Regressió i correlació lineal. Teoria i pràctica

Francesc Carreras i Antoni Magaña
Departament de Matemàtiques,
Universitat Politècnica de Catalunya
Autor per a la correspondència.
Adreça electrònica: antonio.magana@upc.edu.

Resum

El principi dels mínims quadrats permet construir les dues rectes de regressió d'una distribució bidimensional. El coeficient de correlació entre les dues variables coincideix amb l'arrel quadrada del producte dels pendents de les dues rectes. La generalització a més variables segueix un patró similar. Després d'una revisió comentada d'aquest material teòric, l'apliquem a l'estudi de diverses situacions concretes i interpretem els resultats obtinguts.

Abstract

The principle of least squares makes it possible to construct the two regression lines of a two−dimensional distribution. The correlation coefficient between the two variables coincides with the square root of the product of the slopes of the two lines. Generalization to more variables follows a similar pattern. After a commented review of this theoretical material, we apply it to the study of several concrete situations and interpret the results obtained.

Paraules clau: distribucions bi i tridimensionals, correlació i regressió lineal, error quadràtic, sudoku, targetes de crèdit, venda lliure, competició futbolística, qualificacions acadèmiques, temperatura de xafogor.

Keywords: *two− and three−dimensional distributions, correlation and linear regression, quadratic error, sudoku, credit cards, free sale, football competition, academic grades, temperature−humidity index.*

Codi(s) MSC2010: primari 62 Estadística; secundaris 62J05 (Regressió Lineal) i 62P99 (Aplicacions).

MSC2010 Code(s): primary 62 Statistics; secondary 62J05 (Linear Regression) and 62P99 (Applications).

1. Introducció

Les matemàtiques constitueixen un camp científic molt singular. Per una banda, són una ciència independent, en el sentit que és possible definir conceptes, establir propietats i plantejar i resoldre problemes, tot expressat en termes estrictament matemàtics. Aquesta *internalitat* atrau i satisfà molts investigadors, que centren en ella els seus esforços i arriben a resultats admirables.

Per altra banda, hi ha matemàtics que admeten un vessant d'*externalitat*, segons el qual les matemàtiques són d'utilitat per estudiar problemes plantejats fora del seu àmbit estricte, modelitzar—los, i aplicar—hi la potència deductiva i el rigor típicament matemàtics per obtenir resultats interpretables i profitosos en l'àmbit que ha inspirat el problema. També molts investigadors es dediquen a aquests processos perquè, en certa manera, la connexió amb la realitat externa a les matemàtiques els dóna un plus de satisfacció. Per descomptat, hi ha matemàtics que treballen en ambdós vessants.

En aquest article hem adoptat una postura relativament mixta i ens proposem analitzar amb eines matemàtiques situacions ben conegudes que semblen allunyades de les matemàtiques. Hem seleccionat diverses situacions extretes del dia a dia i les hem analitzat amb eines estadístiques per obtenir conclusions sobre cada una d'elles. Entre les situacions que estudiem es troben, per exemple, la dificultat en la resolució d'un sudoku depenent del nombre de dades inicials o el comportament dels equips en el campionat de lliga de la Primera Divisió espanyola de futbol masculí. L'estudi d'aquests casos és la nostra aportació a l'externalitat que hem comentat abans. Tanmateix, també hem inclòs una secció tècnica prèvia on es justifiquen i es comenten amb deteniment els conceptes i les propietats que s'utilitzen després. Això pertany a l'àmbit de la internalitat.

En particular, com a matemàtics ens sembla admirable el *principi dels mínims quadrats*. Suposem que volem quantificar la dispersió d'un conjunt de dades, $\{x_1, x_2, \ldots, x_n\}$, respecte a una mesura de centralització x que resumeix el conjunt. És a dir, volem calcular la distància entre el conjunt de dades i aquesta mesura x. Una manera podria ser considerar la mitjana dels valors absoluts de les diferències entre x i cadascuna de les x_i, és a dir, prendre la *desviació mitjana*:

$$\frac{|x - x_1| + |x - x_2| + \cdots + |x - x_n|}{n}.$$

Tanmateix, el valor absolut és una funció no derivable en els punts on s'anul·la. Això va conduir a prendre la mitjana dels quadrats de les desviacions i, per conservar la magnitud en què s'expressen les dades, aplicar finalment l'arrel quadrada:

$$\sqrt{\frac{(x - x_1)^2 + (x - x_2)^2 + \cdots + (x - x_n)^2}{n}}.$$

Aquesta és l'anomenada *desviació típica* σ_x (respecte a x) del conjunt de dades.

De fet, la mateixa idea de *mitjana aritmètica* ja és una aplicació del principi dels mínims quadrats: la mitjana aritmètica $\overline{x}$ d'un conjunt de dades, definida com

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n},$$

i utilitzada amb tanta freqüència en qualsevol context, és l'única mesura de centralització que minimitza la suma dels quadrats de les desviacions respecte a ella. Aquesta propietat es pot demostrar fàcilment amb eines de càlcul infinitesimal o, alternativament, d'àlgebra lineal. Fins i tot la *mitjana geomètrica* $\overline{x}_g$, que es defineix com

$$\overline{x}_g = \sqrt[n]{x_1 x_2 \cdots x_n},$$

es basa en el fons en el mateix principi, ja que, aplicant logaritmes,

$$\log \overline{x}_g = \frac{\log x_1 + \log x_2 + \cdots + \log x_n}{n}.$$

El principi dels mínims quadrats és també el fonament de la recta de regressió d'una distribució bidimensional i, més en general, permet definir i calcular les solucions aproximades de qualsevol sistema d'equacions lineals. Ens ha semblat oportú mostrar, en la secció tècnica, com es construeix la recta de regressió, tant des del punt de vista del càlcul infinitesimal com des del de l'àlgebra lineal.

Després d'aquest prefaci, passem a descriure l'organització de l'article. En la secció 2 presentem, com a motivació, un dels exemples que estudiarem després. La secció 3, dividida en diverses subseccions, és la part tècnica, on recollim: un sumari dels conceptes i propietats que aplicarem; un complement dedicat a la recta de regressió d'una distribució bidimensional; un altre amb els comentaris addicionals que hem cregut adients, i una última subsecció on, més breument, descrivim les eines per estudiar distribucions tridimensionals.

Les seccions següents estan dedicades, cada una, a un exemple concret: els sudokus (secció 4), l'ús de les targetes de crèdit en un comerç (secció 5), el paper de la venda lliure en una farmàcia, on hi ha una remuneració oficial provinent del CatSalut (secció 6), el comportament dels equips de futbol en la lliga de Primera Divisió espanyola (secció 7), les qualificacions acadèmiques en tres assignatures estretament relacionades (secció 8) i el concepte de temperatura de xafogor (secció 9). En els exemples 4—7 i en la primera part del 8, tractem amb distribucions bidimensionals, mentre que en la segona part del 8 i en les quatre parts del 9 estudiem distribucions tridimensionals. La secció 10 presenta les conclusions del treball. Finalment, donem una bibliografia comentada sobre els temes tractats en l'article.

En tots els exemples el tractament és similar. Calculem els principals paràmetres de les distribucions rellevants: coeficient de correlació, recta de regressió (pla de regressió, en el casos tridimensionals finals), centre de gravetat i error(s) quadràtic(s). En els casos bidimensionals afegim el diagrama de dispersió. En tots els exemples comentem els resultats i, si escau, les particularitats específiques de cada situació.

Ens agradaria que aquest article animés alguns lectors a aplicar les matemàtiques a qualsevol nivell, des del més elemental, com el del nostre treball, fins al molt superior de la intel·ligència artificial, l'anàlisi de dades massives (*big data*) o l'aprenentatge automàtic (*machine learning*), «les joies de la corona» actuals de l'opció que hem anomenat externalitat.

2. Un exemple il·lustratiu: sudokus

Tothom sap què és un sudoku i quin problema planteja. És un trencaclosques matemàtic. Es tracta d'omplir amb les xifres de l'1 al 9 una quadrícula de 9×9 cel·les dividida en 9 subquadrícules de 3×3. No es pot repetir cap número en cap fila, en cap columna ni en cap subquadrícula. En cada sudoku concret, certes cel·les ja porten d'entrada el número posat (són les *dades*).

Hi ha dues intuïcions que són *vox populi* entre els aficionats als sudokus: (a) amb menys de 17 dades la solució no serà mai única; (b) amb 17 dades o més sempre hi ha una solució única. L'any 2014, juntament amb dos col·legues, el matemàtic Gary McGuire, de la Universitat de Dublín, va demostrar computacionalment que cap sudoku amb *exactament* 16 dades té solució única, però no que amb 17 dades o més es garanteixi que la solució és única. De fet, nosaltres hem trobat un sudoku amb 53 dades (!) que té tres solucions o més, cosa que desmunta de forma radical la intuïció (b). McGuire també va constatar que la majoria de sudokus tenen unes 25 dades. Un exemple de sudoku amb 24 dades es pot veure a la figura 1, on curiosament (no és freqüent) notem que hi ha simetria horitzontal i també vertical, i per tant també simetria central, en la posició de les dades.

	9			6			2	
			4	3	9			
		1				8		
9				7				2
			8		1			
3				5				4
		8				6		
			5	9	3			
	5			8			1	

Figura 1. Un sudoku.

Sembla lògic pensar que, com menys dades tingui un sudoku, més difícil serà resoldre'l. Això no és cap propietat general. Sovint hi ha valoracions de la dificultat dels sudokus amb asteriscos, però generalment les fan les persones que els proposen i, per tant, poden ser subjectives.

A la secció 4 tornarem a aquest exemple i ens plantejarem si hi ha relació entre el nombre de dades d'un sudoku i el seu grau de dificultat expressat numèricament (i, per força, subjectiu).

3. Fonaments teòrics

En aquesta secció exposem els conceptes i resultats de l'estadística descriptiva que seran d'utilitat en el nostre estudi d'exemples concrets. Hi afegim dues subseccions on tractem amb més deteniment alguns punts i proveïm comentaris i demostracions de determinats resultats. Posem més èmfasi en el que es refereix a distribucions bidimensionals i analitzem més breument el que fa referència a distribucions tridimensionals, com una mostra de la generalització natural del cas anterior a més dimensions.

3.1. Distribucions bidimensionals

3.1.1. Recta de regressió i correlació lineal

Tenim dues *sèries de dades* amb n termes: $\vec{x} = (x_1,x_2,\ldots,x_n)$ i $\vec{y} = (y_1,y_2,\ldots,y_n)$. Considerades conjuntament, formen una *distribució bidimensional*. De vegades interessa saber si hi ha relació entre les dades de la primera sèrie i les de la segona i, més concretament, si aquesta relació és aproximadament lineal. En cas que la resposta sigui afirmativa, té sentit construir l'anomenada *recta de regressió*, que dona explícitament la relació entre les variables x i y. Aquesta recta permet, entre altres coses, fer prediccions: donat un valor de x diferent de les dades de la primera sèrie, la imatge de x en la recta de regressió ens proporciona el possible valor de y que li correspondria.

L'única restricció que imposarem d'entrada és que els punts no es trobin tots en una mateixa vertical, és a dir, que no tinguem $x_1 = x_2 = \cdots = x_n$, perquè en aquest cas no existiria recta de regressió de y sobre x. Se sap que les *mitjanes aritmètiques* són:[1]

$$\overline{x} = \frac{\sum x_i}{n} \quad \text{i} \quad \overline{y} = \frac{\sum y_i}{n}.$$

I que les *desviacions típiques* σ_x i σ_y queden definides per les *variàncies* respectives:

$$\sigma_x^2 = \frac{\sum (x_i - \overline{x})^2}{n} \quad \text{i} \quad \sigma_y^2 = \frac{\sum (y_i - \overline{y})^2}{n}.$$

També se sap que

$$\sigma_x^2 = \overline{x_2} - \overline{x}^2 \quad \text{i, analògament,} \quad \sigma_y^2 = \overline{y_2} - \overline{y}^2,$$

on $\overline{x_2}$ i $\overline{y_2}$ són les respectives mitjanes aritmètiques dels quadrats. Així, la variància és «la mitjana dels quadrats menys el quadrat de la mitjana». La demostració és, per exemple per a les x:

1. $\sum$ significarà sempre $\sum_{i=1}^{n}$. En qualsevol altre cas, els límits del sumatori estaran especificats.

$$n\sigma_x^2 = \sum (x_i - \overline{x})^2 = \sum x_i^2 - 2\overline{x}\sum x_i + n\overline{x}^2 = \sum x_i^2 - n\overline{x}^2.$$

Fins aquí, els paràmetres s'apliquen a cada sèrie per separat. A partir d'ara, els paràmetres descriuran la distribució bidimensional i dependran de les dues sèries alhora.

La *covariància* de les dues sèries és $cov(x,y) = \sum (x_i - \overline{x})(y_i - \overline{y})$. És fàcil veure que

$$cov(x,y) = \vec{x}.\vec{y} - n\,\overline{x}\,\overline{y},$$

és a dir, el producte escalar dels vectors menys n vegades el producte de les mitjanes.

El *coeficient de correlació* $r(x,y)$ de la distribució bidimensional és

$$r(x,y) = \frac{cov(x,y)}{n\sigma_x\sigma_y} = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum (x_i - \overline{x})^2 \sum (y_i - \overline{y})^2}}.$$

Sempre tenim $r(x,y) \in [-1,1]$, i la segona expressió és la més còmoda per calcular-lo tabularment. Per a més detalls sobre correlació i covariància, vegeu més avall la secció 3.1.3.

