3. Correlació lineal i predicció

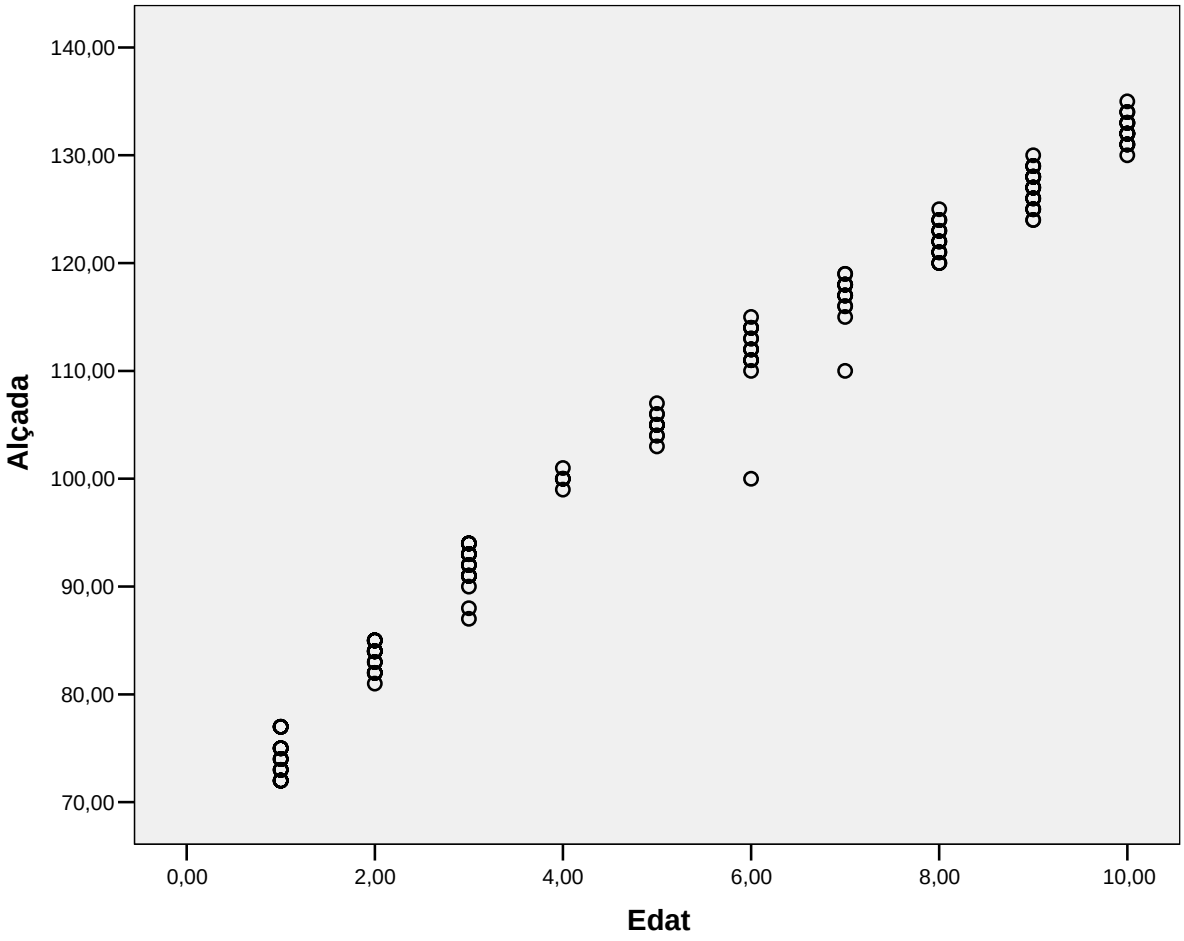
3.1. Introducció

Fins ara hem estat treballant amb una sola variable. Quan buscàvem els índexs de tendència central i de dispersió els buscàvem per a una variable cada vegada.

Però ens pot interessar estudiar conjuntament dues variables. Una de les anàlisis que podem fer és el de la *correlació lineal* entre dues variables.

Per entendre què és això de la correlació, vegem com es representaria gràficament la relació entre dues variables, "edat" i "alçada" d'una mostra de 150 nens i nenes d'entre 1 i 10 anys.

