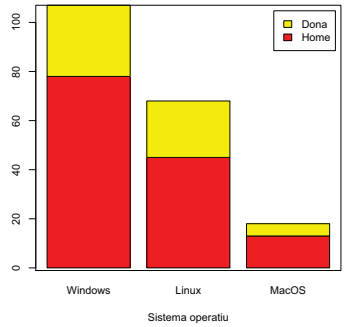
Exemple

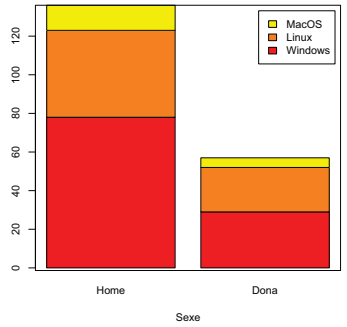
Una enquesta sobre SEXE i SISTEMA OPERATIU resulta:

SEXE \ SO	*Linux*	*MacOS*	*Windows*	*Fr.abs.(rel.)*
Dona	42(0.210)	18(0.090)	37(0.185)	97(0.485)
Home	47(0.235)	23(0.115)	33(0.165)	103(0.515)
Fr.abs.(rel.)	89(0.445)	41(0.205)	70(0.350)	200(1.000)

- *Quin % de persones de la mostra és dona i usa Linux?*
- *Quin % de persones de la mostra usa Linux? I Windows?*
- *Quin % de persones de la mostra és dona? I home?*
- *Quin % de persones de la mostra és dona o usa Linux?*
- *Quin % de persones que usen Linux és dona?*
- *Quin % de les dones usa Linux? I dels homes?*

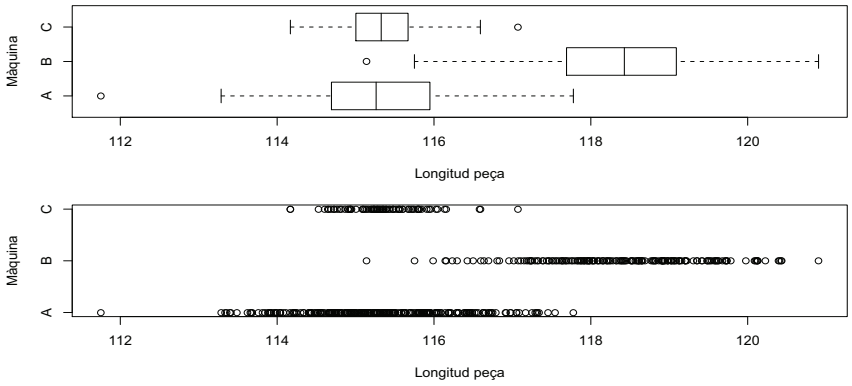
QUALITATIVA vs. QUALITATIVA





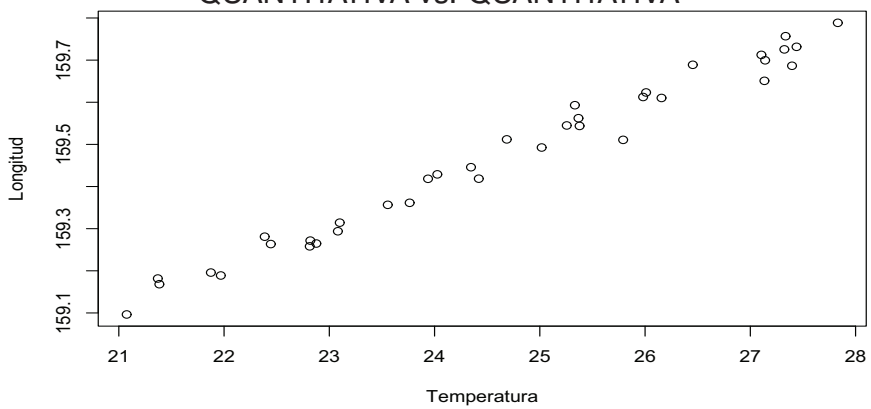
Et sembla que el valor de la variable SEXE afecta la distribució de la variable SO?

QUALITATIVA vs. QUANTITATIVA



Et sembla que la màquina escollida afecta la distribució de les longituds de les peces?

QUANTITATIVA vs. QUANTITATIVA



Et sembla que el valor de la temperatura afecta la longitud de la peça observada?

