Pràctica 8. Anàlisi descriptiva de la relació entre variables numèriques: correlació i regressió

Aquesta pràctica comença amb el diagrama de dispersió (o núvol de punts), que ens dóna informació visual de la relació entre dues variables numèriques. Cal que les variables corresponguin als mateixos casos (o a casos aparellats).

També calcularem el coeficient de correlació, que quantifica la relació lineal. Finalment, veurem la recta de regressió quan s'escaigui (és a dir, quan el coeficient de correlació sigui l'adequat).

▶ Recordeu activar les opcions d'edició:

- "Mostrar comandos en anotaciones" a la pestanya de "Visor".
- "Nombre y etiquetas" per a les variables i "Valores i etiquetas" per als valors l'apartat de "Etiquetado de tablas pivot" de la pestanya de "Etiquetas de resultados".

▶ En aquesta pràctica treballarem amb l'arxiu *poblacio.sav* primerament, i després amb *exemple de regressio vot PSC-inmigracio.sav*. Per començar obrim el fitxer **poblacio.sav**, (amb dades de la pàgina web de les Nacions Unides).

0. Disseny: dades aparellades

La base de dades ha de contenir variables numèriques, on cada cas correspongui a un mateix individu, objecte, entitat (mateix país en aquest arxiu), o a individus aparellats.

L'objectiu és analitzar la relació entre parelles de variables numèriques. Per exemple veurem si hi ha relació entre les variables *espvidam* (esperança de vida masculina) i *espvidaf* (esperança de vida femenina), o bé entre *espvidam* i *mortinf* (mortalitat infantil), etc...

Les eines per a analitzar la relació són: el diagrama de dispersió, el coeficient de correlació i la recta de regressió (si s'escau).

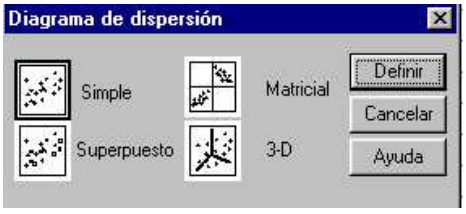
1. Diagrama de dispersió

És l'eina gràfica bàsica. En un diagrama d'eixos cartesians, es representa un núvol de punts: cada punt representa un cas, la projecció del punt sobre cada eix és el valor de la variable corresponent en aquest cas.

Per fer el diagrama, anem a la barra del menú i seleccionem:

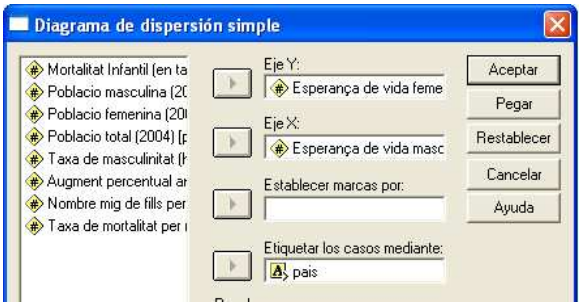
Gráficos −−> Dispersión

Activem l'opció **Simple** i cliquem a **Definir**



A la casella Eje X posem la variable *espvidam* i a Eje Y posem la variable *espvidaf*.

A l'opció Etiquetar los casos mediante posem la variable *pais*, per que surti el nom del país enlloc del número de posició.



Obtenim el següent diagrama:

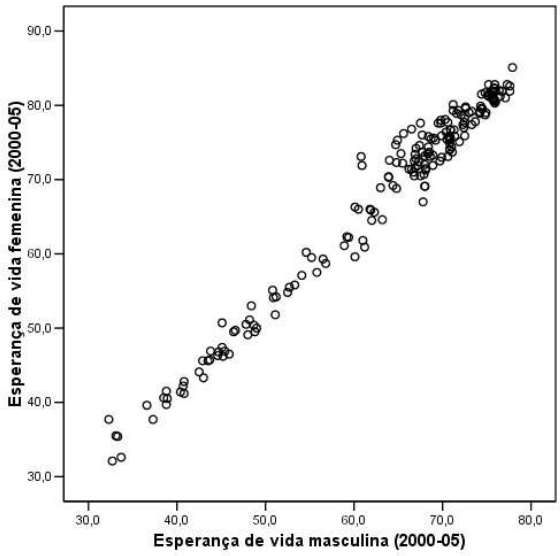
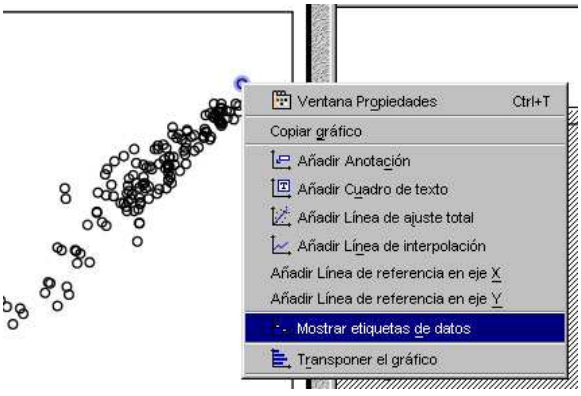


Figura 1

Si volem saber quin és el país amb esperances de vida més altes, editem el gràfic, seleccionem *un sol punt* i amb el menú contextual, triem l'opció **Mostrar etiquetas de datos**.



I otenim l'etiqueta del Japó.

De manera anàloga obtenim els vuit gràfics següents:

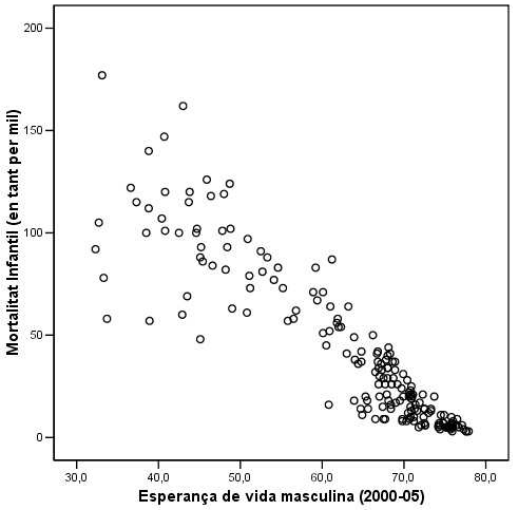


Figura 2

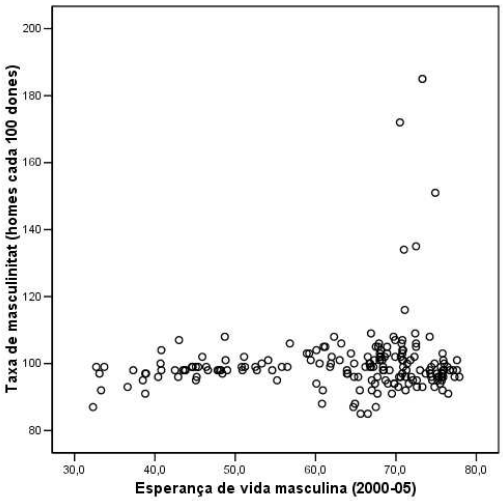


Figura 3

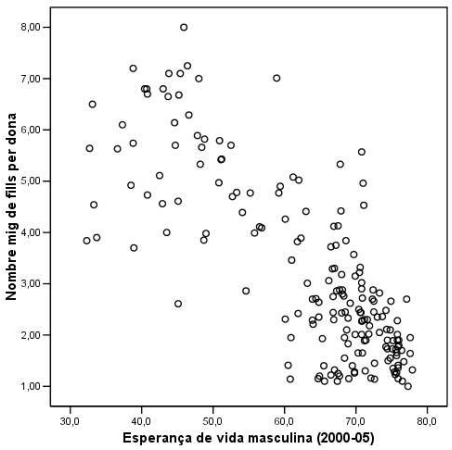


Figura 4

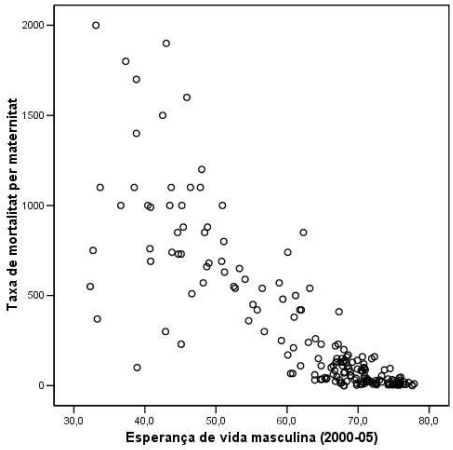


Figura 5

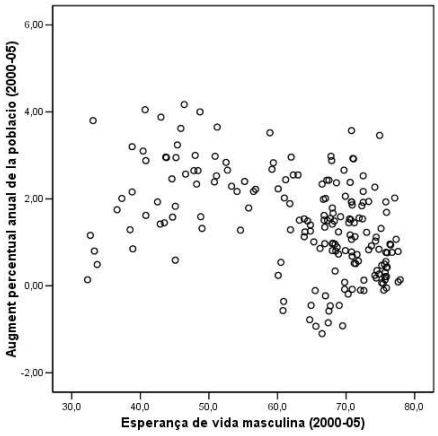


Figura 6

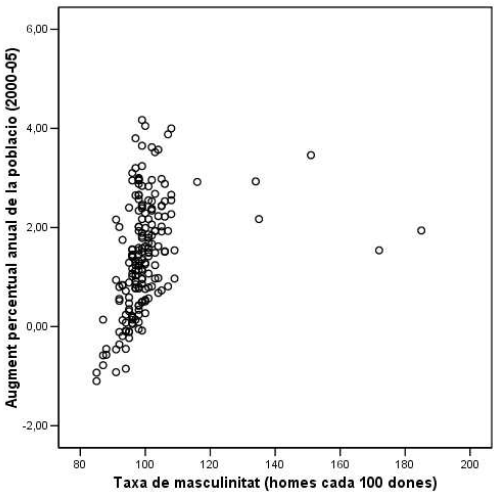


Figura 7

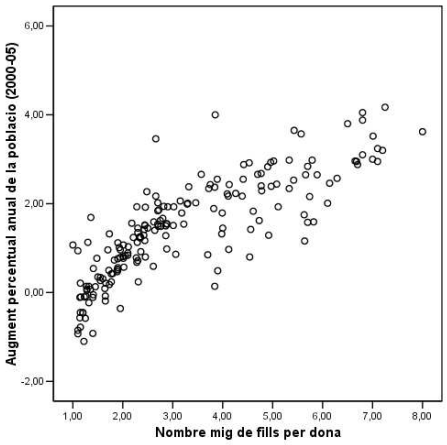


Figura 8

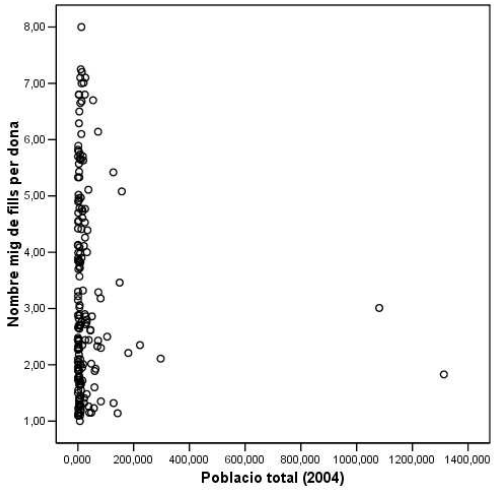
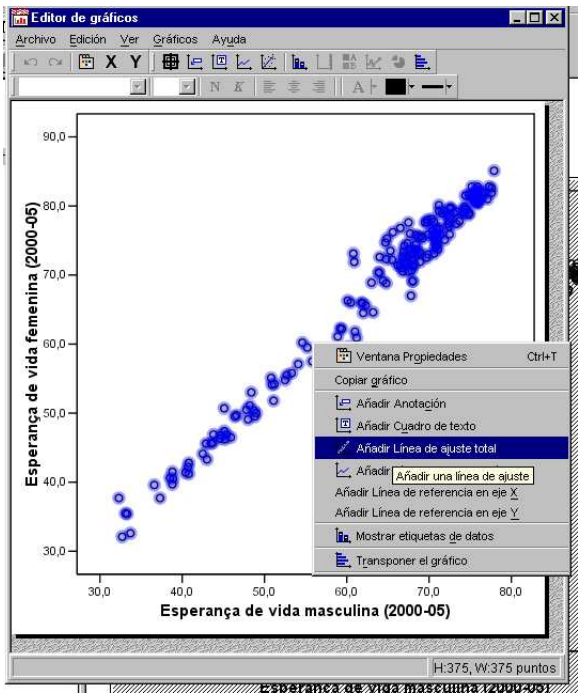
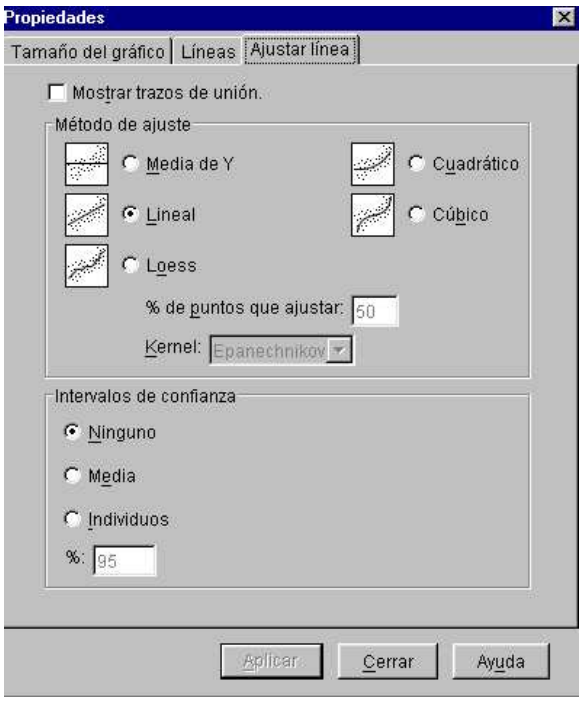


Figura 9

► Aquests gràfics es poden retocar, editant-los de la manera habitual. Per exemple, podeu fer que l'escala vertical de la figures 4 i 9 comencin en el zero, podeu canviar els noms dels eixos, canviar els marcadors dels punts, el seu color, etc.

També podem fer que s'hi ajusti la recta de regressió dins del menú d'edició de la gràfica. Editem la gràfica de la figura 1, seleccionem els punts, i amb el menú contextual seleccionem l'opció **Añadir Línea de ajuste total**.





A la finestra "Propiedades", a la pestanya Ajustar línea escollim l'opció **Lineal**.

Obtenim la figura 1-bis, i observem un bon ajust lineal. En canvi en la figura 8-bis la corba cúbica s'ajusta millor.