Gràfic 3.1: Diagrama de dispersió de les variables "edat" i "alçada" d'una mostra de nens i nenes d'entre 1 i 10 anys

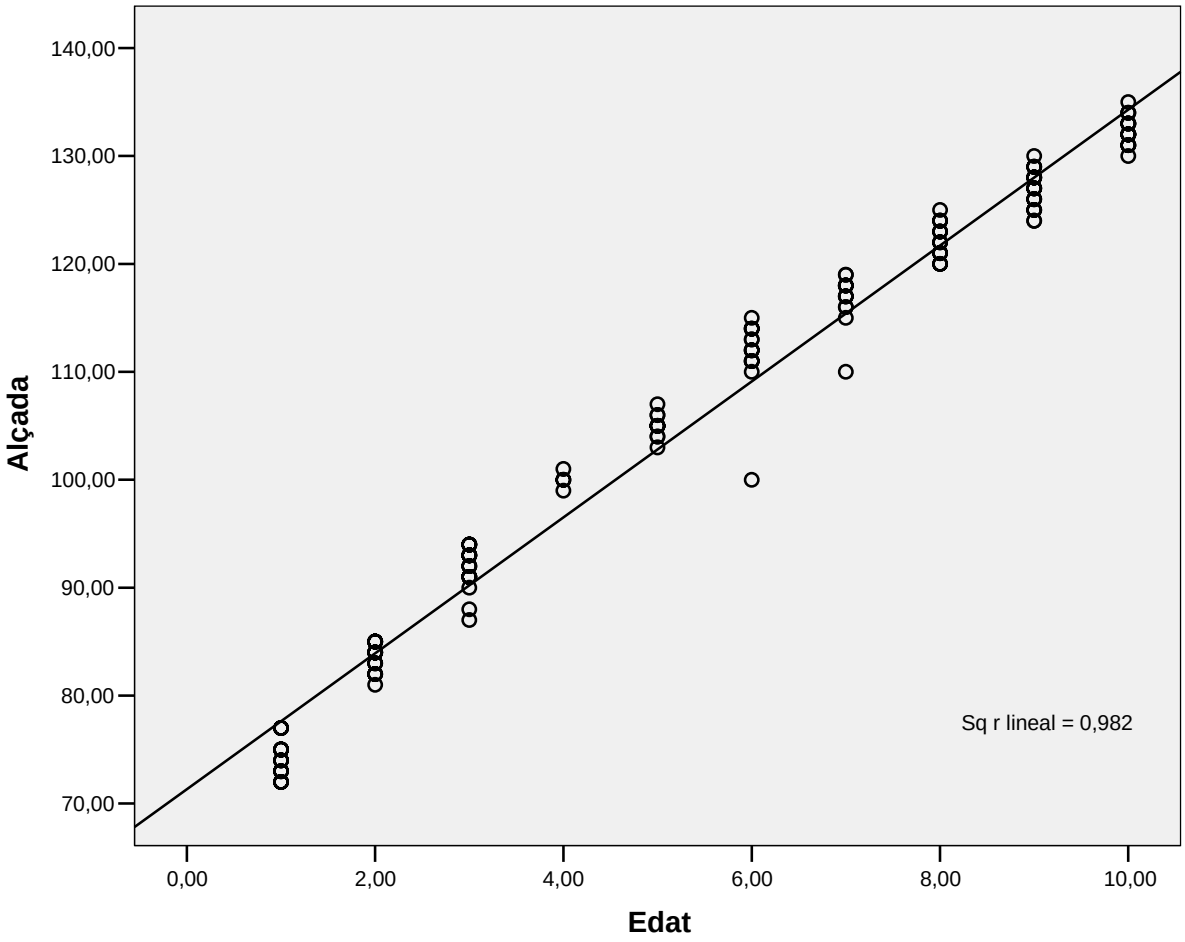


Aquest diagrama ens indica on se situa cada un dels 150 casos, creuant l'edat amb l'alçada. Una vegada col·locats tots els punts, podem veure que tendeixen a situar-se en

uns llocs concrets del diagrama i que segueixen una tendència determinada, de manera que els nens més grans tenen més alçada i els nens més petits en tenen menys. Aquest cas és força obvi, però altres vegades voldrem saber si entre dues variables hi ha alguna relació d'aquest tipus, i no resultarà tan obvi com entre l'edat i l'alçada.