La *recta de regressió de y sobre x* és de la forma

$$y - \overline{y} = m(x - \overline{x}),$$

on $(\overline{x},\overline{y})$ és el *centre de gravetat* de la distribució i el *pendent* m és

$$m = \frac{cov(x,y)}{n\sigma_x^2} = \frac{\vec{x}.\vec{y} - n\,\overline{x}\,\overline{y}}{n\sigma_x^2} = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\sigma_x^2} = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2},$$

on $\overline{xy} = \dfrac{\sum x_i y_i}{n}$ és la mitjana dels productes de les dues sèries. La quarta expressió és la més còmoda si estem fent els càlculs tabularment. Sol ser interessant i il·lustratiu dibuixar el *diagrama de dispersió* de les dades sobre uns eixos rectangulars $x-y$ juntament amb la gràfica de la recta de regressió. Per a més detalls sobre el fonament de la recta de regressió, vegeu la secció 3.1.2.

L'*error quadràtic* ε d'aquesta recta de regressió és

$$\varepsilon = \sqrt{\sum (y_i - y_i^*)^2}, \quad \text{on} \quad y_i^* = \overline{y} + m(x_i - \overline{x}) \quad \text{per a cada} \quad (x_i,y_i).$$

L'*error per capita* $\varepsilon_0 = \varepsilon/n$ sembla més interessant a efectes comparatius, ja que reparteix l'error total entre les n dades (x_i,y_i). Va bé per comparar distribucions bidimensionals amb diferent nombre de dades n. Quan volem comparar dues distribucions on els rangs de variació de y són molt diferents, un tercer error sembla adequat: consisteix a dividir l'error per capita per la mitjana aritmètica $\overline{y}$, que prenem com a representativa de tots els valors de y en cada cas, i obtenim, doncs, l'*error (per capita) normalitzat* $\varepsilon_0^* = \varepsilon_0/\overline{y}$.

3.1.2. La recta de regressió de y sobre x

Hi ha dos mètodes principals per fonamentar i obtenir la recta de regressió: un és l'*analític*, basat en el càlcul infinitesimal, i l'altre és l'*algebraic*, basat en l'àlgebra lineal. Tots dos estan inspirats pel *principi dels mínims quadrats*: la recta de regressió minimitza la suma dels quadrats de les desviacions respecte a la sèrie de les *y* o, equivalentment, l'error quadràtic definit unes línies més amunt.

A. Mètode analític. La recta de regressió és de la forma

$$y = a_0 + a_1 x.$$

Idealment, tots els punts de la distribució haurien de satisfer aquesta equació, és a dir,

$$y_i = a_0 + a_1 x_i \quad \text{per a } i = 1,2,\ldots,n.$$

Tanmateix, usualment aquest sistema de n equacions amb dues incògnites a_0, a_1 no té solució perquè els n punts no estan alineats. Considerem la suma dels quadrats dels errors comesos entre els valors reals y_i, donats per la distribució, i els que donaria la recta de regressió, $y_i^* = a_0 + a_1 x_i$. Tindrem un «error total» E, donat per

$$E = \sum (a_0 + a_1 x_i - y_i)^2.$$

Aquest error total E, que depèn de a_0 i a_1, és una funció contínua, diferenciable i no negativa, i no té cap màxim absolut perquè, quan a_0, a_1 o tots dos tendeixen cap a ∞, E tendeix també cap a ∞. En canvi, comprovarem que té un únic mínim absolut, que podrem localitzar com a mínim local. Primer imposem que les dues derivades parcials de E s'anul·lin (condició necessària de punt crític o extrem local):

$$\begin{cases} \dfrac{\partial E}{\partial a_0} = 2\sum (a_0 + a_1 x_i - y_i) = 0, \\ \dfrac{\partial E}{\partial a_1} = 2\sum (a_0 + a_1 x_i - y_i) x_i = 0, \end{cases} \quad \text{és a dir,} \quad \begin{cases} \sum y_i = n a_0 + a_1 \sum x_i, \\ \sum x_i y_i = a_0 \sum x_i + a_1 \sum x_i^2. \end{cases} \tag{1}$$

Aquest és un sistema de dues equacions lineals amb les incògnites a_0 i a_1. Ara fem les derivades parcials segones:

$$\frac{\partial^2 E}{\partial a_0^2} = 2n, \qquad \frac{\partial^2 E}{\partial a_1 \partial a_0} = 2\sum x_i,$$

$$\frac{\partial^2 E}{\partial a_0 \partial a_1} = 2\sum x_i, \quad \frac{\partial^2 E}{\partial a_1^2} = 2\sum x_i^2.$$

Aquestes quatre derivades parcials segones, disposades tal com estan, defineixen la matriu hessiana $H(E)$ de la funció E. Aquesta matriu és també la matriu dels coeficients del siste-

ma (1) donat per l'anul·lació de les primeres derivades parcials, i observem que és constant (independent de a_0 i a_1). Calculem ara el seu determinant (per simplicitat, prescindim dels coeficients 2 de totes les derivades segones, cosa que no afectarà el raonament):

$$\det H(E) = n\sum x_i^2 - \left(\sum x_i\right)^2.$$

Per veure que sempre és positiu, comprovarem primer que

$$\det H(E) = \sum_{1\leqslant i<j\leqslant n} (x_i - x_j)^2. \tag{2}$$

La demostració és per inducció sobre $n \geqslant 2$. (a) Per a $n = 2$ tenim

$$\det H(E) = 2(x_1^2 + x_2^2) - (x_1 + x_2)^2 = (x_1 - x_2)^2.$$

(b) Suposem $n \geqslant 3$ i que la fórmula (2) val per a $2,\ldots,n-1$. Aleshores,

$$\det H(E) = n\textstyle\sum_1^n x_i^2 - \left(\sum_1^n x_i\right)^2 = n\sum_1^{n-1} x_i^2 + nx_n^2 - \left(\sum_1^{n-1} x_i + x_n\right)^2 =$$

$$= n\textstyle\sum_1^{n-1} x_i^2 - \left(\sum_1^{n-1} x_i\right)^2 - 2x_n\sum_1^{n-1} x_i - x_n^2 + nx_n^2 =$$

$$= \left[(n-1)\textstyle\sum_1^{n-1} x_i^2 - \left(\sum_1^{n-1} x_i\right)^2\right] + \sum_1^{n-1} x_i^2 - 2x_n\sum_1^{n-1} x_i + (n-1)x_n^2.$$

Per hipòtesi d'inducció, coneixem la suma dels dos primers termes (entre claudàtors); per tant,

$$\det H(E) = \sum_{1\leqslant i<j\leqslant n-1} (x_i - x_j)^2 + R,$$

on R recull els dos termes restants. Ara tenim

$$R = \textstyle\sum_1^{n-1} x_i^2 - 2x_n\sum_1^{n-1} x_i + (n-1)x_n^2 = \sum_1^{n-1}(x_i - 2x_n)x_i + (n-1)x_n^2 =$$

$$= \textstyle\sum_1^{n-1}(x_i - 2x_n)x_i + \sum_1^{n-1} x_n^2 = \sum_1^{n-1}(x_i^2 - 2x_ix_n + x_n^2) =$$

$$= \textstyle\sum_1^{n-1}(x_i - x_n)^2.$$

Finalment,

$$\det H(E) = \sum_{1\leqslant i<j\leqslant n-1} (x_i - x_j)^2 + \sum_1^{n-1}(x_i - x_n)^2 = \sum_{1\leqslant i<j\leqslant n} (x_i - x_j)^2.$$

Com que no tots els punts de la distribució es troben en una única vertical, almenys hi haurà un parell *i,j* tals que $x_i \neq x_j$, i $\det H(E)$ serà positiu. Per tant, hi haurà només un punt crític.

Aplicant, per exemple, la regla de Cramer al sistema (1), trobem els coeficients i l'equació de la recta de regressió, que és

$$y = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2} x + \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2}.$$

En el cas $n = 2$, l'equació anterior es redueix a

$$y = \frac{y_1 - y_2}{x_1 - x_2} x + \frac{x_1 y_2 - x_2 y_1}{x_1 - x_2},$$

que és simplement la recta que passa pels dos punts (x_1,y_1) i (x_2,y_2).

Ara comprovem que en el punt crític hi ha un mínim absolut de E. La forma quadràtica definida per la matriu $H(E)$ és definida positiva segons el criteri de Sylvester, ja que el seu primer coeficient i el seu determinant són positius. Això implica que la fórmula de Taylor de grau 2 de la funció E en el punt crític $p = (a_0,a_1)$ és, per a qualsevol punt (x,y) del pla,

$$E = f(x,y) = f(p) + \frac{1}{2} \begin{pmatrix} x - a_0 & y - a_1 \end{pmatrix} H(E) \begin{pmatrix} x - a_0 \\ y - a_1 \end{pmatrix} > f(p),$$

ja que $H(E)$ és definida positiva (el residu de la fórmula és 0 perquè E és una funció polinòmica de segon grau de a_0,a_1). Això demostra que en el punt crític $p = (a_0,a_1)$ hi ha (l'únic) mínim absolut de E.

Tornant al cas general, és immediat comprovar amb la primera de les equacions de (1) que la recta de regressió passa pel centre de gravetat de la distribució, ja que

$$\frac{\sum y_i}{n} = a_0 + a_1 \frac{\sum x_i}{n}.$$

B. Mètode algebraic. Tornem a suposar que la recta que busquem és de la forma

$$y = a_1 x + a_0.$$

Novament, tots els punts de la distribució haurien de satisfer aquesta equació. Això ens donaria un sistema de n equacions lineals amb dues incògnites: a_1 i a_0. Escrit en forma breu seria

$$a_1 x_i + a_0 = y_i \quad \text{per } i = 1,2,\ldots,n.$$

En forma matricial el sistema s'escriu

$$CX = D$$

o, més explícitament,

$$\begin{pmatrix} x_1 & 1 \\ x_2 & 1 \\ \dots & 1 \\ x_n & 1 \end{pmatrix} \begin{pmatrix} a_1 \\ a_0 \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix}.$$

En general, el sistema serà incompatible llevat que els n punts estiguin alineats. En termes vectorials, busquem dins de l'espai $\mathbb{R}^n$ un vector u' del pla H generat per $c_1 = (x_1,x_2,\dots,x_n)$ i $c_2 = (1,1,\dots,1)$ tal que $u' = u$, i és, doncs,

$$u' = a_1(x_1,x_2,\dots,x_n) + a_0(1,1,\dots,1) = (y_1,y_2,\dots,y_n) = u.$$

Que el sistema sigui incompatible significa que $u = (y_1,y_2,\dots,y_n) \notin H$, i per tant la relació precedent és impossible. L'alternativa proposada pel *principi dels mínims quadrats* consisteix a buscar el vector $u' \in H$ *més pròxim* a u, i per tant u' ha de ser la *projecció ortogonal* de u sobre H perquè és la que *minimitza* $||u - u'||$ entre tots els vectors $u' \in H$. La condició que ha de complir $u' \in H$ per ser la projecció ortogonal de u sobre H és que $u - u' \perp H$, que es tradueix en

$$u - u' \perp c_1 \quad \text{i} \quad u - u' \perp c_2.$$

Això equival a imposar

$$u \cdot c_1 = u' \cdot c_1 \quad \text{i} \quad u \cdot c_2 = u' \cdot c_2,$$

que podem expressar matricialment com

$$C^tCX = C^tD,$$

o, introduint $A = C^tC$ i $B = C^tD$ (A resulta quadrada 2×2 i simètrica),

$$AX = B,$$

sistema compatible i, a més, determinat (solució única), ja que explícitament queda

$$A = \begin{pmatrix} \sum x_i^2 & \sum x_i \\ \sum x_i & n \end{pmatrix} \qquad \text{i} \qquad B = \begin{pmatrix} \sum x_iy_i \\ \sum y_i \end{pmatrix},$$

de manera que

$$\det A = \det H(E) = \sum_{1 \leqslant i < j \leqslant n} (x_i - x_j)^2 \neq 0$$

i la solució única del sistema $AX = B$ és la mateixa que dona el mètode analític. A més,

$$\varepsilon = ||u - u'|| = \sqrt{E},$$

que és el que s'anomena *error quadràtic* de la recta de regressió obtinguda.

No és difícil veure que el pendent m de la recta de regressió definit en la secció 3.1.1 coincideix amb el pendent a_1 calculat en aquesta secció. En efecte, dividint per n^2 numerador i denominador de a_1 tenim, recordant que $\sigma_x^2 = \overline{x_2} - \overline{x}^2$,

$$a_1 = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2} = \frac{\frac{\sum x_i y_i}{n} - \frac{\sum x_i}{n}\frac{\sum y_i}{n}}{\frac{\sum x_i^2}{n} - \left(\frac{\sum x_i}{n}\right)^2} = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\sigma_x^2} = m,$$

segons la tercera expressió de m donada en la secció 3.1.1.

3.1.3. Covariància cov(x, y) i coeficient de correlació r(x, y)

Per començar, convé notar que la covariància és simètrica, és a dir, $cov(y,x) = cov(x,y)$. Si les dades estan totes en una mateixa vertical o bé en una mateixa horitzontal, és fàcil veure que $cov(y,x) = 0 = cov(x,y)$.