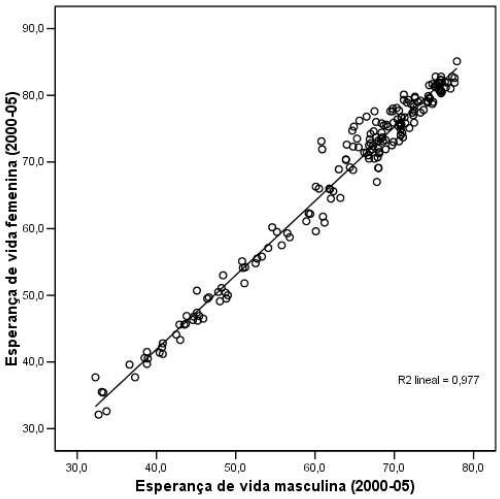


Figura 1-bis

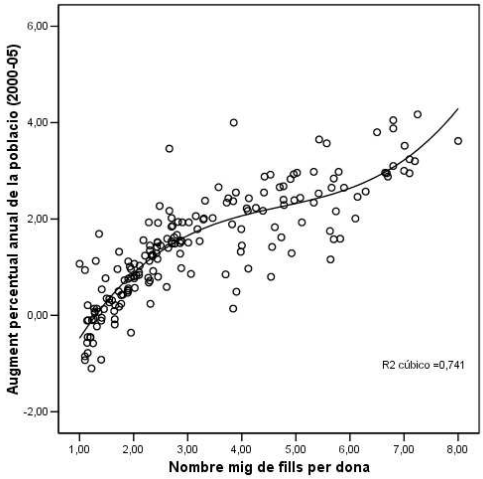


Figura 8-bis

Analitzem les gràfiques:

- La figura 1 indica una tendència de relació creixent i lineal (s'ajusta bé a la recta). Veiem que als països amb *espvidam* elevada els correspon també *espvidaf* elevada, els que la tenen baixa en una també la tenen baixa en l'altra. Per tant ens indica una relació lineal creixent entre les dues variables, i a més sembla que el grau d'intensitat de la relació és elevat, doncs l'ajust és força bo a la recta.
- La figura 2 indica una tendència de relació decreixent i lineal, però menys intensa.
- La figura 8 indica una tendència de relació creixent, però l'ajustament lineal encara sembla més feble.
- La resta de figures presenten una relació no lineal o en tot cas molt feble.
- També detectem la presència de punts influents (outliers). En concret a les figura 3 i 7 destaquen els punts amb una taxa de masculinitat superior al 120%. Els podem localitzar a la base de dades: els Emirats Àrabs Units, Qatar, Kuwait, Oman i el Sahara Occidental (per ordre decreixent).
- En la figura 9 també destaquen els outliers de població.
- A la figura 5, la taxa de mortalitat per maternitat està expressada en nombre de morts de la mare per cada 100.000 naixements vius. Observem que aquesta taxa té valors molt concentrats en els països amb una esperança de vida masculina elevada, però molt dispersos en els països amb una esperança de vida masculina baixa.
- A la figura 6 observem que la població només disminueix en països que tenen esperança de vida masculina de 60 anys o més.

▶ QÜESTIÓ:

- A la figura 4, quin és el país amb un nombre de fills per dona més elevat?
- A la figura 9, quins són els dos països amb més població?
- A la figura 8-bis, identifica els dos països que, amb un augment percentual anual de la població elevat, s'aparten més del grup.

2. Coeficient de correlació

L'apreciació de la intensitat de relació que fem a partir del diagrama de dispersió no és gaire precisa i pot dependre de l'escala en que fem la gràfica. L'eina numérica que avalua la intensitat de relació lineal és el coeficient de correlació de Pearson.

- Se sol denotar **r** i també ρ (rho), i és un indicador de la intensitat que no té unitats i que pren valors entre -1 i 1.
- Si **r** es positiu indica relació directa o creixent.
- Si **r** es negatiu indica relació inversa o decreixent.
- Si el mòdul de **r** (sense signe) és proper a 1, aleshores indica relació lineal intensa, més intensa com és a prop de 1.
- En canvi, si **r** és proper a 0, aleshores indica un manca de relació lineal, la qual cosa *no vol dir manca de relació*.

