Relació entre dues variables numèriques

Continuant amb l'anàlisi de relació entre variables X i Y, avui considerarem el cas en que les dues variables són numèriques.

Recordeu que ens podem trobar en la situació de mostres independents (aleshores utilitzàvem les eines de l'anàlisi estadística d'una variable per estudiar la relació) o mostres aparellades. En aquest últim cas tenim més eines al nostre abast.

El nostre objectiu és l'estudi conjunt entre de dues variables numèriques per tal d'avaluar-ne la seva eventual relació i, si s'escau, determinar el tipus de relació existent. Buscarem relacions estadístiques, és a dir, no relacions funcionals exactes, sinó relacions d'associació. Per exemple, relacions en sentit directe (X gran aleshores Y gran o X petita aleshores Y petita) o invers (X gran aleshores Y petita o X petita aleshores Y gran). Ara bé, cal tenir en compte que les relacions que obervarem no seran sempre de causa-efecte.

Eines i pasos a realitzar en l'estudi conjunt de dues variables numèriques Si volem estudiar la relació entre les dues variables X i Y utilitzarem les següents eines:

1. **Diagrama de dispersió**: ens permet analitzar visualment la distribució conjunta de dues variables.

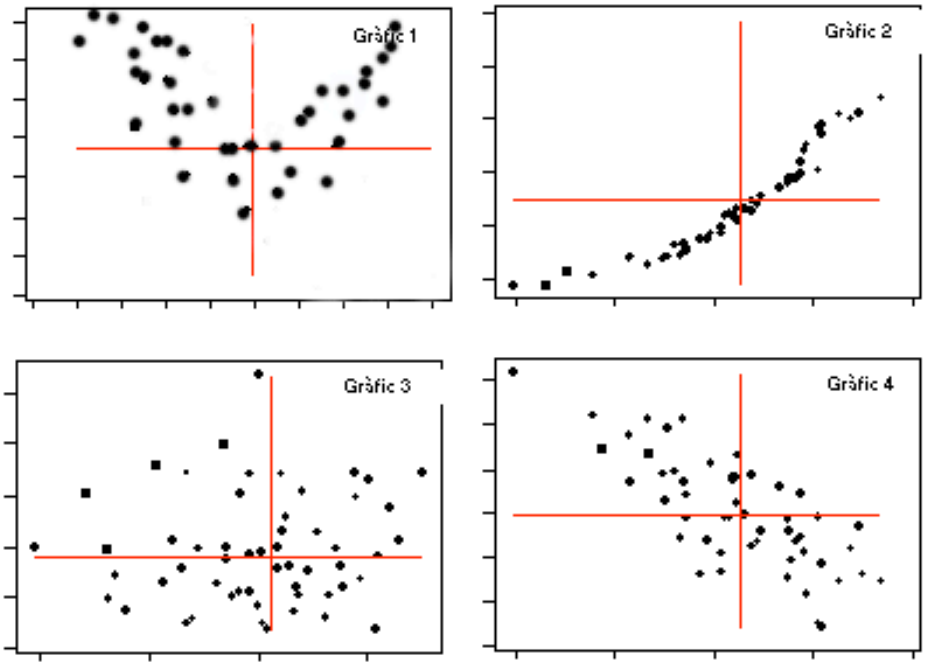
2. **Coeficient de correlació de Pearson**: és un valor numèric que ens informa del grau o intensitat de la relació lineal entre les variables ($Y \simeq aX + b$) i de la tendència de la relació.

3. **Models de regressió:** és una temptativa d'ajustar una funció a les dades (gràficament, al diagrama de dispersió). Ens concentrarem en el cas de la recta de regressió (ajust lineal).

Diagrama de dispersió: ens uns eixos X i Y representem els punts (x_i, y_i) per cada individu de la mostra on x_i i y_i són els valors que prenen les variables X i Y a l'individu i-èssim.

És una eina molt útil que ens permet analitzar visualment la existència o no de relació entre les dues variables i el tipus de relació.