Observeu que cada punt del gràfic representa dos valors per a un mateix nen: l'edat i l'alçada. La tendència que segueixen aquests valors es pot representar amb una línia recta que sintetitza la inclinació del diagrama de dispersió i la seva direcció, d'esquerra a dreta, en aquest cas. Aquesta línia recta s'anomena *recta de regressió*.

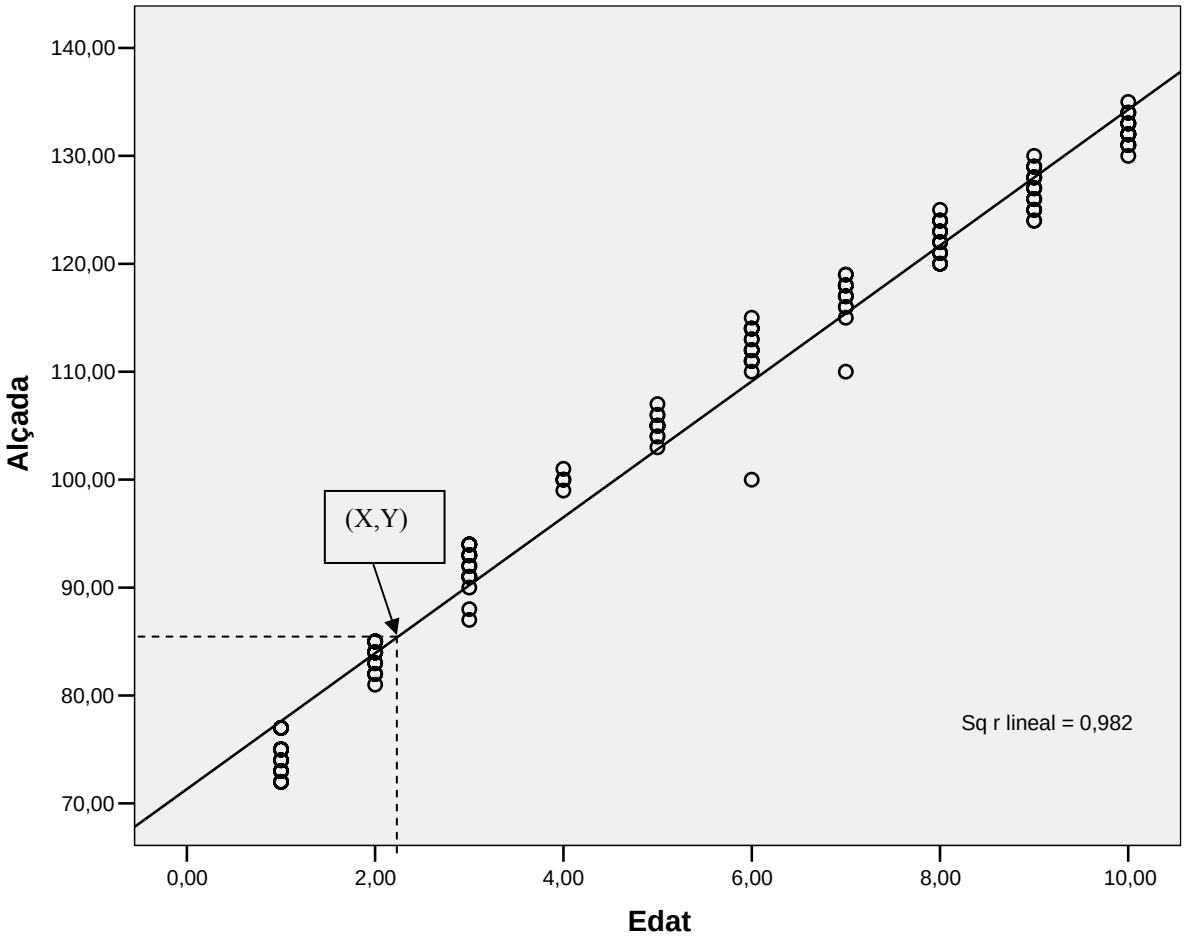
Gràfic 3.2: Diagrama de dispersió edat-alçada amb la recta de regressió



Quan sabem que dues variables estan correlacionades, trobar aquesta línia recta que resumeix la tendència dels punts ens permetrà predir el valor d'una variable en una persona si coneixem el valor de l'altra.