Calculem el coeficient de correlació de les variables de la figura 1. Activem:

Analizar $-->$ Correlaciones $-->$ Bivariadas

S'obre una finestra, on posem les variables *espvidam* i *espvidaf*.

i obtenim

Correlaciones

		Esperança de vida masculina (2000-05)	Esperança de vida femenina (2000-05)
Esperança de vida masculina (2000-05)	Correlación de Pearson	1	,989**
	Sig. (bilateral)		,000
	N	188	188
Esperança de vida femenina (2000-05)	Correlación de Pearson	,989**	1
	Sig. (bilateral)	,000	
	N	188	188

**. La correlación es significativa al nivel 0,01 (bilateral).

Veiem que $\boxed{r = 0.989}$, que indica una relació lineal intensa i directa o creixent.

▶ Exercici:

Obteniu els coeficients de correlació de les altres 8 parelles de variables de les figures 2 a 9, i compareu-ne el valor amb l'apreciació gràfica.

3. Recta de regressió

Quan el diagrama de dispersió ens indica que l'ajust lineal pot ser bo, i el coeficient de correlació (en valor absolut) és proper a 1 (posem 0.7 com a llindar), podem determinar quina és la recta que s'ajusta al núvol de punts, a fi i efecte de fer prediccions.

L'eqüació d'una recta és $\boxed{y = \mathbf{a} + \mathbf{b}\,x}$. On:

$$\begin{cases} \mathbf{b} \text{ és el pendent de la recta} \\ \mathbf{a} \text{ és l'ordenada a l'origen (valor de } y \text{ quan } x = 0) \text{ o constant} \end{cases}$$

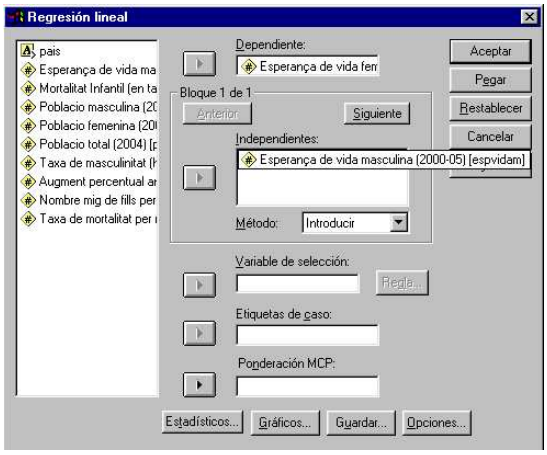
El programa SPSS ens calcula els valors del pendent **b** i la constant **a**.

Fem-ho per a les parelles de variables de les figures 1 i 2:

Per a la figura 1 utilitzarem:

- X =*espvidam* (variable independent) i
- Y =*espvidaf* (variable dependent).

Activem: **Analizar** $-->$ **Regresión** $-->$ **Lineal**
I entrem les variables:



De tots els resultats obtinguts, i sabent prèviament que $r = 0.989$, ens quedem amb el de la taula de coeficients (podeu esborrar la resta de taules):

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	-2,630	,798		-3,297	,001
	Esperança de vida masculina (2000-05)	1,113	,012	,989	89,870	,000

a. Variable dependiente: Esperança de vida femenina (2000-05)

Veiem que $\mathbf{a = -2.630}$ i $\mathbf{b = 1.113}$

Per tant la relació entre les variables s'ajusta a una recta de la forma següent:

$$\boldsymbol{espvidaf = -2.630 + 1.113\ (espvidam)}$$

Podem utilitzar la recta per fer prediccions.

Si en una regió ***espvidam*=68.5**.

Aleshores: $\widehat{\textbf{espvidaf}} = -\mathbf{2.630} + \mathbf{1.113}(\mathbf{68.5}) = \mathbf{73.6}$

és la predicció de l'esperança de vida en aquest lloc, en base a la recta de regressió.