Ara cal justificar les dues expressions d'aquest paràmetre que apareixen a la fórmula del coeficient de correlació $r(x,y)$ a la secció 3.1.1. Per a això, només cal comprovar la igualtat dels numeradors, ja que els denominadors són, òbviament, idèntics. I, en efecte,

$$cov(x,y) = \vec{x}.\vec{y} - n\,\overline{x}\,\overline{y} = \sum x_i y_i - n\,\overline{x}\,\overline{y} = n\,\overline{xy} - n\,\overline{x}\,\overline{y} = n(\overline{xy} - \overline{x}\,\overline{y}),$$

mentre que

$$\sum (x_i - \overline{x})(y_i - \overline{y}) = \sum x_i y_i - \overline{y}\sum x_i - \overline{x}\sum y_i + n\overline{x}\,\overline{y} = n(\overline{xy} - \overline{x}\,\overline{y}).$$

Passem ara al coeficient de correlació. També per simetria tenim

$$r(y,x) = r(x,y)$$

i, en els dos casos extrems, l'horitzontal i el vertical, $r(x,y) = 0$.

Es defineix la *variació total* de y com

$$n\sigma_y^2 = \sum (y_i - \overline{y})^2.$$

Ara comprovarem que

$$\sum(y_i - \overline{y})^2 = \sum(y_i - y_i^*)^2 + \sum(y_i^* - \overline{y})^2.$$

Comencem per la identitat

$$y_i - \overline{y} = (y_i - y_i^*) + (y_i^* - \overline{y}).$$

Elevant al quadrat els dos termes i sumant per a tot *i* obtenim

$$\sum(y_i - \overline{y})^2 = \sum(y_i - y_i^*)^2 + \sum(y_i^* - \overline{y})^2 + 2S,$$

on

$$S = \sum(y_i - y_i^*)(y_i^* - \overline{y}) = \sum(y_i - a_0 - a_1x_i)(a_0 + a_1x_i - \overline{y}) =$$

$$= a_0\sum(y_i - a_0 - a_1x_i) + a_1\sum(y_i - a_0 - a_1x_i)x_i - \overline{y}\sum(y_i - a_0 - a_1x_i) = 0$$

aplicant les equacions (1) als coeficients de a_0,a_1,$\overline{y}$.

Així doncs, la variació total queda dividida en dos termes:

$$n\sigma_y^2 = \sum(y_i - \overline{y})^2 = \sum(y_i - y_i^*)^2 + \sum(y_i^* - \overline{y})^2.$$

El primer terme del segon membre és la *variació no explicada*, mentre que el segon terme és la *variació explicada* perquè segueix un patró de referència, cosa que no fa el primer.

El *coeficient de correlació* $r(x,y)$ es defineix com l'arrel quadrada del quocient entre la variació explicada i la variació total, amb signe positiu o negatiu segons el signe del pendent *m* de la recta de regressió. Simbòlicament:

$$r(x,y) = \pm\sqrt{\frac{\sum(y_i^* - \overline{y})^2}{\sum(y_i - \overline{y})^2}}\,.$$

Aquesta és una definició raonada de $r(x,y)$. Haurem de comprovar que aquesta expressió és equivalent a la de $r(x,y)$ que apareix a la secció 3.1.1. Abans, observem que $r(x,y) \in [-1,1]$ i que, com a casos particulars extrems: (a) si la variació explicada és nul·la, $r(x,y) = 0$; (b) si la variació no explicada és nul·la, $r(x,y) = \pm 1$.

Fem la comprovació anunciada. De l'equació de la recta de regressió tenim

$$y_i^* - \overline{y} = a_1(x_i - \overline{x}).$$

Per tant,

$$r^2(x,y) = \frac{\sum(y_i^* - \overline{y})^2}{\sum(y_i - \overline{y})^2} = \frac{a_1^2 \sum(x_i - \overline{x})^2}{\sum(y_i - \overline{y})^2}.$$

Ara bé, segons la quarta expressió del pendent m en la secció 3.1.1,

$$a_1 = m = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2}\,.$$

Substituint a l'equació prèvia i simplificant un factor $\sum(x_i - \overline{x})^2$ resulta

$$r^2(x,y) = \frac{\left[\sum(x_i - \overline{x})(y_i - \overline{y})\right]^2}{\sum(x_i - \overline{x})^2 \sum(y_i - \overline{y})^2}$$

i, en definitiva,

$$r(x,y) = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2 \sum(y_i - \overline{y})^2}}\,.$$

El signe $\pm$ ja va incorporat al numerador d'aquesta expressió.

Considerem finalment la *covariància per capita*

$$\sigma_{xy} = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{n}\,.$$

Aleshores en resulta una expressió simplificada del coeficient de correlació:

$$r(x,y) = \frac{\sigma_{xy}}{\sigma_x \sigma_y}\,.$$

També el pendent m de la recta de regressió de y sobre x admet expressions més simples:

$$m = \frac{\sigma_{xy}}{\sigma_x^2} = r(x,y)\frac{\sigma_y}{\sigma_x}\,.$$

Si els punts de la distribució no estan tampoc en una mateixa horitzontal, pensem en la recta de regressió de x sobre y, que és anàloga *mutatis mutandis* a la de y sobre x que hem estat considerant exclusivament fins ara, i escrivim

$$x - \overline{x} = m'(y - \overline{y}).$$

És obvi que

$$m' = r(x,y)\frac{\sigma_x}{\sigma_y}\,,$$

i, per tant,

$$m\, m' = r(x,y)^2,$$

expressió que es podria adoptar com a definició del coeficient de correlació. En particular, si les dues rectes de regressió coincideixen, $m = m'$, hi ha una correlació lineal perfecta i, per tant, $r(x,y) = 1$. En canvi, si $r(x,y) = 0$, les rectes de regressió són perpendiculars. Es pot interpretar que, en general, el coeficient de correlació ve a ser una mesura de l'angle α que formen les dues rectes, encara que la relació no és senzilla (lineal):

$$\tan(\alpha) = \frac{m - m'}{1 + m\, m'} = \frac{r}{1 + r^2} \frac{\sigma_y^2 - \sigma_x^2}{\sigma_x \sigma_y}.$$

3.2. Distribucions tridimensionals

Per completar aquest resum teòric considerarem el cas d'una distribució amb tres variables. Tenim, doncs, tres *sèries de dades* amb n termes cada una: $\vec{X} = (X_1, X_2, \ldots, X_n)$, $\vec{Y} = (Y_1, Y_2, \ldots, Y_n)$ i $\vec{Z} = (Z_1, Z_2, \ldots, Z_n)$. Considerades conjuntament, formen una *distribució tridimensional*. L'única restricció que imposarem d'entrada és que els punts (X_i, Y_i, Z_i) no es trobin tots en una mateixa vertical (paral·lela a l'eix Z), perquè en aquest cas no tindria sentit establir una relació de Z com a funció de X i Y.

Les *mitjanes aritmètiques* $\overline{X}$, $\overline{Y}$ i $\overline{Z}$ es defineixen com en el cas bidimensional.[2] El punt $(\overline{X}, \overline{Y}, \overline{Z})$ serà el *centre de gravetat* de la distribució. Reservarem les corresponents lletres minúscules x, y i z per a les *variables normalitzades*, que són les desviacions de les variables originals respecte a les seves mitjanes aritmètiques:

$$x = X - \overline{X}, \quad y = Y - \overline{Y} \quad \text{i} \quad z = Z - \overline{Z}.$$

També es defineixen com en el cas bidimensional les *desviacions típiques* σ_X, σ_Y i σ_Z, determinades per les respectives *variàncies*. I segueix tenint validesa que cada variància és «la mitjana dels quadrats menys el quadrat de la mitjana». Val la pena remarcar tres punts de fàcil comprovació: (a) la mitjana aritmètica de les variables normalitzades és nul·la, és a dir, $\sum x_i = \sum y_i = \sum z_i = 0$; (b) les seves desviacions típiques coincideixen amb les de les variables originals, $\sigma_x = \sigma_X$, $\sigma_y = \sigma_Y$ i $\sigma_z = \sigma_Z$; (c) el coeficient de correlació de dues qualssevol d'aquestes variables coincideix amb el de les seves variables normalitzades, per exemple, $r(X,Y) = r(x,y)$. Tot això, que també era vàlid en el cas bidimensional, serà d'utilitat aquí per alleugerir la notació en alguns moments.

Ens ocuparem ara d'estudiar tres conceptes principals: (a) el de *pla de regressió lineal* de la distribució tridimensional, que descriu, si existeix, la dependència aproximadament lineal d'una de les variables (escollirem la Z) en funció de les altres dues (X i Y) i es basa en el *principi dels mínims quadrats*; (b) l'*error quadràtic* que dona aquest pla d'equació $Z = f(X,Y)$ quan fem

2. $\sum$ seguirà significant sempre $\sum_{i=1}^{n}$.

servir aquesta funció per estimar valors de Z a partir de valors de X i Y; i (c) el *coeficient de correlació (global)* que mesura la qualitat de la dependència donada per la funció lineal f.

El *pla de regressió lineal* serà de la forma

$$Z = aX + bY + c. \tag{3}$$

Les constants a,b,c són els *coeficients de regressió parcial*. La situació ideal (dependència lineal perfecta) es donaria si

$$Z_i = aX_i + bY_i + c \quad \text{per a } i = 1,2,\ldots,n.$$

Introduint les variables normalitzades x,y,z, l'equació (3) passa a ser

$$z = ax + by.$$

Seguint un procés similar al del cas bidimensional, partim de

$$z_i = ax_i + by_i. \tag{4}$$

Multiplicant ara l'equació (4) separadament, primer per x_i i després per y_i, i sumant en cada cas terme a terme, obtenim

$$\begin{cases} \sum z_i x_i = a\sum x_i^2 + b\sum x_i y_i, \\ \\ \sum z_i y_i = a\sum x_i y_i + b\sum y_i^2, \end{cases} \tag{5}$$

que és el sistema d'equacions lineals que determina els coeficients a,b.[3] Escrit matricialment és:

$$\begin{pmatrix} \sum x_i^2 & \sum x_i y_i \\ \sum x_i y_i & \sum y_i^2 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sum z_i x_i \\ \sum z_i y_i \end{pmatrix}.$$

Aplicant, per exemple, la regla de Cramer, la solució del sistema dona

$$a = \frac{\sum z_i x_i \sum y_i^2 - \sum z_i y_i \sum x_i y_i}{\sum x_i^2 \sum y_i^2 - (\sum x_i y_i)^2} \quad \text{i} \quad b = \frac{\sum z_i x_i \sum x_i^2 - \sum z_i x_i \sum x_i y_i}{\sum x_i^2 \sum y_i^2 - (\sum x_i y_i)^2}.$$

3. El sistema equivalent amb les variables originals és

$$\sum Z_i = a\sum X_i + b\sum Y_i + cn,$$
$$\sum Z_i X_i = a\sum X_i^2 + b\sum X_i Y_i + c\sum X_i,$$
$$\sum Z_i Y_i = a\sum X_i Y_i + b\sum Y_i^2 + c\sum Y_i.$$

El tractament teòric d'aquest sistema és força més complicat que el del sistema (5).

Aleshores, l'*equació del pla de regressió de Z sobre X i Y* és

$$Z - \overline{Z} = a(X - \overline{X}) + b(Y - \overline{Y}).$$

Nota. Tanmateix, és convenient indicar que, en un exemple numèric, és millor preparar una taula numèrica amb les dades originals i els càlculs addicionals necessaris i resoldre el sistema detallat en la nota 4, per obtenir l'equació del pla amb la forma

$$Z = aX + bY + c.$$

D'aquesta manera reduïm els petits errors que indefectiblement impliquen les aproximacions. També arrosseguem errors si introduïm els anomenats *coeficients de correlació binària*, que són els coeficients de correlació de les tres distribucions bidimensionals subjacents a la tridimensional que estem considerant:

$$r(X,Y) = r(x,y), \quad r(X,Z) = r(x,z) \quad \text{i} \quad r(Y,Z) = r(y,z).$$

Després de tenir en compte les relacions

$$\sum x_i^2 = n\sigma_x^2, \quad \sum y_i^2 = n\sigma_y^2 \quad \text{i} \quad \sum x_i y_i = n\sigma_x \sigma_y r(x,y)$$

i, anàlogament,

$$\sum x_i z_i = n\sigma_x \sigma_z r(x,z) \quad \text{i} \quad \sum y_i z_i = n\sigma_y \sigma_z r(y,z),$$

i simplificar alguns termes, és cert que les equacions (5) queden

$$a\sigma_x + b\sigma_y r(x,y) = \sigma_z r(x,z)$$

$$a\sigma_x r(x,y) + b\sigma_y = \sigma_z r(y,z),$$

o, en forma matricial,

$$\begin{pmatrix} \sigma_x & \sigma_y r(x,y) \\ \sigma_x r(x,y) & \sigma_y \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \sigma_z \begin{pmatrix} r(x,z) \\ r(y,z) \end{pmatrix}.$$

Aplicant la regla de Cramer obtenim les expressions de *a,b* i, com a *equació del pla de regressió de z sobre x i y*,[4]

$$\frac{z}{\sigma_z} = \frac{r(x,z) - r(x,y)r(y,z)}{1 - r(x,y)^2} \frac{x}{\sigma_x} + \frac{r(y,z) - r(x,y)r(x,z)}{1 - r(x,y)^2} \frac{y}{\sigma_y}.$$

4. És fàcil veure que, eliminant per exemple la *y*, aquesta equació es redueix a

$$\frac{z}{\sigma_z} = r(x,z)\frac{x}{\sigma_x},$$

que és l'equació de la recta de regressió bidimensional de *z* sobre *x*.

No obstant l'aspecte agradable i simètric d'aquesta equació, els coeficients de correlació binària que hi apareixen arrosseguen errors d'aproximació pel seu propi càlcul, i per tornar a les variables originals del problema cal, a més, introduir les mitjanes aritmètiques. Per això en la nota de més amunt s'ha inclòs el comentari sobre els exemples numèrics.