Quan les dues variables (X, Y) són quantitatives podem resumir la mostra en una posicio central i un grau de dispersió:

- Per la posició, usem la mitjana, ara de dues variables:

$$\overline{(X,Y)} = \frac{\sum_i (x_i, y_i)}{n} = \left(\frac{\sum_i x_i}{n}, \frac{\sum_i y_i}{n}\right) = (\overline{x}, \overline{y})$$

- Per la dispersió, si usem la variància tenim i una dispersió

$$\frac{\sum_i \left((x_i, y_i) - \overline{(x,y)}\right)^2}{n-1}$$

Treballant amb els vectors, el cuadrat es transforma en el producte per la transposta, amb el qual es té com resultat

$$\begin{pmatrix} \frac{\sum_i (x_i-\overline{x})^2}{n-1} & \frac{\sum_i (x_i-\overline{x})(y_i-\overline{y})}{n-1} \\ \frac{\sum_i (y_i-\overline{y})(x_i-\overline{x})}{n-1} & \frac{\sum_i (y_i-\overline{y})^2}{n-1} \end{pmatrix} = \begin{pmatrix} s_X^2 & s_{XY} \\ s_{YX} & s_Y^2 \end{pmatrix}$$

Definició (Covariància i Correlació)

Es defineix la covariància mostral entre les variables X i Y com

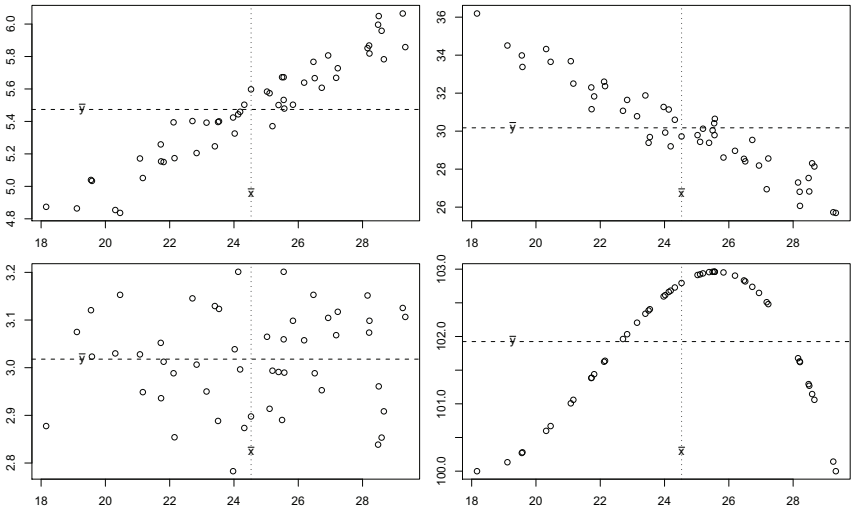
$$s_{XY} = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{n-1} = \frac{\sum_i x_i \cdot y_i}{n-1} - \frac{n}{n-1}\overline{x} \cdot \overline{y}$$

Per altra banda es defineix el coeficient de correlació entre les variables X i Y com

$$r_{XY} = \frac{s_{XY}}{s_X s_Y}$$

La covariància canvia si canviem les unitats de les variables. El coeficient de correlació no. A més $r_{XY} \in [-1.0, 1.0]$ sempre!

El signe de la covariància es pot deduir del núvol de punts



El gràfic és una eina potent per apreciar intuitivament la "possible relació" entre dues variables.

Definició (Independència estadística entre variables)

Quan la distribució d'una variable Y no canvia al condicionar-la als distints valors de l'altra variable X, es diu que Y és ***independent*** *de X.*

En el cas que Y siga independent de X, es pot provar matemàticament que també X és independent de Y, i aleshores es diu directament que X i Y són independents.

Exercici

Una mostra bivariant (variables "sexe" i "sistema operatiu ususal de treball") vé representada en la taula de freqüències següent:

	SISOPER		
SEXE	Windows	Linux	MacOS
Home	78	45	13
Dona	29	23	5

Escriu les distribucions de la variable SEXE condicionades als valors de l'altra i avalua el grau de independència entre les variables.

Quan no hi ha independència estadística, una de les variables sols venir afectada pel valor de l'altra. Segons el context:

- La variable que **rep la influència** es diu variable **explicada**, **resposta**, **dependent**,...
- La variable que **exerceix la influència** es diu variable **explicativa**, **de control**, **independent**,...

Exemples:

- La velocitat de conexió depén del nombre d'usuaris conectats.
- El temps de caiguda lliure d'un cos depén de l'alçada.
- La longitud depén de la temperatura (dilatació).
- La salut d'una persona depén (entre altres) del consum de tabac (OMS).
- La nota en el examen dependrà del temps emprat en la preparació de l'assignatura.