Figura 1 Els eixos de coordenades tenen origen el punt $(\overline{X}, \overline{Y})$



1. **Relació directa o positiva:** Quan valors grans de la variable X van emparellats amb valors grans de Y i els valors petits d'una s'associen amb valors petits de l'altra. Per a visualitzar-ho convé dibuixar uns eixos en el punt mitjà de la distribució i veure si la majoria de punts cauen en el primer i el tercer quadrants, **Figura 1-gràfic 2**.

   En aquest cas podrà ser apropiat un model de regressió lineal amb una recta de regressió creixent.

2. **Relació inversa o negativa:** Quan valors grans de la variable X van emparellats amb valors petits de Y i els valors petits de X s'associen amb valors grans de Y. Per a visualitzar-ho convé dibuixar uns eixos en el punt mitjà de la distribució i veure si la majoria de punts cauen en el segon i el quart quadrants, **Figura 1-gràfic 4**.

   En aquest cas podrà ser apropiat un model de regressió lineal amb una recta de regressió decreixent.

3. **Relació nula:** Quan els valors presenten una pauta clarament no lineal o no presenten cap pauta, de manera que valors grans de X s'alternen amb valors grans i petits de Y, i vice-versa. Al dibuixar un eixos pel punt mitjà, els punts cauen en tots els quadrants, **Figura 1-gràfic 3**.

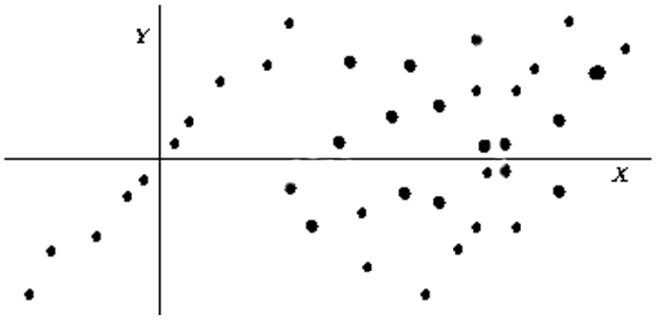
   En aquest cas no serà apropiat cap model de regressió.

4. **Relació de tendència no-lineal**: Com per exemple la tendència parabólica clara de la **Figura 1-gràfic 1**.

   En aquest cas podria ser apropiat un model de regressió no-lineal, potser un model quadràtic (paràbola).

(B.) Podem veure si el grup d'individus es pot considerar homogeni o bé hi ha subgrups diferenciats que caldria tractar per separat, **Figura 2**:

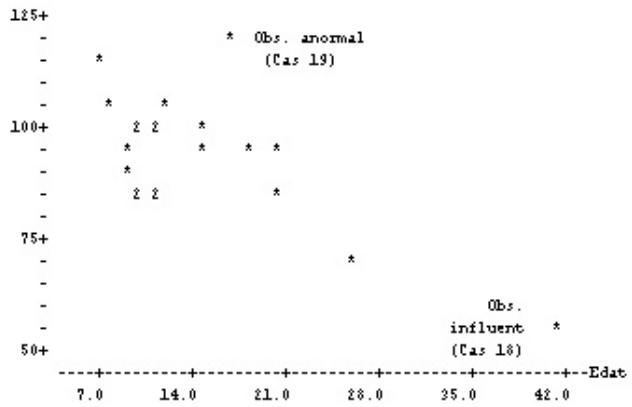
Figura 2



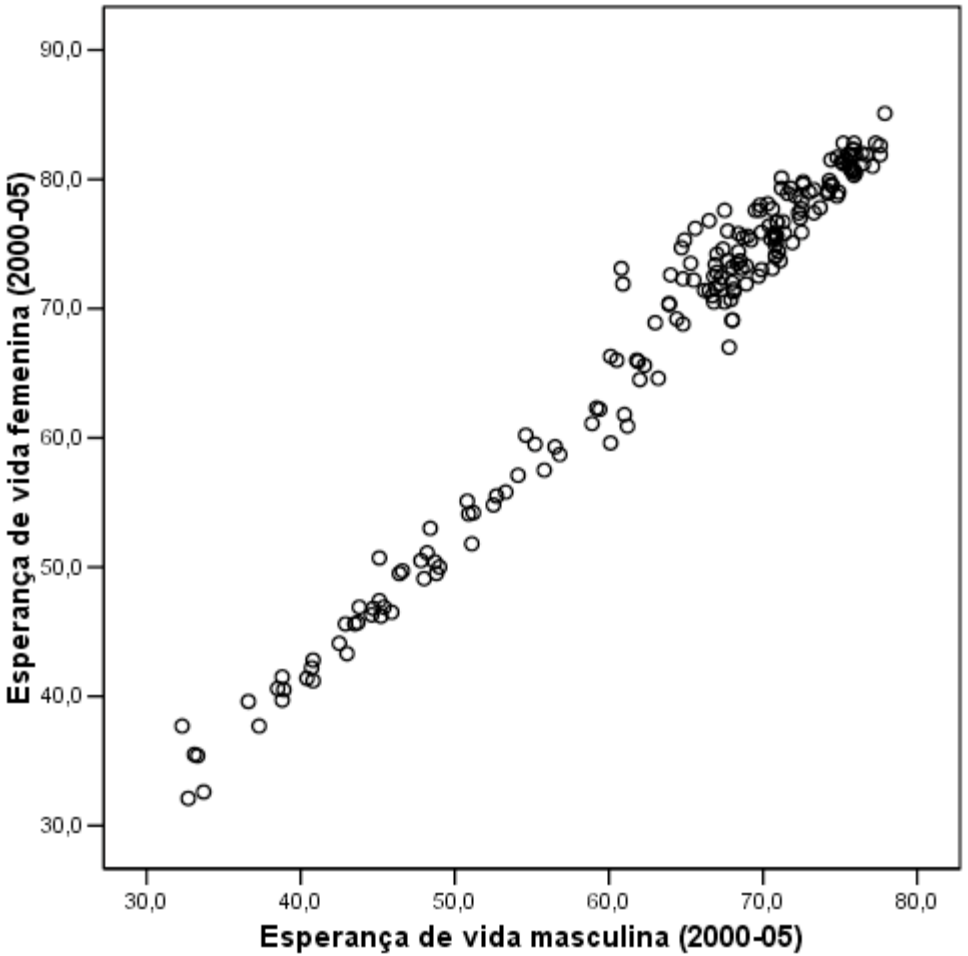
En la figura és veu un primer subgrup a l'esquerra amb una forta relació directa i lineal, i un segon subgrup a la dreta amb una relació més feble.