L'*error quadràtic* ε de Z sobre X i Y es defineix per

$$\varepsilon_Z(X,Y) = \sqrt{\sum (Z_i - Z_i^*)^2}, \quad \text{on} \quad Z_i^* = \overline{Z} + a(X_i - \overline{X}) + b(Y_i - \overline{Y}) \quad \text{per a cada} \quad (X_i, Y_i).$$

Tanmateix, és més habitual considerar l'*error típic de l'estima*, que ve a ser un *error per capita* i és donat per

$$\varepsilon_Z^0(X,Y) = \sqrt{\frac{\sum (Z_i - Z_i^*)^2}{n}}.$$

El *coeficient de correlació múltiple* $R_Z(X,Y)$ de la distribució tridimensional es defineix com

$$R_Z(X,Y) = \sqrt{1 - \frac{\varepsilon_Z^0(X,Y)^2}{\sigma_Z^2}}.$$

Aquest coeficient de correlació múltiple sempre varia entre 0 i 1. Com més s'acosta a 1, millor és la relació lineal entre les variables. Com més s'acosta a 0, la relació lineal és pitjor. Si el coeficient és 1, la relació lineal és perfecta. Si és 0, no hi ha relació lineal però n'hi pot haver d'un altre tipus. (Hi ha expressions alternatives d'aquests dos nous paràmetres en termes dels coeficients de correlació binària que, per les raons ja explicades, no farem servir.)

4. Exemple: sudokus

Ens plantegem donar resposta a la qüestió següent: hi ha relació entre el nombre de dades d'un sudoku i el seu grau de dificultat expressat numèricament?

Un de nosaltres va tenir ocasió de resoldre els 108 sudokus d'un quadern publicat per Edigrama el 2021, on els graus de dificultat anunciats eren: *** (16 problemes), **** (46 problemes) o ***** (46 problemes). Els va completar en, aproximadament, un mes. Un d'ells era l'exemple de la figura 1 (secció 2) i tots tenien el mateix tipus de simetria horitzontal i vertical.

Després de resoldre cada sudoku, va valorar-ne la dificultat, segons aquestes opcions: *F* (fàcil), *R* (regular), *D* (difícil) i *D** (molt difícil). Són els *graus de dificultat*, en sentit creixent. Evidentment, aquesta valoració també és subjectiva. De fet, es va observar que, sovint, la dificultat també depèn una mica de la «lucidesa» de qui intenta resoldre el sudoku en el moment de trobar la solució, i això també és lògic. No és estrany que d'un problema que es deixa encallat a la nit, se'n trobi la solució l'endemà al matí sense excessives dificultats.

No ens aturarem a explicar la diferència entre 'difícil' i 'molt difícil', però és cert que és notable. Per tant, a l'hora de quantificar els quatre graus, l'assignació ha estat: $F = 1$, $R = 2$, $D = 3$ i $D^* = 5$. Aquí també hi ha subjectivitat, és clar.

Vegem en la taula 1 la graella de freqüències. La variable *x* descriu les dades de cada sudoku, des de 21 fins a 31, mentre que la variable *y* descriu el grau de dificultat. Dins de cada cel·la hi ha la freqüència de cada parell (dades, grau). Fins aquí arriba la descripció de la part empírica, és a dir, els resultats de l'experiment.

Taula 1. Freqüències de cada parell (dades, grau de dificultat).

$y \downarrow x \rightarrow$	21	22	23	24	25	26	27	28	29	30	31	sumes
$D^* = 5$				1		1		1				3
$D = 3$			1		4	1		1				7
$R = 2$	1		3	1	5	9	5	3				27
$F = 1$				5	8	12	20	18	2	2	4	71
sumes	1		4	7	17	23	25	23	2	2	4	108

Per exemple, dels 25 sudokus amb $x = 27$ dades, n'hi ha 20 que són fàcils ($F = 1$) i 5 de regulars ($R = 2$).

Passem ara a la part analítica. Es tracta de comprovar si 'menys dades, més difícil' és una hipòtesi plausible i quina és la recta de regressió lineal de *y* sobre *x*, és a dir, si es pot preveure amb certa aproximació el grau de dificultat d'un sudoku a partir de les seves dades. Amb aquest objectiu hem estudiat la distribució bidimensional *x,y*.

Els resultats són els següents:

- El *coeficient de correlació* $r = r(x,y)$ entre dues variables, que sempre està a l'interval $[-1,1]$, dona aquí $r = -0{,}29$. Significa que hi ha poca correlació entre *y* i *x*, i el signe negatiu indica que, molt *grosso modo*, com més dades, menys dificultat.
- La *recta de regressió de y sobre x* dona $y = 5{,}20 - 0{,}14x$. Es pot veure representada en el diagrama de la figura 2. Com està previst, passa pel centre de gravetat $(\overline{x},\overline{y})$ = (26,53, 1,49).
- L'*error quadràtic* d'aquesta recta és $\varepsilon = 8{,}383$. Més interessant sembla l'*error quadràtic per capita* ε_0, que reparteix equitativament l'error total entre tots els sudokus estudiats i dona $\varepsilon_0 = 0{,}078$ de mitjana per a cada un.

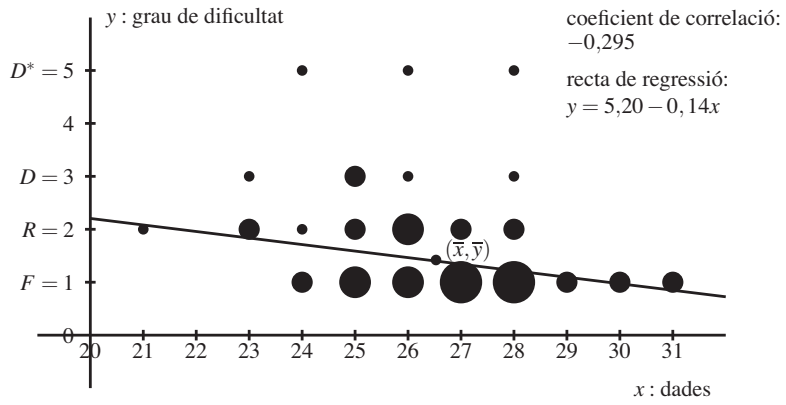

Figura 2. Diagrama de dispersió dels sudokus estudiats.

La figura 2 ens dona el *diagrama de dispersió* de l'experiment. Cal advertir que el gruix de cada cercle és aproximadament proporcional a la freqüència de la parella (dades, grau) corresponent. Es pot observar que la recta de regressió decreix i travessa el nucli de la distribució. A la secció següent compararem aquests resultats amb els que s'obtenen en un cas totalment diferent.

5. Exemple: venda amb targetes de crèdit

El 2008 es va inaugurar una farmàcia en un barri modest de la tercera ciutat catalana. El «farmacèutic consort», amic nostre, va rebre diversos encàrrecs i ens en va parlar recentment: un d'ells era revisar cada dia els tiquets dels pagaments amb targeta fets pels clients, per comprovar si el banc abonava l'import net corresponent després de quedar-se la comissió del 0,3 % sobre l'import brut. Des de l'inici de la pandèmia feia aquesta feina no a diari sinó un cop per setmana, cosa que suposava comptar cada vegada 6 feixos de targetes (de dilluns a dissabte, llevat dels festius).

Després de 14 anys, segons va explicar, deu haver revisat milers de tiquets, però mai no els havia comptat. L'any 2021 es va decidir a fer-ho: disposa, doncs, per a cada mes del 2021, del nombre de compres fetes amb targeta i l'import total net. Total: 4.400 targetes. La sensació aproximada que té és que «com més targetes hi ha, més puja l'import total», encara que aquesta no és una regla rígida. Com podríem comprovar amb les seves dades quin grau de validesa té la seva sensació intuïtiva?

Tenim dues sèries de *dades*, cada una amb 12 components: *el nombre de targetes de cada mes* i *l'import total net mensual*. Són valoracions objectives. Tanmateix, l'experiment queda condicionat per molts factors: l'emplaçament de la farmàcia, el tipus de clientela que té, l'elecció que fa cada client entre pagar en efectiu o amb targeta, l'efecte de la pandèmia, etc. I només tenim les dades d'un any, mes per mes. Per tant, els resultats difícilment serien extrapolables, no només a altres farmàcies sinó fins i tot a altres anys d'aquesta.

La taula 2 ens dona les dades de la distribució bidimensional que estudiarem.

Taula 2. Nombre mensual de targetes de crèdit i imports nets corresponents.

mes	gener	febrer	març	abril	maig	juny
targetes (x)	294	321	419	361	376	371
import (y)	4.261,83	3.983,60	5.713,97	4.802,22	5.624,81	5.561,59
mes	juliol	agost	setembre	octubre	novembre	desembre
targetes (x)	366	281	344	420	381	466
import (y)	5.452,62	4.070,60	5.136,48	5.196,56	4.575,67	6.612,53

El nombre de targetes (x) és una quantitat de tres xifres, mentre que l'import en euros (y) ve donat amb quatre xifres i dos decimals. Aquesta diferència fa que el diagrama de dispersió hagi de tenir a l'eix de les y una escala diferent que a l'eix de les x.

Passem ara a la part analítica. Volem comprovar si «com més targetes, més import net total» era una hipòtesi plausible i quina era la recta de regressió lineal de *y* sobre *x*, és a dir, si es podia preveure amb certa aproximació l'import mensual a partir del nombre de targetes.

Els resultats són els següents:

- El *coeficient de correlació r* entre dues variables, que sempre està a l'interval $[-1,1]$, dona aquí $r = 0{,}85$. Significa que hi ha una correlació positiva notable entre *y* i *x*, cosa que *grosso modo* confirma la intuïció del nostre amic.
- La *recta de regressió de y sobre x* dóna $y = 12{,}5x + 500$. Es pot veure representada en el diagrama de dispersió de la figura 3.
- L'*error quadràtic* d'aquesta recta és $\varepsilon = 1{,}346{,}20$ i l'*error quadràtic per capita* dona $\varepsilon_0 = 112{,}18$ de mitjana per a cada mes. Aquest és un error notable a primera vista, és a dir, en termes absoluts, però cal tenir en compte el rang dels valors de *y*.

La figura 3 dona el *diagrama de dispersió* de l'experiment. Es pot observar que la recta de regressió creix i s'acosta força a la majoria dels 12 punts (x,y). El número que acompanya cada punt en aquesta figura representa el mes corresponent.

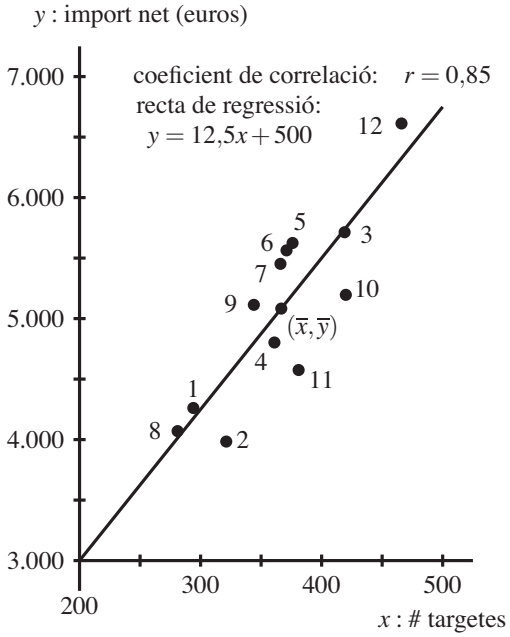

Figura 3. Diagrama de dispersió de les vendes amb targeta.

Comparem els resultats d'aquest estudi de les targetes amb els dels sudokus. En el cas dels sudokus, la correlació $r = -0{,}29$ és baixa en termes absoluts, és a dir, obviant el signe, i l'error per capita també és «petit»; en canvi, per a les targetes, la correlació $r = 0{,}85$ és força alta i l'error per capita també és «molt gran». Tanmateix, la recta de regressió per a les targetes sembla més ajustada a les dades que la dels sudokus. El motiu d'aquesta contradicció aparent és que el rang de les *y* és $[1,5]$ per als sudokus i aproximadament $[3{,}983\,,\,6{,}612]$ per a les

targetes. Introduïm un segon factor de normalització, l'error per capita normalitzat $\varepsilon_0^* = \varepsilon_0/\overline{y}$, definit al final de la secció 3.1.1. La taula 3 compara tots els valors rellevants.

Taula 3. Comparació amb l'exemple dels sudokus.

	error quadràtic	error per capita	mitjana de les *y*	error normalitzat
	ε	ε_0	$\overline{y}$	ε_0^*
sudokus	8,383	0,078	1,49	0,0523
targetes	1.346,20	112,18	5.082,71	0,0221

Ara es veu clar que, adequant en cada cas l'error per capita al rang de les *y* (o, equivalentment, a l'entorn de la mitjana aritmètica de les *y*), l'error per capita normalitzat de les targetes és bastant menor que el dels sudokus, com suggereixen les gràfiques de les rectes de regressió respectives. Sembla, doncs, que aquest últim paràmetre reflecteix millor els diagrames de dispersió. La diferència encara seria més gran prenent com a paràmetre $\varepsilon/\overline{y}$: donaria 5,6226 per als sudokus i 0,2649 per a les targetes.

6. Exemple: venda lliure

Un dels paràmetres importants per a la valoració d'una farmàcia (per exemple, a l'hora d'una compravenda) és la proporció que presenta de venda lliure (des d'ara, VLL) en el total d'ingressos. Els preus de la VLL no estan regulats (d'aquí ve el qualificatiu): cada farmàcia els decideix i pot modificar—los en tot moment, encara que hi ha una certa uniformitat, sobretot entre farmàcies pròximes.