Així, si coneixem la recta de regressió entre edat i alçada, sabent l'edat d'un nen determinat podrem predir amb una certa aproximació l'alçada que té. Per exemple, podrem predir que l'alçada d'un nen de 2 anys i 4 mesos serà, aproximadament, d'uns 86 cm, tal com queda representat al gràfic.

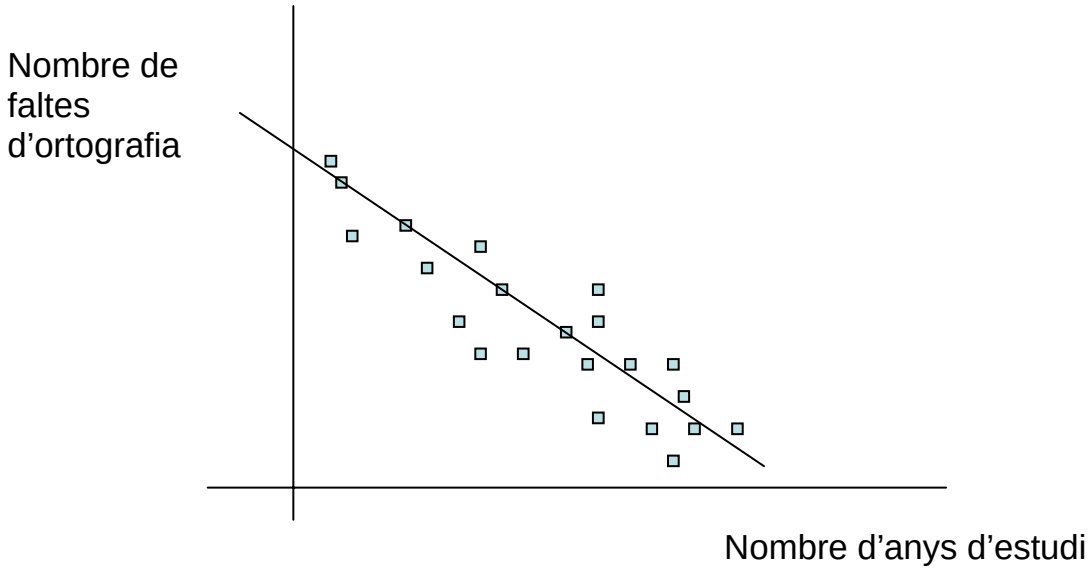
Gràfic 3.3: Representació d'una predicció



Quan dues variables estan correlacionades, podem fer aquesta mena de prediccions. Ara bé, si no ho estan, com seria el cas de l'alçada i el coeficient intel·lectual, no podem fer aquest tipus de prediccions.

No totes les variables tenen el tipus de correlació que hem descrit fins aquí. Posem per cas que hem agafat una mostra de persones d'entre 8 i 20 anys i volem veure si hi ha relació entre el nombre d'anys d'estudis d'aquestes persones i el nombre de faltes d'ortografia que fan quan escriuen un text dictat de 150 paraules. En principi, si hi hagués correlació, probablement seria en el sentit següent: com més anys d'estudis, menys nombre de faltes d'ortografia. En aquest cas, ho podríem representar com es pot observar al gràfic 3.4. La recta de regressió té una inclinació diferent de la que vèiem al gràfic 3.2. En aquest cas, es diu que hi ha correlació, però és negativa, cosa que significa que la tendència de la distribució conjunta entre les dues variables és que, a mesura que augmenten els valors d'una variable, disminueixen els valors de l'altra, mentre que en la correlació positiva la tendència és que, a mesura que augmenten els valors d'una variable, augmenten també els valors de l'altra variable.