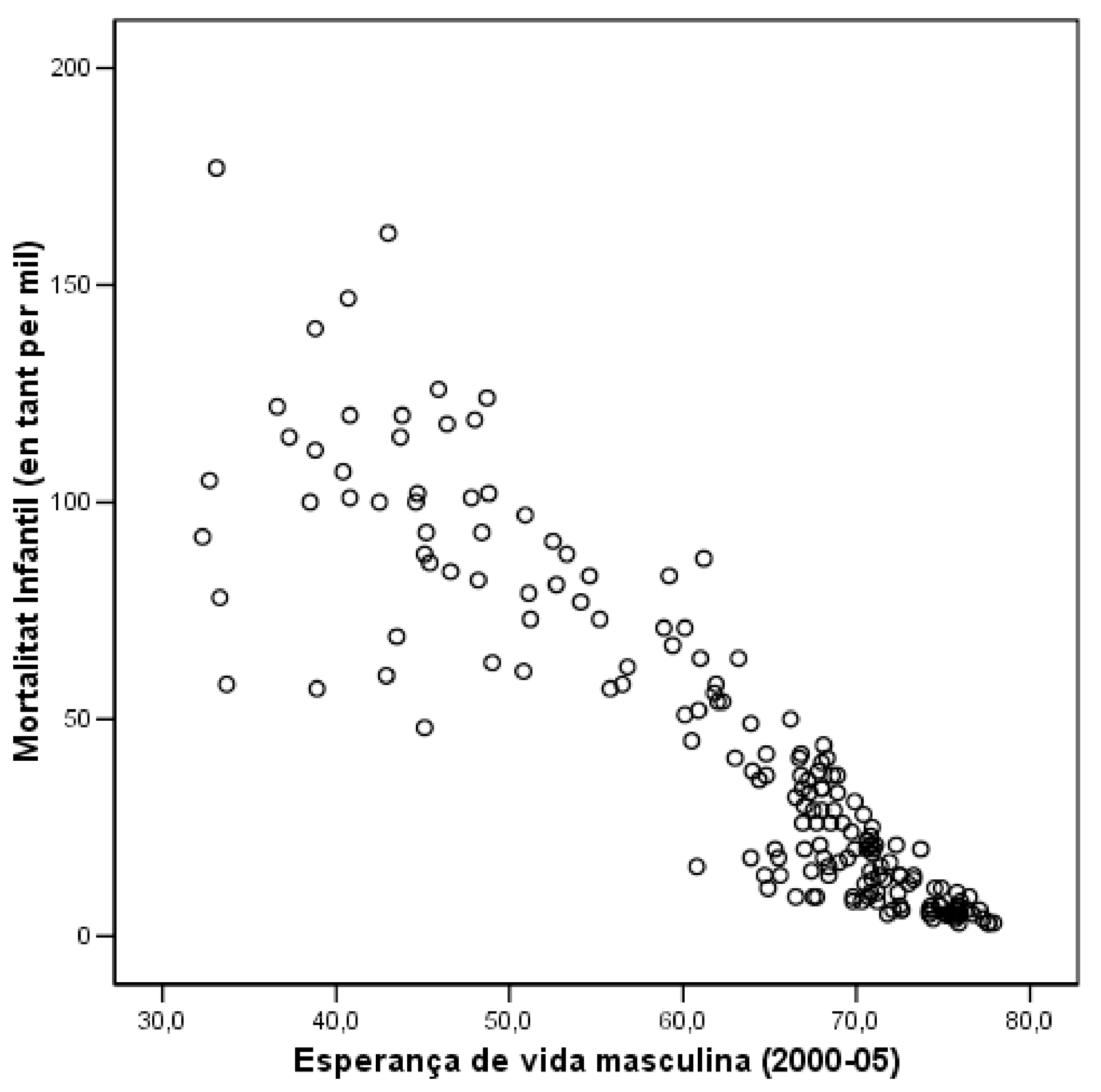
També es pot detectar l'existència d'observacions *influents* o *anòmales*, **Figura 3**:

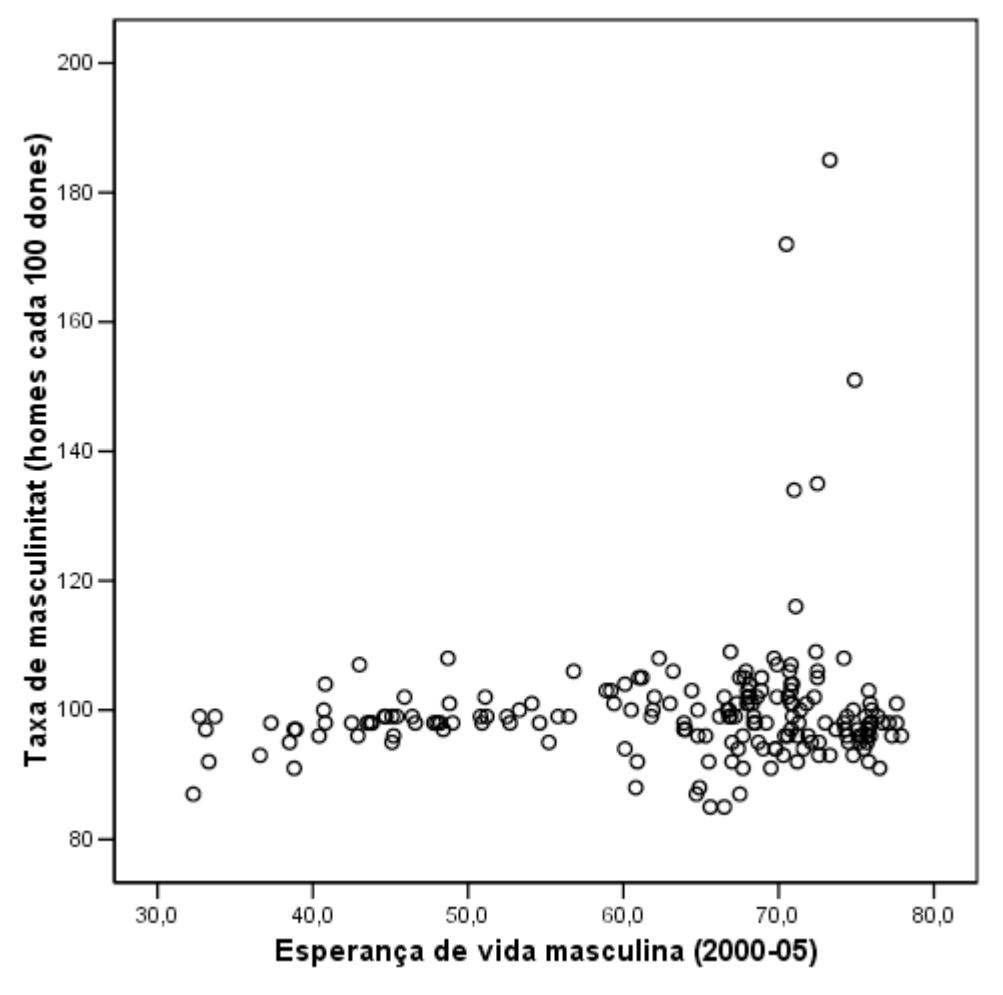
Figura 3

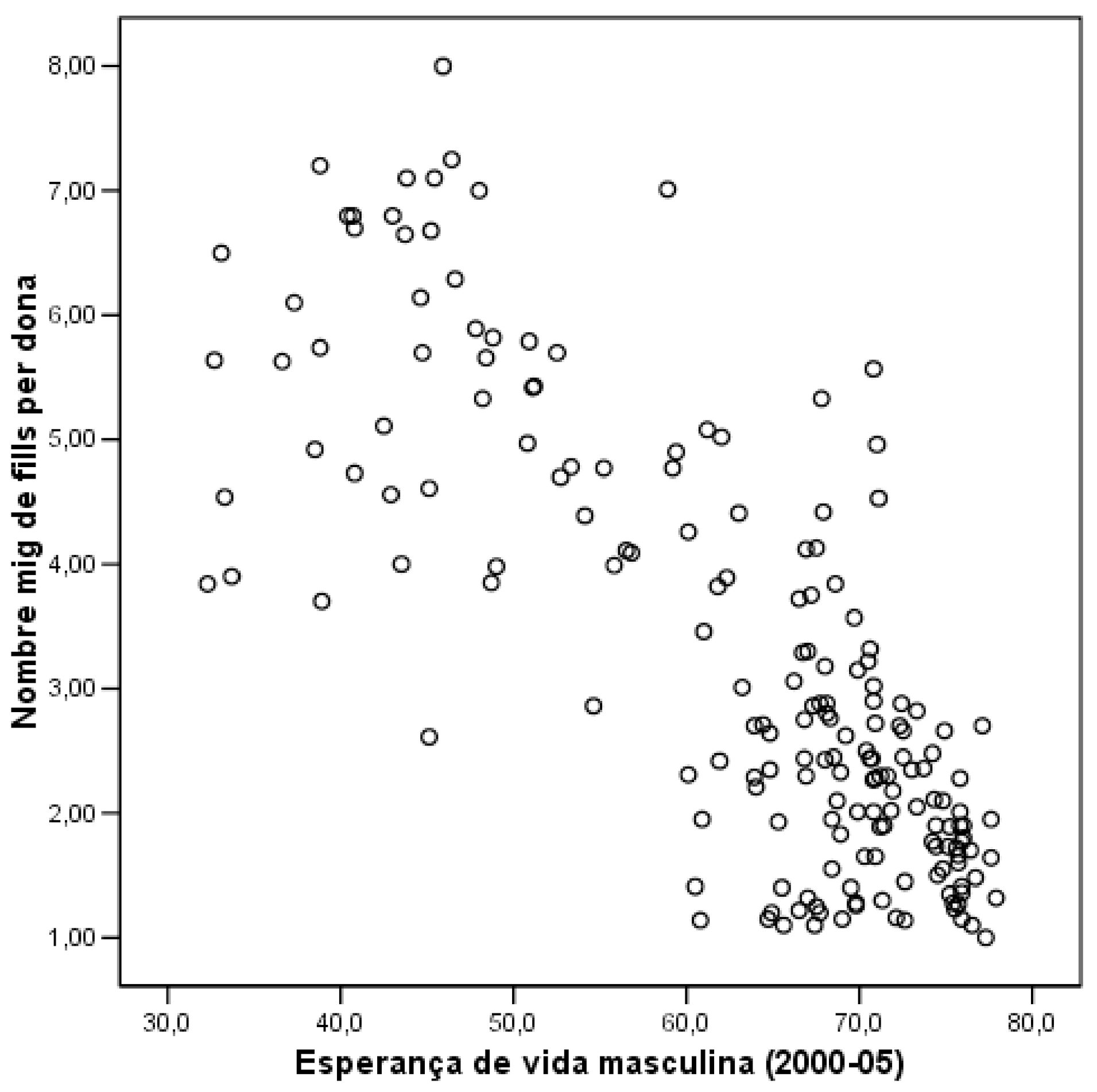


La diferència entre valor *anòmal* i valor *influent* és de vegades subtil. Un valor és influent quan "desvia" la recta de regressió. Les observacions anòmales han d'examinar-se perquè es podrien deure a errades.


90,0
80,0
70,0
60,0
50,0
40,0
30,0
Esperança de vida femenina (2000-05)
30,0
40,0
50,0
60,0
70,0
80,0
Esperança de vida masculina (2000-05)


Mortalitat Infantil (en tant per mil)
200
150
100
50
0
30,0
40,0
50,0
60,0
70,0
80,0
Esperança de vida masculina (2000-05)


Taxa de masculinitat (homes cada 100 dones)
200
180
160
140
120
100
80
30,0
40,0
50,0
60,0
70,0
80,0
Esperança de vida masculina (2000-05)


8,00
7,00
6,00
5,00
4,00
3,00
2,00
1,00
Nombre mig de fills per dona
30,0
40,0
50,0
60,0
70,0
80,0
Esperança de vida masculina (2000-05)

2 El coeficient de correlació lineal de Pearson

En l'apartat anterior hem vist que el diagrama de dispersió és una eina per visualitzar si els valors de dues variables presenten algun tipus d'associació o relació.

El següent pas és obtenir una avaluació numèrica del grau de relació entre les variables.

Els coeficients de correlació (n'hi ha diversos) són uns estadístics que avaluen numèricament la relació entre les variables. El més apropiat per a variables numèriques és el coeficient de correlació de Pe
Per a dades ordinals s'usa el coeficient de correlació de Spearman.

Tots els coeficients de correlació (en particular el de Pearson) tenen en comú:

- Són estàndards (sense unitats) i poden prendre qualsevol valor de l'interval $[-1, +1]$.
- El *signe* del coeficient indica el *sentit* de la relació:
  (a) Sí és positiu, relació directa o creixent (**Fig 1-Gràfic 2**).
  (b) Sí és negatiu, relació inversa o decreixent (**Fig 1-Gràfic 4**).
- El *valor absolut* de coeficient indica el *grau o intensitat* de la relació:
  (a) Valor absolut = 1, relació màxima (punts perfectament alineats).
  (b) Valor absolut = 0, relació lineal nul.la; en la **Fig 1-Gràfics 1 i 3** el coeficient de correlació seria aproximadament 0.
  (c) Com més gran és el valor absolut, més intensa és la relació; la **Fig 1-Gràfic 2** indica una relació més intensa que la **Fig 1-Gràfic 4**.