Els supermercats tenen productes que també es venen a les farmàcies (parafarmàcia, per exemple) i poden oferir—los a preus inferiors perquè compren més a l'engròs. Fins al punt que algunes farmàcies, després d'indagar en els supermercats pròxims, acaben venent al mateix preu o amb pèrdues alguns d'aquests productes per no deixar escapar clients.

Hi ha tres emplaçaments típics per a una farmàcia: a prop d'un centre d'atenció primària (CAP), cèntrica o en una barriada. Les farmàcies pròximes a un CAP despatxen moltes receptes, com és lògic; les cèntriques tenen una alta VLL, fins i tot superior a les receptes del Sistema Català de Salut (des d'ara, SCS); finalment, les farmàcies de barriada no destaquen per la VLL.

El nostre objectiu és l'estudi de la relació entre l'import total que paga el SCS i el de la VLL. Ens cenyirem a les dades mensuals del 2021 de la mateixa farmàcia de l'exemple 5.

Tenim, doncs, dues sèries de *dades*, cada una amb 12 components mensuals: *l'abonament del SCS* i *l'import de la VLL*. La taula 4 ens dona les dades de la distribució bidimensional que estudiarem.

Taula 4. Dades del SCS i de la VLL.

mes	gener	febrer	març	abril	maig	juny
SCS (x)	37.763,17	34.490,70	42.904,53	36.429,43	40.741,81	44.417,78
VLL (y)	8.202,35	8.092,48	9.443,87	8.926,15	9.156,85	9.620,87

mes	juliol	agost	setembre	octubre	novembre	desembre
SCS (x)	39.931,27	37.189,63	41.038,80	39.765,75	42.393,33	41.965,91
VLL (y)	9.797,36	7.252,55	9.282,99	9.825,66	8.400,31	12.010,76

Encara que els imports del SCS tenen cinc xifres enteres i dos decimals, mentre que la majoria dels de la VLL són de quatre xifres enteres i els rangs de les variables són diferents, en el diagrama de dispersió l'escala de l'eix de les y serà la mateixa que a l'eix de les x. Com a la figura 3, l'artifici per no desaprofitar espai ha consistit a situar l'origen dels eixos en el punt (34,7).

Passem ara a la part analítica. Volem veure si hi ha una bona relació entre les dues sèries de dades i quina és la recta de regressió de y sobre x.

Els resultats són els següents:

- El *coeficient de correlació r* entre dues variables, que sempre està a l'interval $[-1,1]$, dona aquí $r = 0{,}5396$. Per tant, la VLL està només relativament relacionada amb els pagaments del SCS.
- El *centre de gravetat* de la distribució és el punt $(\overline{x},\overline{y}) =$(39.914,34, 9.167,68).
- L'equació de la *recta de regressió de y sobre x* és $y = 0{,}2186x + 442{,}6184$.
- L'*error quadràtic* d'aquesta recta és $\varepsilon = 3{,}326{,}34$, l'*error quadràtic per capita* dona $\varepsilon_0 = 277{,}19$ de mitjana per a cada mes i l'*error normalitzat* és $\varepsilon_0^* = 0{,}0302$. Aquest últim està comprès entre el de les targetes (0,0221) i el dels sudokus (0,0527).

La figura 4 dona el *diagrama de dispersió* de l'experiment. La recta de regressió creix lentament i es distancia de la meitat dels 12 punts (x,y). També hi ha marcada amb traç discontinu la poligonal que va unint successivament, seguint l'ordre de les x creixents, els punts de la distribució (el número al costat de cada punt indica el mes de l'any). Pensem que la singular posició del punt 12 s'explica perquè al desembre la gent incrementa la compra de productes de venda lliure com a regals típics de l'època. En canvi, la baixa proporció de venda lliure dels mesos d'agost i novembre és deguda, respectivament, a les vacances d'estiu i al *black friday*, que no es practica a les farmàcies.

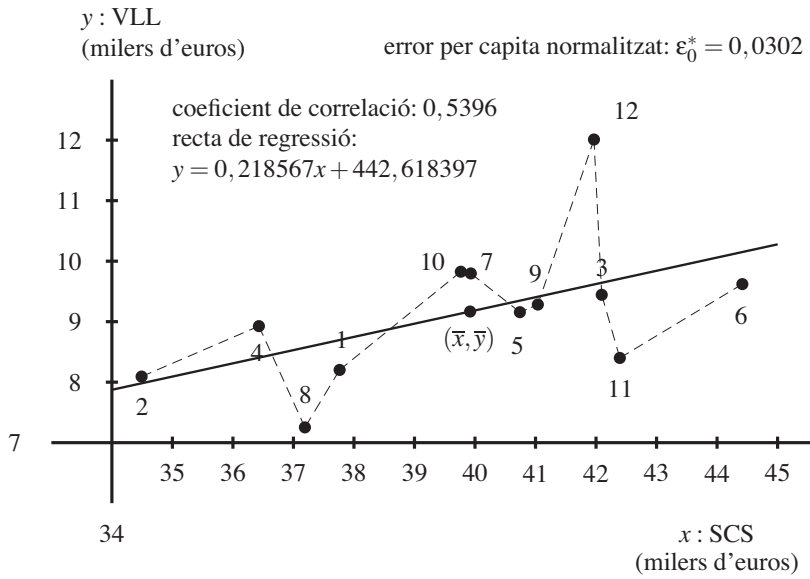

Figura 4. Diagrama de dispersió de la VLL vs. el SCS.

7. Exemple: futbol de la Primera Divisió espanyola

La Primera Divisió del Campionat de Lliga d'Espanya la formen 20 equips. Durant la primera volta, cada jornada es juguen, doncs, 10 partits, cada un al camp d'un dels dos contrincants. A la segona volta fan el mateix però al camp de l'altre. Per tant, en total hi ha 38 jornades, 19 a la primera volta i 19 més a la segona. Quan un equip guanya a un altre, el primer rep 3 punts i el segon 0. Si empaten, reben 1 punt cada un.

Els quatre primers classificats jugaran la Champions League la temporada següent; el cinquè i el sisè jugaran l'Europa League. A l'altre extrem, els tres darrers classificats baixen a la Segona Divisió la temporada següent; els substituiran els tres primers classificats d'aquesta divisió (els dos primers directament, mentre que el tercer surt d'eliminatòries entre els quatre següents, els que ocupen les posicions tercera, quarta, cinquena i sisena).

Com és lògic, els resultats de cada equip a la segona volta solen ser diferents dels de la primera volta. L'objectiu d'aquest estudi és analitzar els comportaments a la primera i la segona volta. Prendrem les dades de la temporada 2020−2021 i ens fixarem només en els equips més forts d'una banda i en els més dèbils de l'altra. Els primers estan interessats a quedar en els «llocs europeus»; els segons, a fugir dels tres llocs més baixos de la classificació final.

De millor a pitjor, en acabar la primera volta els equips més forts van ser els següents: Atlético de Madrid, Real Madrid, Barcelona, Sevilla, Villarreal, Real Sociedad, Granada i Betis. El Villarreal tenia garantit l'accés a la Champions League perquè havia guanyat l'Europa League la temporada anterior 2019−2020 (norma europea). També de millor a pitjor, a meitat de temporada els equips més dèbils van ser els següents: Getafe, Athletic de Bilbao, Valencia, Eibar, Valladolid, Alavés, Elche, Osasuna i Huesca. En tots dos casos, al final de la competició hi va haver alguns canvis d'ordre.

La doble taula 5 ens dona la puntuació de cada equip en acabar la primera volta (jornada 19) i al final de la temporada (jornada 38). Això ens donarà les dades *x,y* de la distribució bidimensional que estudiarem, primer per als equips forts i després per als dèbils. Per raons d'espai, hem abreujat els noms d'alguns equips.

Taula 5. Puntuacions a mig campionat (jornada 19) i al final (jornada 38).

equip	AMadrid	RMadrid	Barça	Sevilla	Villa	RSoc	Grana	Betis
J 19 (*x*)	48	40	37	36	33	30	28	26
J 38 (*y*)	86	84	79	77	58	62	46	61

equip	Geta	AthB	Valen	Eibar	Valla	Alav	Elche	Osasu	Huesca
J 19 (*x*)	23	21	20	19	19	18	17	16	12
J 38 (*y*)	38	46	43	30	31	38	36	44	34

Amb aquestes dades ja es veuen diferències de comportament notables entre els equips. A la meitat del campionat, els cinc primers tenien plaça per a la Champions League, i el sisè i el setè, per a l'Europa League, però no hi havia res definitiu. Quedaven molts punts per disputar. El Betis, per exemple, podia aspirar a anar a l'Europa League.

A la banda baixa, també tot podia canviar. De moment estaven en risc Elche, Osasuna i Huesca, però també perillaven Alavés, Valladolid, Eibar, Valencia... Degut a les diferències entre el rang de variació de *x* i el de *y* en tots dos casos, l'escala per a la *y* serà diferent de la de la *x* en els diagrames corresponents inclosos més avall.

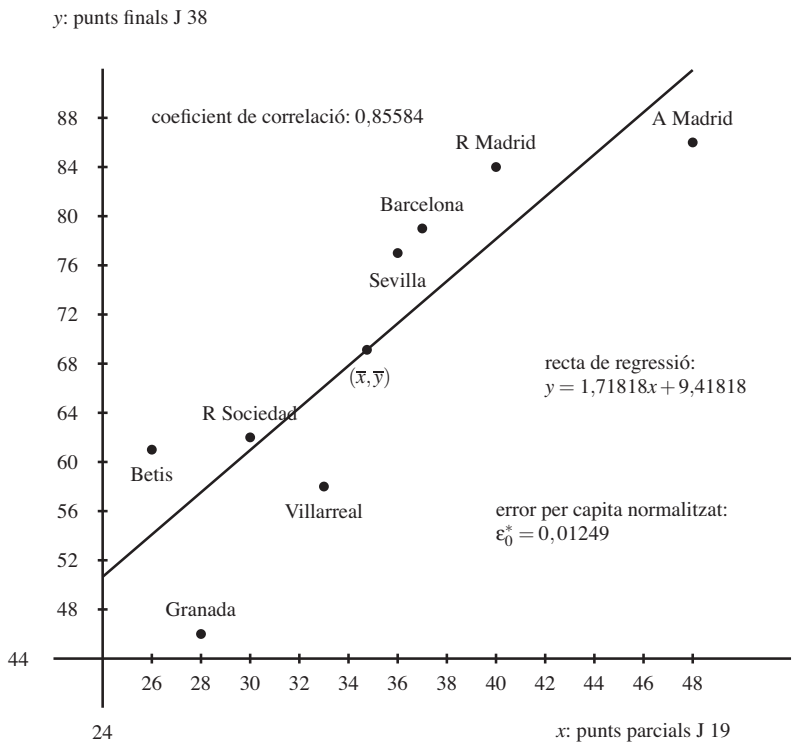

Figura 5. Diagrama de dispersió dels equips amb aspiracions europees.

Passem ara a la part analítica. Hem estudiat en els dos casos la distribució bidimensional *x,y*. En primer lloc considerem el grup dels equips forts. Els resultats són els següents:

- El *coeficient de correlació r* entre dues variables, que sempre està a l'interval $[-1,1]$, dona aquí $r = 0{,}85584$. Significa que hi ha molta correlació entre *y* i *x*. És a dir, els resultats de la segona volta estan, en general, en consonància amb els de la primera.
- El *centre de gravetat* de la distribució és el punt $(\overline{x},\overline{y}) = (34{,}75,\ 69{,}13)$.
- La *recta de regressió de y sobre x* dóna $y = 1{,}71818x + 9{,}41818$. Es pot veure representada en el primer diagrama de dispersió (figura 5).
- L'*error quadràtic* d'aquesta recta és $\varepsilon = 6{,}90909$, l'*error quadràtic per capita* dona $\varepsilon_0 = 0{,}86364$ de mitjana per a cada equip i l'*error normalitzat* és $\varepsilon_0^* = 0{,}01249$. Aquest últim és força baix en comparació amb exemples previs.

Passem al grup d'equips dèbils. La figura 6 dona el seu diagrama de dispersió.

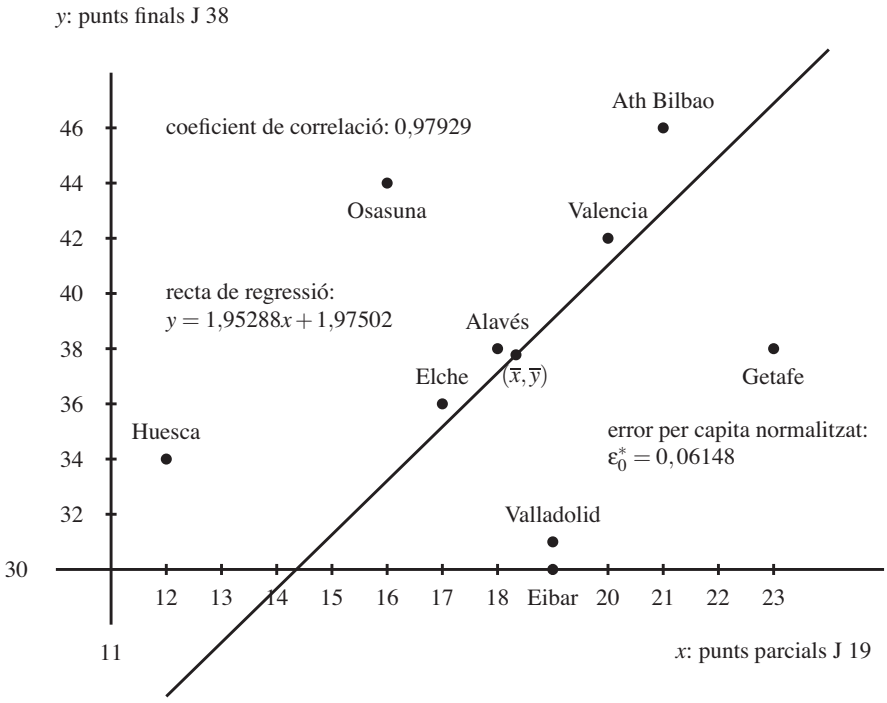

Figura 6. Diagrama de dispersió dels equips en risc de descens.