Gràfic 3.4: Representació d'una correlació negativa

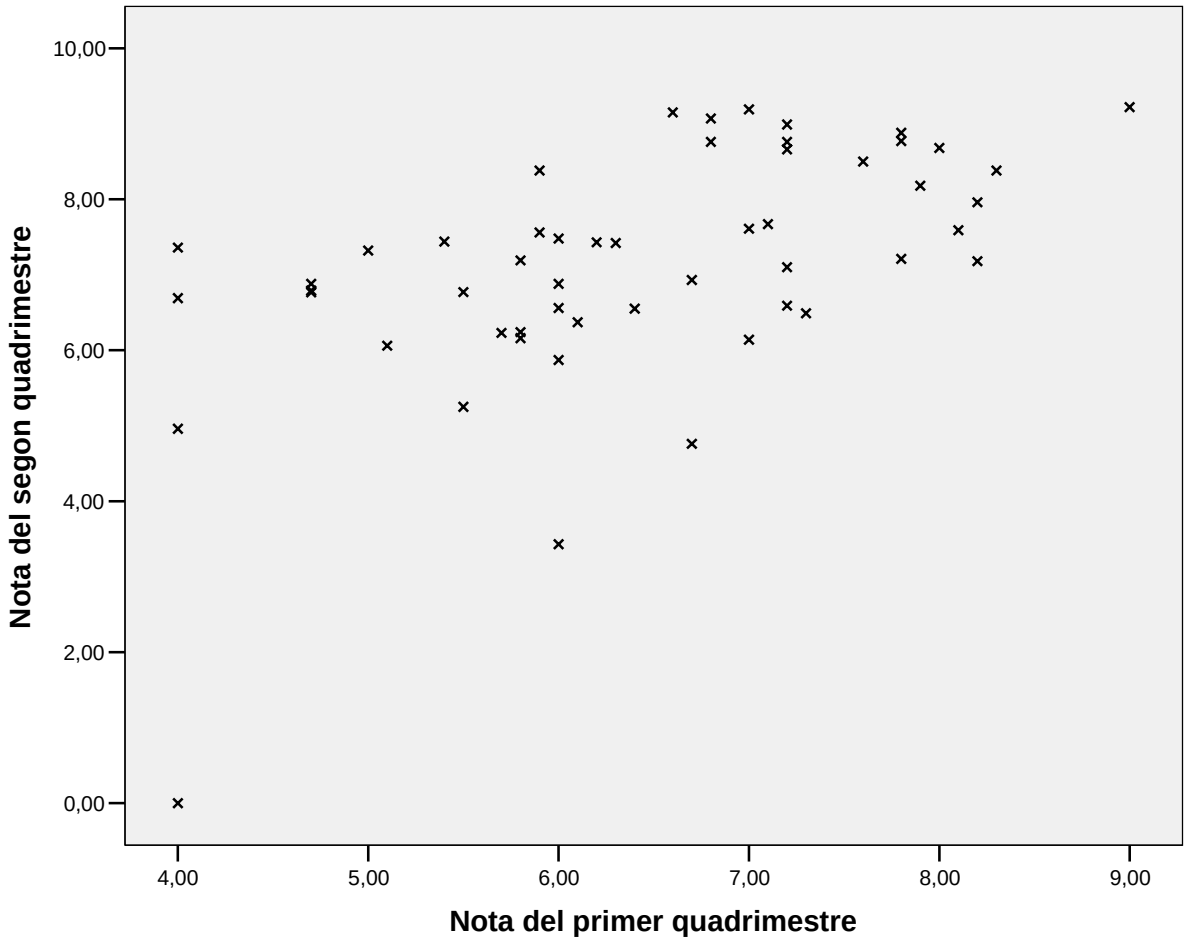


Un altre exemple

Hem recollit les notes de l'assignatura de Bases metodològiques de la investigació educativa del curs 2004-2005 i volem veure si hi ha relació entre les notes que els estudiants treuen en finalitzar el primer quadrimestre i les que treuen en finalitzar el segon. El diagrama de dispersió d'aquestes dades el trobem al gràfic 5. En aquest cas, s'observa una certa tendència dels punts, però ens és molt més difícil que en el cas de les variables edat-alçada determinar, a simple vista, si és probable que hi hagi una relació entre aquestes dues variables.