Quan s'intuieix certa dependència entre dues variables, interessa saber si eixa dependència és casualitat o obeix a alguna relació entre elles (causalitat: causa-efecte).
Encara que la regressió es pot fer sempre, per que siga un procediment científic es necesari que:

1. Vinga motivada a priori per sospites fondamentades en la teoría on s'enmarca l'estudi.
2. El resultat es contraste amb les dades mitjançant un gràfic o un coeficient de "bondat d'ajustament".
3. Les conclusions que es traguen no se n'isquen dels màrgens de les dades de la mostra.

Si Y és la variable resposta i X és la de control, trobar la relació entre les variables consisteix en trobar una equació $Y^* = f(X)$ que siga el més consistent possible amb les dades de la mostra.

X	x_1	x_2	x_3	$\cdots$	x_n
Y	y_1	y_2	y_3	$\cdots$	y_n
Y^*	$y_1^* = f(x_1)$	$y_2^* = f(x_2)$	$y_3^* = f(x_3)$	$\cdots$	$y_n^* = f(x_n)$
E	$e_1 = y_1 - y_1^*$	$e_2 = y_2 - y_2^*$	$e_3 = y_3 - y_3^*$	$\cdots$	$e_n = y_n - y_n^*$

La "millor" funció f es tria amb l'objectiu de:

1. Que siga sencilla.
2. Que tinga mínima discrepància amb les dades, es a dir, minimitzant, per exemple, s_E^2.

En molts casos el núvol de punts informa a l'investigador del tipus de funció que convé per fer la regressió, encara que per conéixer-la siga necessari fer càlculs. Per exemple:

1. **Tipus lineal**: $Y^* = a + bX$.
2. Tipus exponencial: $Y^* = ae^{bX}$.
3. Tipus potencial ($Y^* = aX^b$).
4. Tipus parabòlic ($Y^* = a + bX + cX^2$).
5. ...

on a i b són números per determinar amb la condició de mínim error total.
Usant derivades s'obté que **la recta** (tipus lineal) **que millor s'aproxima a un núvol de punts** (x_i, y_i) **és** aquella que:

$$b = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{n\sum x_i^2 - (\sum_i x_i)^2} = \frac{\mathrm{s}_{XY}}{\mathrm{s}_X \mathrm{s}_Y} \qquad a = \overline{y} - b\overline{x}$$

$$\underbrace{\sum_i (y_i - \overline{y})^2}_{\mathrm{SQ}_y} = \underbrace{\sum_i (y_i - y_i^*)^2}_{\mathrm{SQ}_{\mathrm{error}}} + \underbrace{\sum_i (y_i^* - \overline{y})^2}_{\mathrm{SQ}_{\mathrm{regres}}}$$

Definició (Bondat d'ajustament)

Es defineix el ***coeficient de determinació lineal*** *com*

$$R^2 = 1 - \frac{\mathrm{SQ}_{\mathrm{error}}}{\mathrm{SQ}_y}$$

En el cas lineal es té que $R^2 = r_{XY}^2$.

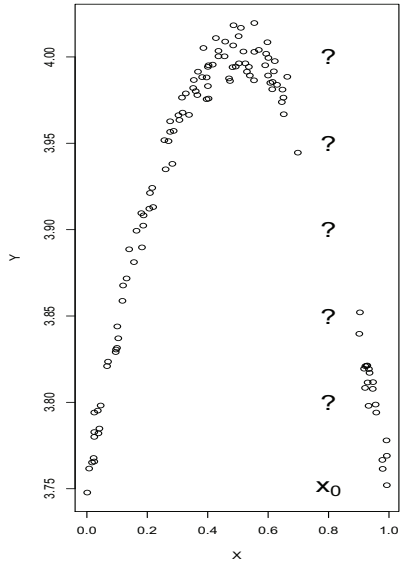
Propietat

$R^2 \in [0.0 - 1.0]$ *sempre!*

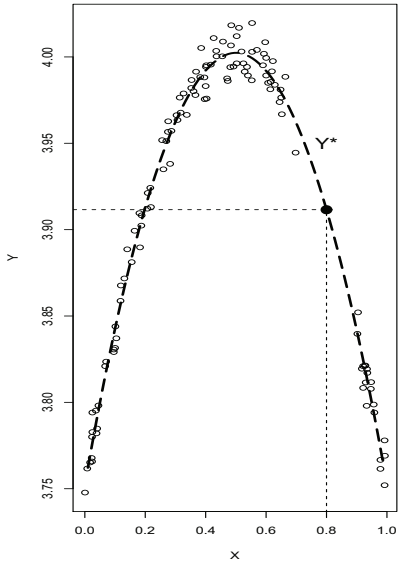
La qualitat de l'ajustament ve reflectida al R^2:

- A major R^2, major qualitat.
- La perfecció és $R^2 = 1$.