Els resultats aquí són els següents:

- El *coeficient de correlació r* entre les dues variables dona aquí $r = 0{,}97929$. Significa que encara hi ha més correlació entre *y* i *x*, és a dir, en general, entre els resultats de les dues voltes.
- El *centre de gravetat* de la distribució és el punt $(\overline{x},\overline{y}) = (18{,}33,\ 37{,}78)$.
- La *recta de regressió de y sobre x* dona $y = 1{,}95288x + 1{,}97502$. Es pot veure representada en el segon diagrama de dispersió (figura 6).

- L'*error quadràtic* d'aquesta recta és $\varepsilon = 20{,}90301$, l'*error quadràtic per capita* dona $\varepsilon_0 = 2{,}32256$ de mitjana per a cada equip i l'*error normalitzat* surt $\varepsilon_0^* = 0{,}06148$. Aquest últim també és baix en comparació amb alguns exemples previs (no pas tots).

Comparem ara els dos grups d'equips. El *coeficient de correlació*, curiosament, és més alt per als equips dèbils. Això pot ser degut que, en general, els equips forts tenen motivacions i plantilles superiors per «mantenir i superar el ritme» durant la segona volta, mentre que als dèbils els passa més aviat el contrari. Tot i així, els dos *centres de gravetat* comparteixen la propietat que la segona component (mitjana de puntuacions finals) és el doble de la primera (mitjana de la primera volta).

En relació amb la *recta de regressió*, cinc equips forts estan per sobre, i això significa que s'han esforçat més a la segona volta, mentre que els altres tres han perdut pistonada. Analitzats individualment, l'Atlético de Madrid no ha fet gaire els deures i ha acabat campió molt justet. Real Madrid, Barcelona i Sevilla han perseverat i han mantingut el seu ordre particular. Els quatre s'han classificat per a la Champions League. La Real Sociedad i el Betis també han apretat i s'han classificat per a l'Europa League, superant el Villarreal, que s'ha relaxat perquè tenia la Champions garantida per norma de la UEFA. El Granada és el que menys esforç ha fet i ha quedat desbancat pel Betis.

Entre els dèbils, sis s'han esforçat i els altres tres no. Aquests són el Getafe, que s'ha vist superat per l'Athletic i el Valencia, i sobretot el Valladolid i l'Eibar, que cauen a Segona Divisió juntament amb l'Huesca, que ha fet un esforç gran però insuficient perquè venia de molt avall. Dels altres sis, a banda dels ja comentats, l'Alavés ha igualat el Getafe, mentre que Elche i Osasuna han sortit del pou.

Finalment, els *errors quadràtic* i *per capita* dels equips dèbils són el triple dels corresponents als forts, mentre que l'*error normalitzat* és sis vegades superior. Això indica que més de la meitat dels equips dèbils s'han desviat molt de la recta de regressió, mentre que entre els forts només el Granada ha quedat clarament lluny de la recta.

Com a comentari final, destaca la regularitat i l'estabilitat a la part alta de la taula (equips forts) i la manca d'aquestes propietats a la part baixa (equips dèbils), amb el canvi (dramàtic) dels dos equips que al final acompanyen l'Huesca a Segona Divisió. Hi ha trets comuns, com ara que en els dos centres de gravetat la segona coordenada dobla la primera: això és bo per als forts i dolent per als dèbils, que haurien hagut de millorar durant la segona volta. El coeficient de correlació és alt en ambdós casos, però té significat positiu per als forts i negatiu per als dèbils. Els tres errors quadràtics també mostren que els equips forts han estat més consistents, mentre que els dèbils s'allunyen molt més, a la baixa, d'un comportament regular.

8. Exemple: qualificacions acadèmiques

A l'Escola Superior d'Enginyeries Industrial, Aeroespacial i Audiovisual, ubicada en el Campus de Terrassa de la UPC, hi ha dos graus d'Aeronàutica: el grau en Enginyeria en Tècniques Aeroespacials (GRETA) i el grau en Enginyeria en Vehicles Aeroespacials (GREVA). Tant en l'un com en l'altre hi ha tres assignatures de matemàtiques en els dos quadrimestres inicials:

Àlgebra i Càlcul I durant el primer (tardor) i Càlcul II durant el segon (primavera). No hi ha establerts prerequisits entre elles.

L'objecte d'aquest estudi és la possible relació entre les notes finals de les tres assignatures. Ens centrarem en el GREVA i, més concretament, en les notes obtingudes pels estudiants durant els quadrimestres de la tardor de 2016 i de la primavera de 2017. Els dos professors que van impartir aquestes tres assignatures tenien criteris de qualificació semblants.

A les actes finals consten 62 estudiants matriculats tant d'Àlgebra com de Càlcul I, quatre dels quals no es van presentar a Àlgebra però sí a Càlcul I, mentre que dos d'aquests quatre i un tercer no surten a l'acta de Càlcul II, que, per tant, tenia 59 matriculats. D'entrada, assignarem un 0 en una assignatura a cada estudiant no matriculat o no presentat (d'ara endavant, NP, genèricament). Si convé, modificarem aquest criteri durant l'estudi.

En la primera part de l'estudi mirarem de relacionar les assignatures per parelles i. per tant, hi haurà tres parts: Àlgebra vs. Càlcul I, Àlgebra vs. Càlcul II, i Càlcul I vs. Càlcul II. En principi, a la vista dels tres temaris no es pot dir que l'Àlgebra i el Càlcul I tinguin gaires punts de contacte, però sí que el Càlcul II fa ús de tècniques estudiades a l'Àlgebra i sobretot, com és lògic, a Càlcul I. Els resultats obtinguts pels estudiants ens poden donar una idea de la seva capacitat (grupal) d'assimilació de cada matèria. En una segona part analitzarem la possible relació *conjunta* de Càlcul II amb Àlgebra i Càlcul I.

Passem ara a la part analítica. Hem estudiat en els tres primers casos distribucions bidimensionals.

8.1. Àlgebra i Càlcul I

En aquest cas estudiarem la distribució bidimensional *x,y*, on *x* representa les notes d'Àlgebra i *y* les de Càlcul I. D'entrada, com ja ha quedat dit, seguirem el criteri d'assignar un 0 als NP, que afectarà quatre estudiants que no es van presentar a Àlgebra. En canvi, els 62 matriculats es van presentar tots a Càlcul I. Els resultats són els següents:

- El *coeficient de correlació* $r \in [-1,1]$ entre les dues variables dona aquí $r = 0{,}0020$. Significa que no hi ha pràcticament cap correlació entre *y* i *x*.
- El *centre de gravetat* de la distribució és el punt $(\overline{x},\overline{y}) =$(6,6387, 5,7081). Sembla que l'Àlgebra resulta una mica més fàcil que el Càlcul I.
- La *recta de regressió de y sobre x* dona $y = 0{,}0013x + 5{,}6992$. Es pot veure representada per una línia fina en el diagrama de dispersió (figura 7); el centre de gravetat és el situat més a l'esquerra dels dos punts destacats per un cercle més gran. És gairebé horitzontal i molt poc representativa de la distribució, com augurava el valor de *r*.
- L'*error quadràtic* d'aquesta recta és $\varepsilon = 12{,}5948$, l'*error quadràtic per capita* dona $\varepsilon_0 = 0{,}2031$ de mitjana per a cada estudiant i l'*error normalitzat* és $\varepsilon_0^* = 0{,}0356$. Aquest últim és relativament baix en comparació amb altres exemples previs on hem tractat temes diversos.

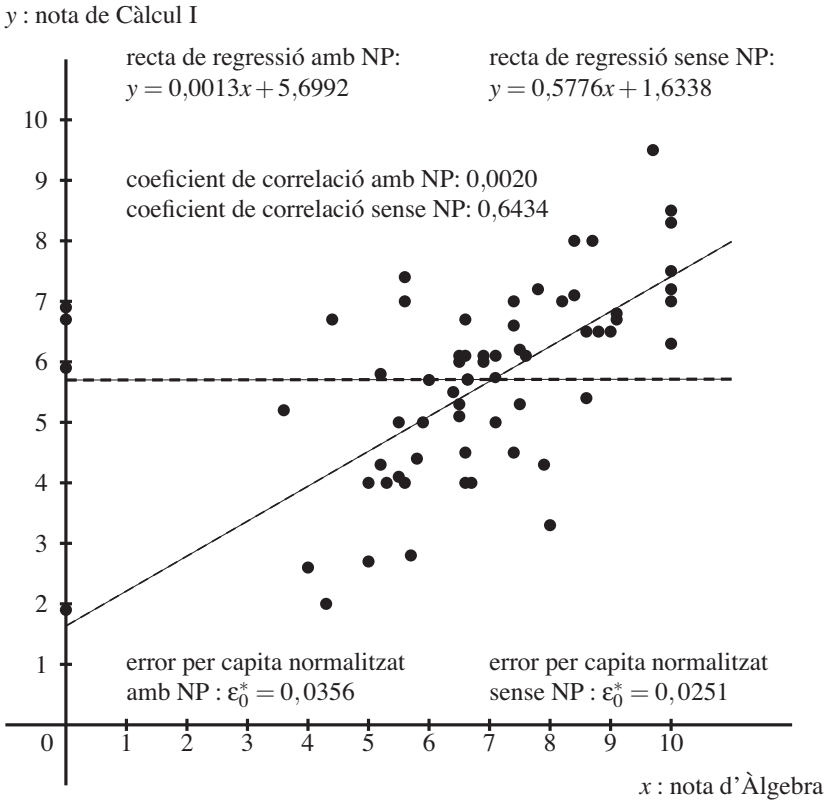

Figura 7. Diagrama de dispersió d'Àlgebra i Càlcul I.

Tanmateix, fixem—nos en els quatre NP d'Àlgebra. Els punts que els representen en el diagrama estan a l'eix *y*. Per veure quina influència té haver—los assignat un 0, hem repetit l'estudi eliminant—los de la distribució bidimensional. Llavors, el nombre de dades s'ha reduït a 58 i els resultats són els següents:

- El *coeficient de correlació* passa a ser $r = 0{,}6434$, que ja indica una correlació més positiva entre *y* i *x*.
- El *centre de gravetat* de la distribució és ara el punt $(\overline{x},\overline{y}) = (7{,}0966,\ 5{,}7328)$, que, lògicament, s'ha mogut cap a la dreta i lleument cap amunt perquè els quatre exclosos tenien una nota mitjana de 5,35 de Càlcul I. La nota mitjana d'Àlgebra arriba ara al notable, mentre que la de Càlcul I puja menys de tres centèsimes.
- La *recta de regressió de y sobre x* dona $y = 0{,}5776x + 1{,}6338$. Ara és clarament creixent i força adequada a l'aspecte del diagrama de punts. Es pot veure representada també en la figura 7 i el seu centre de gravetat és molt pròxim a la intersecció amb la recta anterior.
- L'*error quadràtic* d'aquesta recta és $\varepsilon = 8{,}9367$, l'*error quadràtic per capita* dona $\varepsilon_0 = 0{,}1441$ de mitjana per a cada estudiant i l'*error normalitzat* és $\varepsilon_0^* = 0{,}0251$. Tots tres són menors que els de l'anàlisi anterior, i l'últim és relativament baix en comparació amb altres exercicis previs que hem dut a terme sobre temes diversos.

La conclusió sembla clara: assignar 0 als NP desvirtua fortament la descripció numèrica de la distribució bidimensional dels punts restants. En comptes de donar un altre valor numèric als NP, serà més encertat excloure'ls en els casos següents.

8.2. Àlgebra i Càlcul II

En aquest cas estudiarem la distribució bidimensional *x,y*, on *x* representa les notes d'Àlgebra i *y* les de Càlcul II. Vista l'experiència del cas 8.1, suprimirem els sis estudiants que no es van presentar a Àlgebra, a Càlcul II o a cap de les dues. Quedaran, doncs, 56 estudiants presentats a les dues assignatures. Els resultats són els següents:

- El *coeficient de correlació* $r \in [-1,1]$ entre les dues variables dona aquí $r = 0{,}6182$. Novament indica indica una correlació lleugerament positiva entre *y* i *x*.
- El *centre de gravetat* de la distribució és el punt $(\overline{x},\overline{y}) = (7{,}196429,\ 5{,}683929)$. La mitjana de Càlcul II és un pèl menor que la de Càlcul I trobada en el cas 8.1.
- La *recta de regressió de y sobre x* dona $y = 0{,}5679x + 1{,}5971$. És similar a la del cas 8.1 i es pot veure representada en el diagrama de dispersió (figura 8), igual que el centre de gravetat, indicat per un cercle més gran.
- L'*error quadràtic* d'aquesta recta és $\varepsilon = 8{,}8961$, l'*error quadràtic per capita* dona $\varepsilon_0 = 0{,}1435$ de mitjana per a cada estudiant i l'*error normalitzat* és $\varepsilon_0^* = 0{,}0252$. Tots tres són molt semblants als del cas 8.1, però l'últim és relativament baix en comparació amb altres exercicis previs que hem dut a terme sobre temes diversos.