Seguidament estudiarem el coeficient de correlació de Pearson:

Definició: Donades dues variables numèriques X i Y, definides en els mateixos individus-objectes, i que prenen conjuntament n parelles de valors $(x_1, y_1), \ldots, (x_i, y_i), \ldots, (x_n, y_n)$, es defineix el **coeficient de correlació de Pearson de X i Y** (que denotarem $r_{X,Y}$ o bé $Corr(X, Y)$, o simplement r) com l'estadístic donat per l'expressió següent:

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{X})^2 \sum_{i=1}^{n}(y_i - \overline{Y})^2}} \tag{1}$$

Els criteris que donem a continuació per avaluar la intensitat de la relació lineal a partir del coeficient de correlació s'han d'entendre només com una pauta. El valor del coeficient també depèn del nombre d'observacions que estem considerant, si tenim moltes observacions un valor alt és molt difícil d'assolir.

Criteris de qualificació de la relació, segons el valor de r:

$r = 0$	Relació lineal **nul.la**
$\lvert r\rvert = 1$	Relació lineal **perfecta**, $Y = bX + a$: $\left\{ \begin{array}{ll} \text{creixent} & \text{si } r = 1 \\ \text{decreixent} & \text{si } r = -1 \end{array} \right\}$
$0 < \lvert r\rvert \leq 0.3$	Relació lineal **molt feble**: $\left\{ \begin{array}{ll} \text{creixent} & \text{si } r < 0 \\ \text{decreixent} & \text{si } r > 0 \end{array} \right\}$
$0.3 < \lvert r\rvert \leq 0.7$	Relació lineal **feble**: $\left\{ \begin{array}{ll} \text{creixent} & \text{si } r < 0 \\ \text{decreixent} & \text{si } r > 0 \end{array} \right\}$
$0.7 < \lvert r\rvert \leq 0.87$	Relació lineal **moderada**: $\left\{ \begin{array}{ll} \text{creixent} & \text{si } r < 0 \\ \text{decreixent} & \text{si } r > 0 \end{array} \right\}$
$0.87 < \lvert r\rvert \leq 1$	Relació lineal **forta**: $\left\{ \begin{array}{ll} \text{creixent} & \text{si } r < 0 \\ \text{decreixent} & \text{si } r > 0 \end{array} \right\}$

- El coeficient de correlació, a l'igual que la mitjana i també la desviació típica, es pot veure molt afectat per valors anòmals o influents, aquests valors es poden detectar gràficament (diagrames de caixa i de dispersió), i convé verificar que no es deuen a errades de les dades. Cas que siguin valors vàlids, es convenient fer un estudi amb els valors 'estranys' i sense ells.
- L'existència de dues sub-poblacions o subgrups diferenciats també pertorba el valor del coeficient de correlació.
- Un valor elevat del coeficient **no** significa relació **causal**.
- El coeficient de correlació de Pearson **només mesura el grau de correlació lineal**. Un valor de r proper a zero no indica que no hi hagi una altra mena de relació, com la parabòlica.

Les gràfiques de la pàgina següent corresponen a exemples amb distints tipus de relació segons els anteriors qualificatius. A la taula de correlacions hi podeu veure el valor del coeficient de correlació de Pearson per a distintes parelles de variables. Els quadres ombrejats corresponen a correlacions significatives des d'un punt de vista inferencial.

Qualifiqueu les diferents correlacions de les parelles de variables segons els criteris establerts en el quadre anterior i vegeu a què corresponen gràficament.