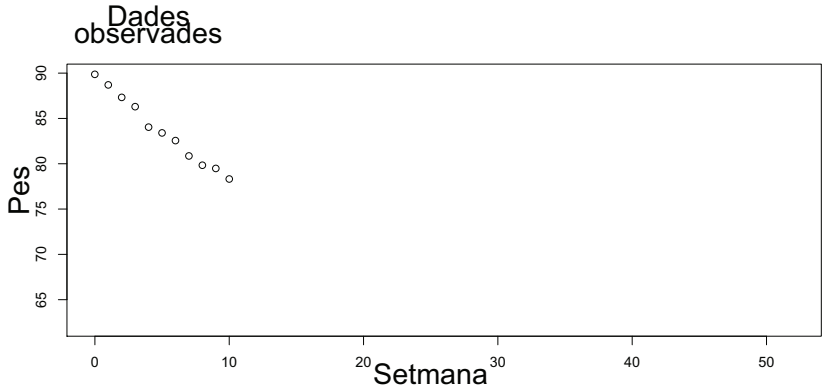
- Una anàlisi de regressió que acaba amb un R^2 pròxim a 1.0 indica que les dades de Y venen ben ajustades per la funció $Y^* = f(X)$.
- Aleshores, per a un nou valor de $X = x_0$ podem fer una **predicció** de valor de Y (que seria $y_0^* = f(x_0)$).
- A major R^2, major confiança en que el valor de la predicció seria pròxim al valor real.



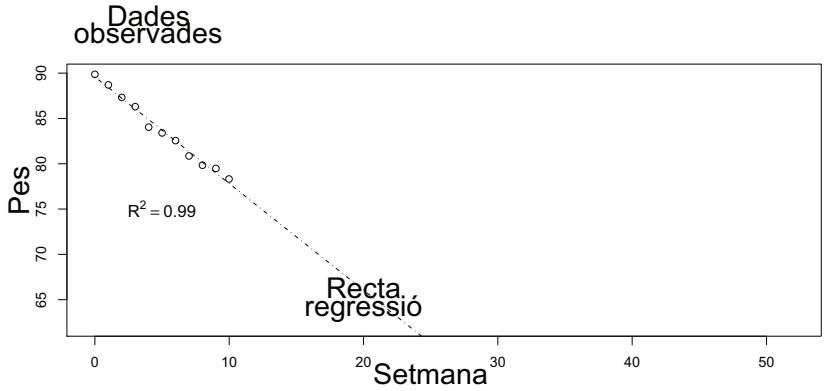
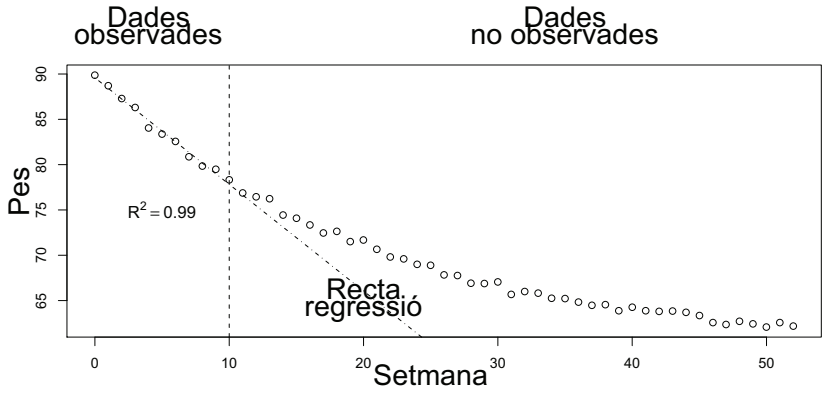
- Un anàlisi de regressió que acaba amb un R^2 pròxim a 1.0 indica que les dades de Y venen ben ajustades per la funció $Y^* = f(X)$.
- Aleshores, per a un nou valor de $X = x_0$ podem fer una **predicció** de valor de Y (que seria $Y^* = f(x_0)$).
- A major R^2, major confiança en que el valor de la predicció seria pròxim al valor real.



Atenció! Les prediccions fetes mitjançant la regressió seran fiables només dins del rang de valors on s'ha treballat.



Atenció! Les prediccions fetes mitjançant la regressió seran fiables només dins del rang de valors on s'ha treballat.



Atenció! Les prediccions fetes mitjançant la regressió seran fiables només dins del rang de valors on s'ha treballat.