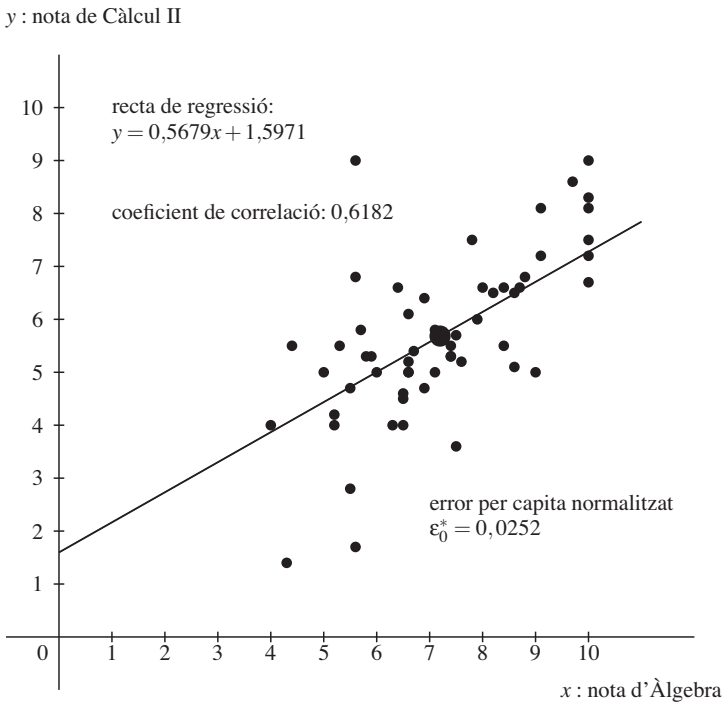

Figura 8. Diagrama de dispersió d'Àlgebra i Càlcul II.

Comparem aquests resultats amb els del cas anterior sense NP. Les relacions entre assignatures són de tipus diferent: Àlgebra i Càlcul I són *paral·leles*, és a dir, s'imparteixen simultàniament, mentre que Àlgebra i Càlcul II son *seqüencials* en el temps, amb el que això implica.

Tanmateix, els paràmetres són molt semblants en els dos casos. Així, per exemple, el coeficient de correlació és molt similar. La mitjana aritmètica de l'Àlgebra ha pujat en el segon cas perquè s'han exclòs dos estudiants més, mentre que les dels dos Càlculs són idèntiques. El pendent de la recta de regressió també és gairebé el mateix, i el mateix passa amb els tres errors. Sí que és cert que la distribució dels punts en aquest cas sembla més compacta o acumulada a l'entorn de la recta de regressió que en el primer cas.

Abans de procedir a l'estudi cas per cas no esperàvem trobar tanta simil·litud.

8.3. Càlcul I i Càlcul II

Finalment, estudiarem la distribució bidimensional *x,y*, on *x* representa les notes de Càlcul I i *y* les de Càlcul II. Seguirem suprimint els estudiants que no es van presentar a Càlcul II (tots 62 es van presentar a Càlcul I). Quedaran, doncs, 59 estudiants presentats a les dues assignatures.

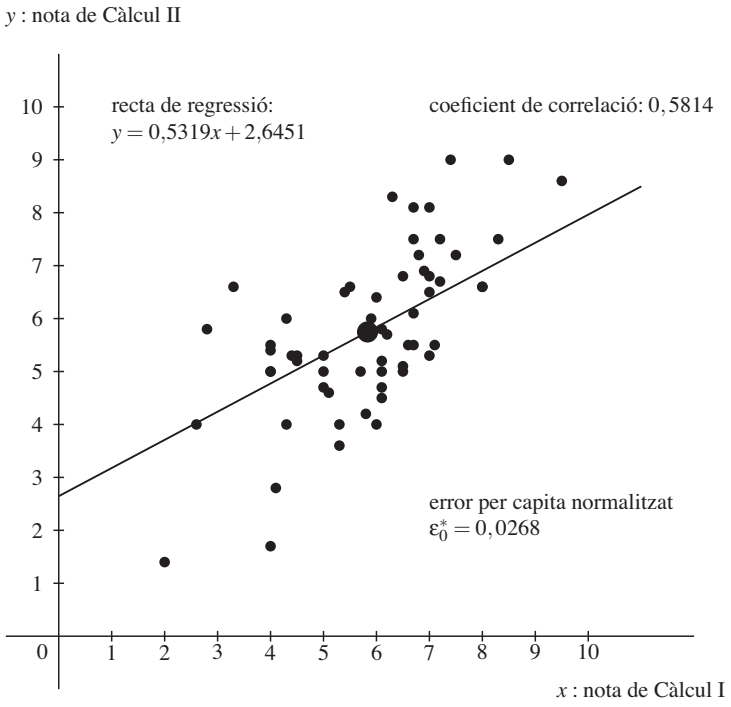

Figura 9. Diagrama de dispersió de Càlcul I i Càlcul II.

Novament, la relació entre les dues assignatures és *seqüencial*, però a priori cal esperar que la segona estigui relacionada més estretament amb Càlcul I que no pas amb Àlgebra, simplement pels continguts dels programes respectius. Els resultats són els següents:

- El *coeficient de correlació* $r \in [-1,1]$ entre les dues variables dona aquí $r = 0{,}5814$. Indica una correlació positiva dèbil entre *y* i *x*.
- El *centre de gravetat* de la distribució és el punt $(\overline{x},\overline{y})$ =(5,832203, 5,747458).
- La *recta de regressió de y sobre x* dona $y = 0{,}5319x + 2{,}6451$. Es pot veure representada en el diagrama de dispersió (figura 9), igual que el centre de gravetat, marcat amb un cercle de doble gruix.
- L'*error quadràtic* d'aquesta recta és $\varepsilon = 9{,}0928$, l'*error quadràtic per capita* dona $\varepsilon_0 = 0{,}1541$ de mitjana per a cada estudiant i l'*error normalitzat* és $\varepsilon_0^* = 0{,}0268$.

Comparem els resultats d'aquest cas amb els dels casos 8.1 (sense NP) i 8.2. La taula 6 recull els paràmetres.

Taula 6. Comparativa d'assignatures.

cas	correlació r	recta de regressió m	centre de gravetat $(\overline{x},\overline{y})$	error quadràtic ε	error normalitzat ε_0^*
1 A vs. C1	0,6434	0,5776	(7,10, 5,73)	8,9367	0,0251
2 A vs. C2	0,6182	0,5679	(7,20, 5,68)	8,8961	0,0252
3 C1 vs. C2	0,5814	0,5319	(5,83, 5,75)	9,0928	0,0268

Contràriament al que es podria esperar segons els temaris de les tres assignatures, la correlació dona valors pròxims però, curiosament, decreixents seguint l'ordre dels casos. La recta de regressió sempre és creixent, però cada vegada amb menys pendent, i el més baix és el de C1 vs C2. Les coordenades del centre de gravetat presenten petites diferències, possiblement degudes als diferents conjunts de NP exclosos en cada cas.

En tot cas, la mitjana de l'Àlgebra, que supera el 7, és molt superior a les dels dos Càlculs, que es mantenen més a prop del 6 que del 5. En els diagrames es nota més dispersió entre A i C1, i menys entre A i C2 i entre C1 i C2, que donen diagrames més «compactes» al voltant de la recta de regressió. Finalment, l'error quadràtic aquí és comparable perquè les dades de les sèries estan totes a l'interval $[0{,}10]$ i les quantitats de dades són homologables: 58, 56 i 59, respectivament. Aquest error és una mica més alt comparant els càlculs i més baix comparant A i C1 o A i C2. L'error normalitzat també augmenta, gairebé dues centèsimes, en comparar els dos càlculs.

8.4. Càlcul II vs Àlgebra i Càlcul I

Aquí ens proposem analitzar la correlació global o *conjunta* de Càlcul II (variable *Z*) respecte a les altres dues assignatures (variables *X* i *Y*, respectivament). Caldrà fer ús de tècniques una mica més complicades que generalitzen les del cas bidimensional, estudiat repetidament en les seccions prèvies.

La mostra consta dels 56 estudiants que es van presentar a les tres assignatures. Recordem que Càlcul II és seqüencial respecte a Àlgebra i Càlcul I. Els càlculs estan fets optimitzant les aproximacions i els resultats són els següents:

- El *coeficient de correlació (global)* de Z respecte a X i Y dona $R_Z(X,Y) = 0{,}7508$. Aquest coeficient sempre està a l'interval $[0,1]$.
- El *centre de gravetat* de la distribució és el punt $(\overline{X},\overline{Y},\overline{Z}) =$(7,1964, 5,7964, 5,6839).
- El *pla de regressió de Z sobre X i Y* dona:

$$Z - 5{,}6839 = 0{,}3753(X - 7{,}1964) + 0{,}4206(Y - 5{,}7964)$$

o, en una forma equivalent,

$$Z = 0{,}3753X + 0{,}4206Y + 0{,}5451.$$

- El fet que la suma dels valors Z^* que dona la recta de regressió coincideixi amb la suma dels valors Z de les dades ens sembla merament anecdòtic.
- Evidentment, aquest pla passa pel centre de gravetat.
- L'*error quadràtic per capita* d'aquest pla és $\varepsilon_Z^0(X,Y) = 1{,}0798$.

Comparant amb els resultats dels estudis bidimensionals resumits en la taula 6 al final de la secció 8.3, observem que el coeficient de correlació global és notablement millor que els tres coeficients de correlació binària. Aquest és un resultat esperat, ja que l'ajustament millora en considerar dues variables independents en comptes d'una de sola. La variació dels coeficients deriva del fet que X i Y estan relacionades entre elles. Finalment, C1 té un coeficient més gran que A, cosa que també sembla raonable vistos els programes de les tres assignatures. També observem que $X = 10$ i $Y = 10$ donen $Z = 8{,}5041$.

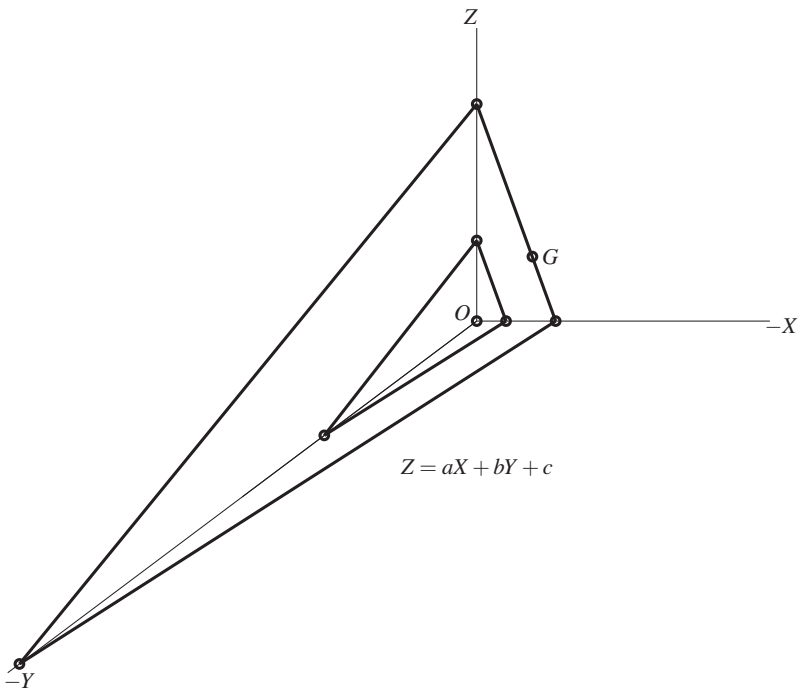

Figura 10. Pla de regressió de Càlcul II respecte a Àlgebra i Càlcul I.

Les coordenades del centre de gravetat *G* són molt semblants a les dels centres de gravetat binaris; de fet, cadascuna d'elles està compresa entre les dels dos binaris corresponents i les diferències poden ser degudes a les petites variacions en el nombre d'estudiants de cada estudi. Segons l'equació del pla de regressió, *Z* creix quan una de les variables *X,Y* creix i l'altra roman constant o totes dues creixen.

En aquest cas tridimensional, no sembla adequat donar la representació gràfica dels 56 punts, però podem visualitzar, a la figura 10, el pla de regressió en una doble imatge inclusiva. Finalment, l'error per capita no és comparable amb els bidimensionals perquè no es defineix com ells; tanmateix, prenent $\sqrt{56}\ \varepsilon_Z^0(X,Y) = 7{,}4833$, obtenim un paràmetre homologable que resulta força menor que tots els corresponents errors quadràtics bidimensionals. Resumint, l'estudi tridimensional sembla un bon complement dels tres estudis bidimensionals.

9. Exemple: temperatura de xafogor

La *temperatura de xafogor* és un concepte que pretén mesurar quantitativament la «sensació tèrmica» que experimenta el cos humà a causa de la combinació de la temperatura ambiental (temperatura de l'aire) i la humitat relativa. La primera notícia sobre aquesta idea la va popularitzar, almenys a Catalunya, Alfred Rodríguez Picó a TV3, on va exercir com a meteoròleg des del 1984 fins al 2002.

En l'actualitat, els que descriuen el temps solen parlar de la (temperatura de) *xafogor* quan mereix un comentari perquè és força més elevada que la temperatura ambiental. De fet, el concepte, degut a R. G. Steadman, va néixer cap al 1979. Només cal anar a la Viquipèdia, per exemple, per obtenir una informació més detallada. En particular, hom hi troba una fórmula (que ha estat relativament criticada) que relaciona la temperatura de xafogor en termes de la temperatura ambiental i la pressió del vapor d'aigua de l'atmosfera.