En general, per calcular el grau de correlació entre dues variables, ens caldrà buscar un coeficient que ens permeti saber si hi ha relació i si aquesta és més o menys intensa. Parlarem d'un tipus de correlació que es coneix amb el nom de *correlació lineal*.

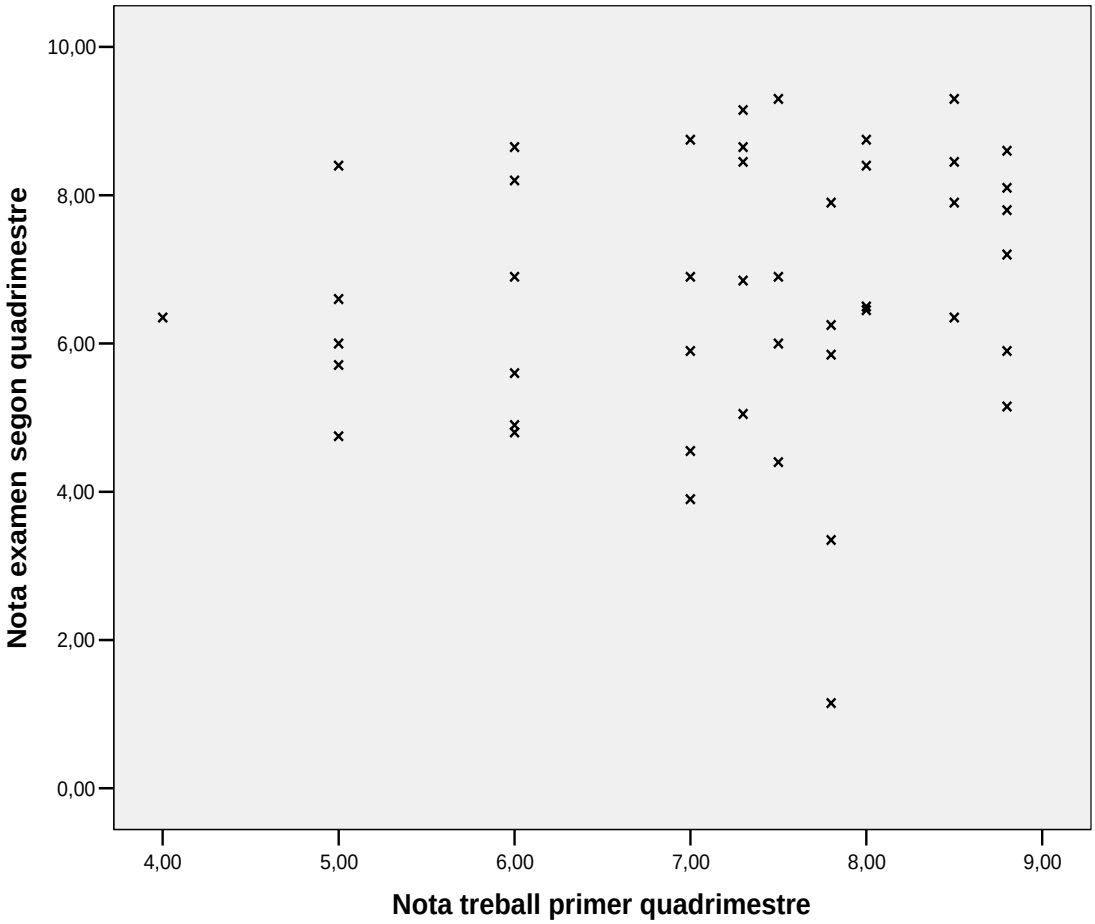
Gràfic 4.5: Diagrama de dispersió per a les qualificacions del primer i del segon quadrimestres de l'assignatura de Bases metodològiques de la investigació educativa. Curs 2004-2005



3.2. Correlació lineal

Tal com diuen Welkowitz, Ewen i Cohen (1981: 204), la relació lineal implica que, si dibuixem un diagrama de dispersió amb els valors de les dues variables, la tendència del núvol de punts obtingut s'ajusta bé a una línia recta. Fixem-nos que els punts que es presenten en els gràfics 3.3 i 3.4 tendeixen a situar-se damunt d'una línia recta. Això és més difícil d'observar en el gràfic 3.5 (nota primer quadrimestre - nota segon quadrimestre) i en el gràfic 3.6 (nota treball primer quadrimestre - nota examen segon quadrimestre).