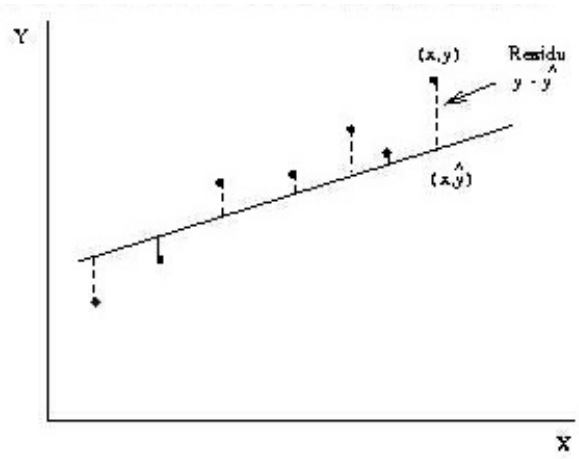
L'SPSS calcula el coeficient de correlació a:

Analizar –> Correlaciones –> Bivariadas

Recta de regressió

Per definició, *és la recta que millor ajusta el núvol de punts en el sentit que minimitza els residus de les prediccions de Y a partir de X.* Qualsevol altra recta produiria residus més grans.

Figura 6



Predicció de valors de Y

No només és interessant obtenir les prediccions i els residus per a valors x_i que pertanyen al conjunt d'observacions, sinó per a valors 'nous' $x*$ de X. Totes Les prediccions es fan segons la recta de regressió obtinguda, tant per a les observacions com per als nous valors.

Per a les observacions x_i de X	$- - - >$	Predicció del valor de Y:	$\hat{y}_i = a + bx_i$
Per als 'nous' valors $x*$ de X	$- - - >$	Predicció del valor de Y:	$\hat{y}* = a + bx*$

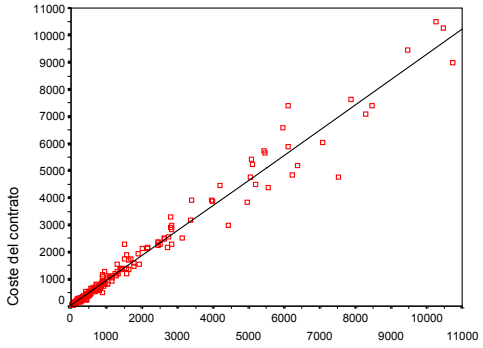
Important: Les prediccions, per a valors de la variable independent molt extrems (allunyats de la mitjana) no són recomanables i poden donar lloc a prediccions absurdes (vegeu l'Exemple 2). El model només és vàlid per un rang restringit d'observacions.

L'SPSS fa el procediment de regressió a:

Analizar –> Regresión lineal
Escollirem:
Coeficientes de regresión −− > Estimaciones
Guardar: Valores pronosticsdos no tipificados
Residuos no tipificados

- El tercer exemple de model lineal és per a les variable independent o predictora $X =$ *Estimació del cost de ...* i la variable dependent o resposta $Y =$ *Cost del contracte*.

  El diagrama de dispersió i els resultats del model de regressió de l'SPSS els teniu a la **Figura**. Veiem que el núvol de punts permet una bon ajust al model lineal sobretot per a valors de X petits, ja que per a valors grans hi ha més dispersió. La relació és directa i forta.



Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,987[a]	,974	,974	313,0885

a. Variables predictoras: (Constante), Estimación del coste de contrucción por el Dpto. Técnico

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados		
		B	Error típ.	Beta	t	Sig.
1	(Constante)	20,907	24,367		,858	,392
	Estimación del coste de contrucción por el Dpto. Técnico	,926	,010	,987	93,886	,000

a. Variable dependiente: Coste del contrato

Veiem que:

$$\textbf{Recta de regressió:} \quad y = .926x + 20.907$$

Les prediccions de valors de Y, fixats certs valors de X, són:

$$\begin{aligned} x* &= 4\,300 --> y* = .926(4\,300) + 20.907 \approx 4\,322 \\ x* &= 10\,500 --> y* = .926(10\,500) + 20.907 \approx 9\,744 \end{aligned}$$

La fiabilitat d'aquesta última estimació és menor perquè es tracta d'un valor més extrem.