3.1. **Exercici.**

- Obteniu els coeficients de regressió **a** i **b** per a la parella de variables de la figura 2: $X =$*espvidam* i $Y =$*mortinf*.
- Escriviu-ne la recta de regressió.
- Feu la predicció de la mortalitat infantil d'una regió en la qual l'esperança de vida masculina sigui de 68.5 anys.
- Feu el mateix per a les variables de la figura 3.

4. Un exemple interessant: VOT al PSC i inmigració a catalunya

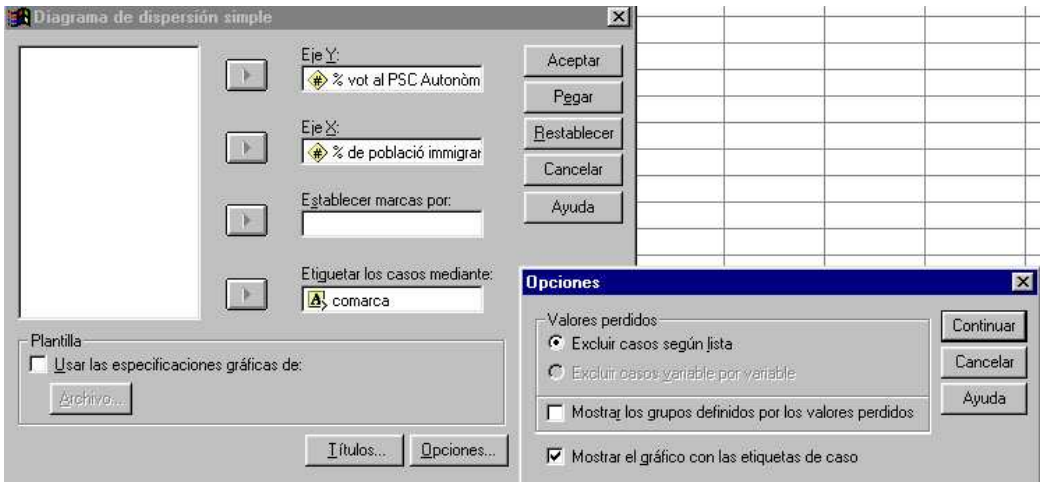
Estudiem la relació entre el percentatge de vot al PSC (a les autonòmiques de 1999) i el percentatge de població inmigrant de la resta d'Espanya a les diferents comarques de Catalunya (cada cas és una comarca). Obrim la base de dades:

Exemple de regressio Vot PSC-inmigracio.sav

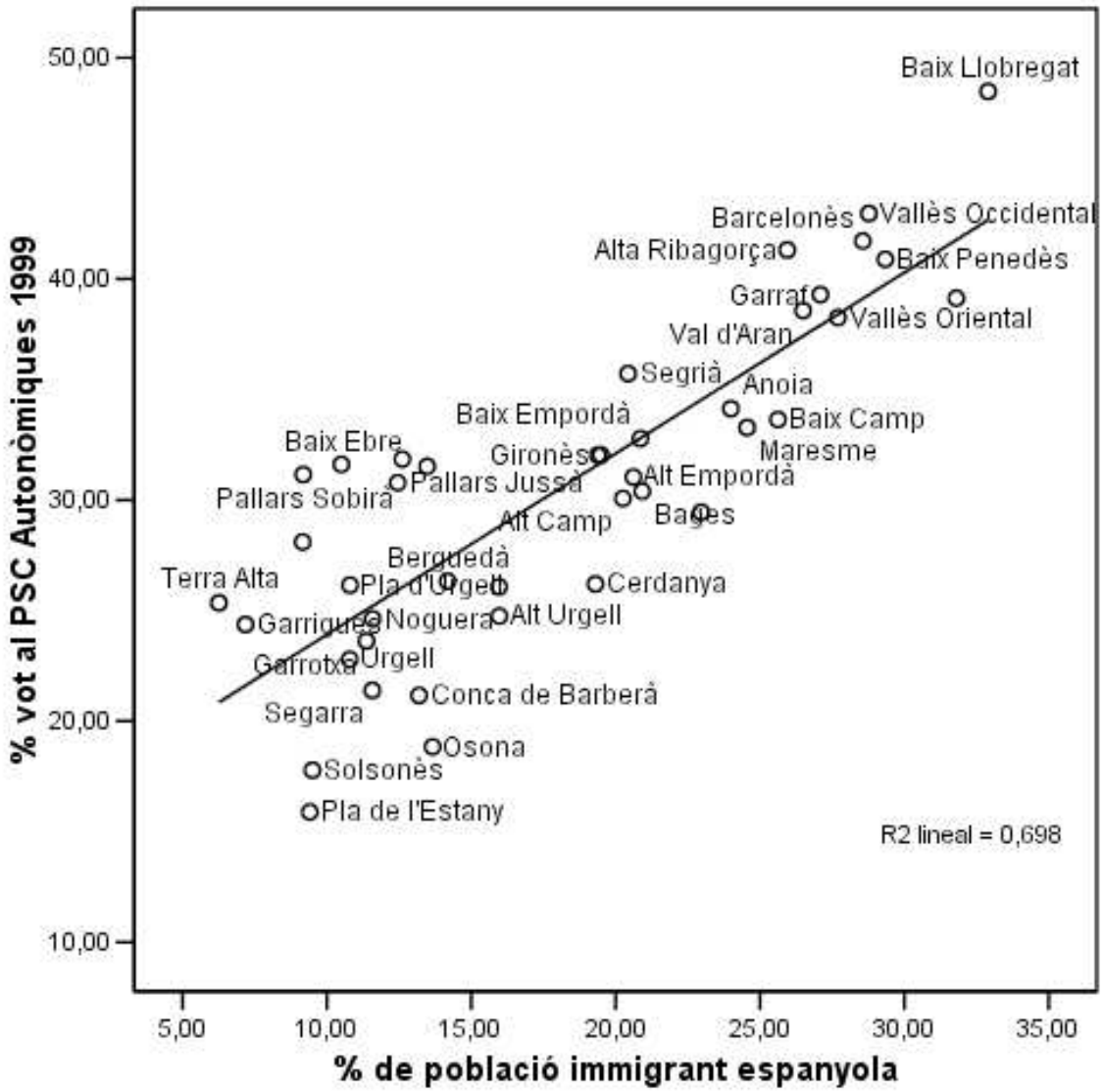
I fem el diagrama de dispersió entre les variables *% de vot al PSC* i *% de població inmigrant*.

Cal tenir clar quina és la variable independent X i quina la dependent Y.

En aquest diagrama volem veure els noms de les comarques (és a dir, activar les etiquetes). Per tant en la finestra de construcció dels diagrames, cliquem en les opcions i activem *mostrar el gráfico con las etiquetas del caso.*



Si a més s'ha fet ajustar la recta de regressió obtenim:



També en podem calcular els coeficients de correlació

Correlaciones

		% vot al PSC Autonòmiques 1999	% de població immigrant espanyola
% vot al PSC Autonòmiques 1999	Correlación de Pearson	1	,835**
	Sig. (bilateral)		,000
	N	41	41
% de població immigrant espanyola	Correlación de Pearson	,835**	1
	Sig. (bilateral)	,000	
	N	41	41

**. La correlación es significativa al nivel 0,01 (bilateral).

i els de la recta de regressió.

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,835[a]	,698	,690	4,10960

a. Variables predictoras: (Constante), % de població immigrant espanyola

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	15,731	1,695		9,280	,000
	% de població immigrant espanyola	,819	,086	,835	9,490	,000

a. Variable dependiente: % vot al PSC Autonòmiques 1999

► QÜESTIONS:

- Té sentit fer la recta de regressió? Per què?
- Quina és la recta de regressió obtinguda?
- Feu la predicció del % de vot del PSC, per a una hipotètica zona on el % de població inmigrada de la resta d'Espanya fos del 25%.