L'objecte del nostre estudi no fa referència a aquesta fórmula. Una taula de dades publicada el 2017 a www.marenostrum.org/meteorologia/xafogor donava 291 valors simultanis de la temperatura ambiental, la humitat relativa i la temperatura de xafogor —-suposem que calculada amb la fórmula esmentada més amunt—-, amb la particularitat de dividir la graella en cinc blocs de diferents tonalitats (vegeu la figura 11).[5]

Hem volgut comparar la correlació de la xafogor respecte a les altres dues variables en cada un dels blocs, numerats de l'1 al 4 de dalt a baix, ja que hem reduït els dos últims a un de sol perquè considerem el cinquè una mica marginal. És a dir, analitzarem cada bloc com una distribució tridimensional on les variables són: *X* = temperatura ambiental, *Y* = humitat relativa i *Z* = temperatura de xafogor. Obtindrem els paràmetres bàsics de cada distribució i els confrontarem. El web de Mare Nostrum assigna les interpretacions següents als colors: 1 = perill extrem; 2 = perill; 3 = precaució; 4 = límit de confort; 5 = confort tèrmic.

5. Ara s'hi pot trobar una taula actualitzada l'1 de gener de 2023 que és idèntica a la presentada aquí.

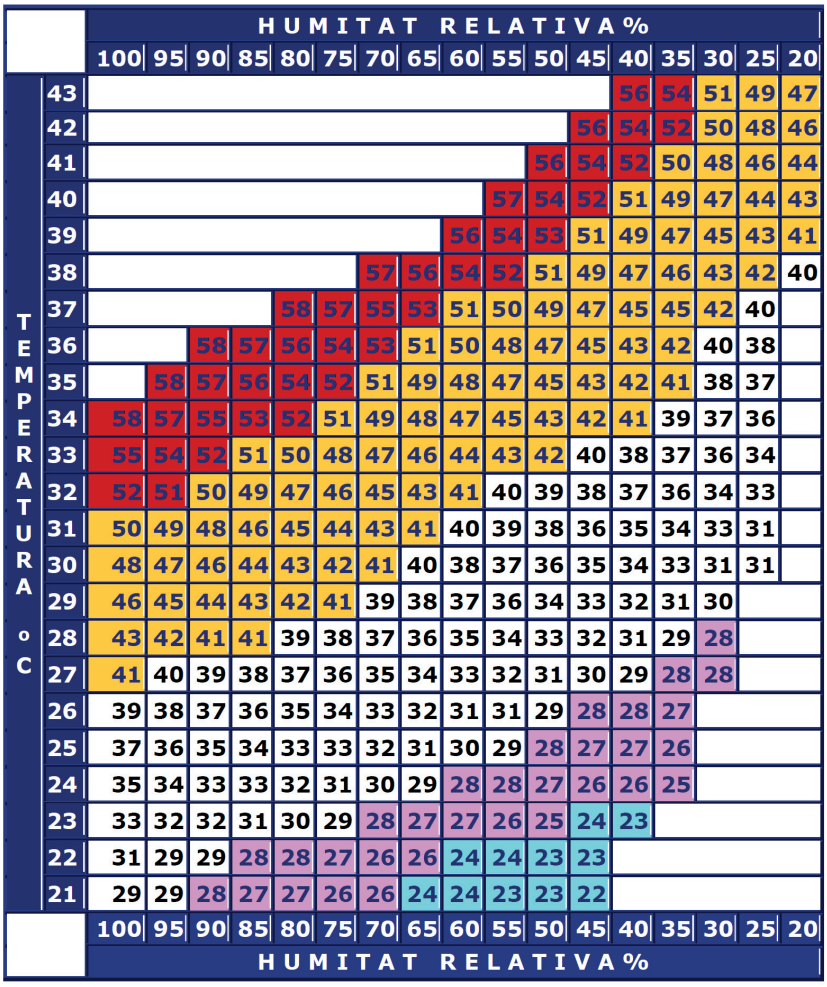

Figura 11. Dades empíriques sobre temperatura, humitat i xafogor. Taula extreta de www.marenostrum.org/meteorologia/xafogor.

Els resultats dels paràmetres s'han arrodonit a un nombre suficient de decimals. Els hem reunit en la taula 7 per poder comparar—los còmodament.

Taula 7. Comparativa de xafogor per blocs.

xafogor	bloc 1	bloc 2	bloc 3	blocs 4 i 5
significat	perill extrem	perill	precaució	confort
dades	42	98	109	42
a	1,875623	1,695420	1,570360	0,862790
b	0,305099	0,267812	0,194376	0,109614
c	−36,284824	−28,585348	−21,406818	0,000000
error/capita	0,748058	1,147926	0,797269	0,801224
correlació	0,897887	0,941676	0,963805	0,871475
centre de gravetat	(37, 70, 55)	(35, 58, 46)	(28, 60, 34)	(23, 55, 26)

Els comentaris són els següents:

- En els quatre blocs els coeficients *a,b* són positius. Això vol dir que, segons el pla de regressió, qualsevol increment de la temperatura ambient i/o de la humitat relativa provoca un augment de la xafogor, i és molt més acusada la influència de la temperatura ambiental ja que $a >> b$. Tanmateix, aquest doble efecte decreix al llarg dels quatre blocs, de manera que és en el bloc marginal on és menys rellevant.
- El coeficient *c* fa una funció de correcció. En valor absolut decreix, i arriba a desaparèixer en el quart bloc.
- L'error per capita és baix segons el rang de la xafogor ($[22,58]$) i no varia gaire llevat del segon bloc, on és molt més alt que en els altres tres (aproximadament entre un 42 % i un 54 % superior).
- El coeficient de correlació també és força alt en tots els casos, sobretot en el segon i el tercer bloc. Això fa pensar que el primer bloc i el quart no són tan representatius de la distribució global.
- Els centres de gravetat apareixen, en cada bloc, entre les dades del bloc, com és lògic, decreixen en les coordenades *X* i *Z* i tenen un comportament més irregular a la coordenada *Y*.
- És probable que el pla de regressió total, és a dir, considerant conjuntament totes les dades de la distribució, sigui una mixtura dels quatre plans per blocs. No hem considerat necessari calcular—lo.

10. Conclusions

Una vegada analitzats tots els exemples, sembla interessant remarcar la varietat d'àmbits als quals fan referència i comparar els principals resultats obtinguts. Hi ha un exemple de joc individual (sudoku), dos de significat econòmic (referits a una mateixa farmàcia), un de dedicat a la lliga de Primera Divisió en dues fases (equips forts i equips dèbils), un altre sobre qualificacions acadèmiques dividit en quatre parts, i un últim sobre un concepte meteorològic, la temperatura de xafogor, també desdoblat en quatre blocs.

Els primers queden definits per distribucions bidimensionals, mentre que els dos últims —l'estudi de tres assignatures simultàniament i els quatre casos de xafogor— corresponen a distribucions tridimensionals. Tots són exemples de casos concrets i, per tant, amb resultats en principi no extrapolables, però sembla clar que la metodologia emprada en tots els casos és perfectament aplicable a qualsevol altre cas amb dades numèriques diferents i, sobretot, a situacions que es puguin modelar en termes anàlegs.

Es posa en relleu, doncs, una reflexió sobre com l'ús d'una eina d'implementació senzilla permet una primera anàlisi que pot ser molt rellevant per descriure situacions econòmiques, empresarials, educatives, etc.; en definitiva, de la societat actual.

La taula 8 permet copsar amb un cop d'ull la informació principal obtinguda en cada exemple: el nombre de dades; el coeficient de correlació, que es mesura de manera diferent en el cas bidimensional (*r*) i en el tridimensional (*R*), com s'indica en la taula (vegeu les seccions

3.1.1 i 3.1.3); el pendent de la recta de regressió en el cas bidimensional —-idea que no s'estén al cas tridimensional perquè passem a tenir un pla de regressió—-; i, finalment, els diversos tipus d'error —-tres en el cas bidimensional, només dos en l'altre—-.[6] El paràmetre ε^{*0} en els casos tridimensionals es defineix, per analogia amb el cas bidimensional, dividint ε^0 per $\overline{Z}$, per poder comparar tots els errors normalitzats. El resultat és que tots aquests errors, llevat d'un, són inferiors a 0,1. I l'única excepció no arriba a 0,2.

Taula 8. Comparativa de tots els exemples estudiats.

ex.	tema	dades	correlació	pendent	error	error	error
					quadràtic	per capita	norm.
		n	r	m	ε	ε_0	ε_0^*
4	sudokus	108	−0,295	−0,140	8,38	0,078	0,052
5	targetes	12	0,853	12,498	1.346,20	112,180	0,022
6	venda lliure	12	0,540	0,219	3.326,34	277,19	0,030
7	equips forts	8	0,856	1,718	6,91	0,864	0,012
7	equips dèbils	9	0,979	1,953	20,90	2,323	0,062
8	A vs C1 (*)	62	0,002	0,001	12,59	0,203	0,036
8	A vs C1	58	0,643	0,578	8,94	0,144	0,025
8	A vs C2	56	0,618	0,568	8,90	0,144	0,025
8	C1 vs C2	59	0,581	0,532	9,09	0,154	0,027
		n	R			ε^0	ε^{*0}
8	A i C1 vs C2	56	0,727			1,080	0,190
9	xafogor bloc 1	42	0,898			0,748	0,014
9	xafogor bloc 2	98	0,942			1,148	0,025
9	xafogor bloc 3	109	0,964			0,797	0,023
9	xafogor bloc 4	42	0,871			0,801	0,031

És destacable la varietat del nombre de dades. També la del coeficient de correlació, que dona un valor negatiu i un altre pròxim a 0 al costat de valors molt pròxims a 1. Quelcom similar es pot dir del pendent de la recta de regressió. Pel que fa als errors, l'error quadràtic dels dos exemples de caire econòmic és elevat perquè són elevades les quantitats que defineixen la sèrie *y* de la distribució. Per això hem introduït l'error per capita, que permet fer comparacions independentment del nombre de dades. I encara més l'error normalitzat, que apareix a l'última columna i dona un resultat que permet comparar tots els exemples.

Agraïments

Volem agrair la revisió duta a terme per Marianna Bosch, que amb els seus comentaris encertats ens ha ajudat a millorar aquest treball. I també la revisió lingüística duta a terme per dos correctors anònims que han millorat força la presentació.

6. L'asterisc del sisè tema indica que l'estudi està fet assignant la qualificació 0 a tots els no presentats, criteri que es descarta en els tres casos següents. Com es pot veure, la correlació creix molt en prescindir d'aquest criteri.

Bibliografia comentada

Les referències que donem són una mica *sui generis*. Aquest article no és dels que, de vegades més aviat marginalment, connecten amb molts treballs previs, propis o d'altres autors, com és habitual en els dedicats a la «internalitat». Es basa en nocions d'estadística força conegudes i estudia situacions no considerades sovint en la literatura científica.

[1] Especial Sudokus nº 235 (2021). Madrid: Ediciones Pléyades. Per a les mateixes seccions.

[2] Jarne, G.; Pérez—Grosso, I.; Minguillón, E. (1997). *Matemáticas para la economía*. Madrid: MacGraw—Hill. Per ajudar a captar el rerefons del nostre enfocament.

[3] Legendre, A. M. (1786). «Mémoire sur la manière de distinguer les Maxima des Minima dans le Calcul des Variations». *Mémoires in Histoire de l'Académie Royale des Sciences*, pàg. 7. Per a la secció 3. Tot i que sembla que Gauss va ser el primer a descobrir el mètode dels mínims quadrats, Legendre va ser el primer que va publicar—lo.

[4] Legendre, A. M. (1819). «Méthode des moindres carrés pour trouver le milieu le plus probable entre les résultats de différentes observations». *Mémoires présentés par divers Savants à la l'Académie des Sciences de l'Institut de France*, pàgs. 149—154. Nova presentació davant d'un auditori diferent.

[5] McGuire, G.; Tugemann, B.; Civario, G. (2014). «There is No 16—Clue sudoku: Solving the sudoku minimum number of clues problem via hittings set enumeration». *Experimental Mathematics* 23 (2), pàgs. 190—217. Per a les seccions 2 i 4.

[6] Peña, D.; Romo, J. (1997). *Introducción a la Estadística para las Ciencias Sociales*. Madrid: McGraw—Hill. Per al contingut de la secció 3.

[7] Spiegel, M. R. (1969). *Estadística*. Madrid: McGraw—Hill. Un clàssic en el que ens hem basat a la secció 3.

[8] Steadman, R. G. (1979). «The assessment of Sultriness. Part I: A temperature—humidity index based on human physiology and clothing science». *Journal of Applied Meteorology* 18 (7), pàgs. 861—873. Per a la secció 9.

[9] Steadman, R. G. (1979). «The assessment of Sultriness. Part II: Effect of wind, extra radiation and barometric pressure on apparent temperature». *Journal of Applied Meteorology* 18 (7), pàgs. 874—885. Per a la secció 9. Continuació de l'article anterior.

[10] Viquipèdia, on es poden trobar molts tutorials sobre els temes concrets d'estadística emprats aquí. Per a la secció 3.

[11] www.rfef.es. Web de la Real Federación Española de Fútbol, on es poden trobar els resultats i les classificacions de la Lliga de Primera Divisió 2020—2021. Per a la secció 7.

[12] m www.marenostrum.org/meteorologia/xafogor (2017). Web on va aparèixer la taula de la figura 11. Actualitzada amb data 1 de gener de 2023. Per a la secció 9.