Gràfic 3.6: Diagrama de dispersió de la relació entre la nota del treball del primer quadrimestre i la nota de l'examen del segon quadrimestre



Que hi hagi correlació lineal entre dues variables no vol dir que tots els punts hagin d'estar situats damunt d'una recta. En l'exemple que estem comentant, l'existència de correlació entre les qualificacions del primer quadrimestre i les qualificacions del segon quadrimestre indicaria que, en general, un estudiant que té una puntuació elevada el primer quadrimestre tendeix a treure una puntuació elevada el segon quadrimestre i que si un estudiant té una puntuació baixa en finalitzar el primer quadrimestre també tendeix a tenir una puntuació baixa en finalitzar el segon quadrimestre. Però això no significaria que tots els estudiants segueixin aquesta tendència. Podria ser que un estudiant hagués tret una puntuació alta en finalitzar el primer semestre i, en canvi, una puntuació baixa en finalitzar el segon, i això no faria variar la tendència de la relació en cas que n'hi hagués. És a dir, el coeficient de correlació lineal ens indicarà quina és *la tendència general del conjunt de valors de les dues variables que s'analitzen*, és a dir, ens indicarà si les dues variables varien conjuntament o no.

El coeficient de correlació lineal oscil·larà entre -1 i +1, segons que la correlació sigui negativa o positiva. Un valor del coeficient molt pròxim a 0 o igual a 0 indicarà que no hi ha correlació.

Així, per exemple, l'índex de correlació corresponent a la relació entre les notes del treball del primer quadrimestre i les notes de l'examen del segon quadrimestre és de 0,163. Aquest índex s'acosta molt a 0, i si observem el gràfic 3.6 veiem que el núvol de punts està molt dispers en el pla configurat pels eixos de coordenades; per tant, en aquest cas no hi ha correlació.

En canvi, tornant a l'exemple de l'edat i l'alçada dels nens d'1 a 10 anys, l'índex de correlació és 0,991. Això indica que és una correlació positiva i que és una correlació elevada. En el diagrama de dispersió corresponent (gràfic 3) veiem que tots els punts tendeixen a situar-se al voltant d'una recta.

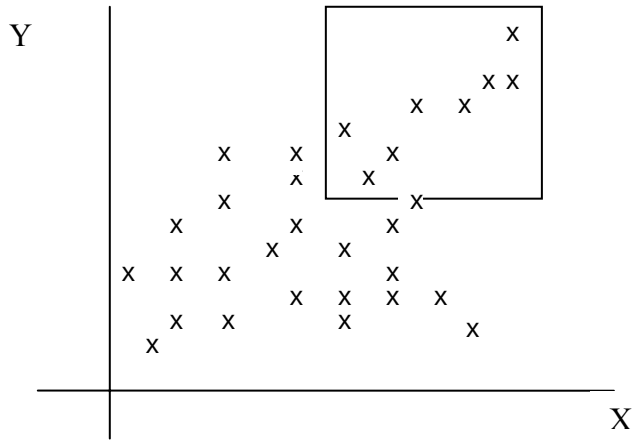
En resum, el coeficient de correlació té les característiques següents:

1. El valor 0 indica que no hi ha relació lineal entre les variables.
2. El valor numèric del coeficient indica la força o intensitat de la relació (els valors absoluts grans indiquen una correlació forta entre les dues variables i els valors absoluts petits indiquen una correlació feble).
3. El signe del coeficient indica la direcció de la correlació.
4. El valor positiu més elevat és +1 i el valor negatiu més elevat és -1.

També cal tenir present que com més gran sigui la mostra més probabilitats hi haurà que el resultat de l'anàlisi correlacional s'acosti al que passa realment a les variables en la població. Si la mostra és petita, pot passar que:

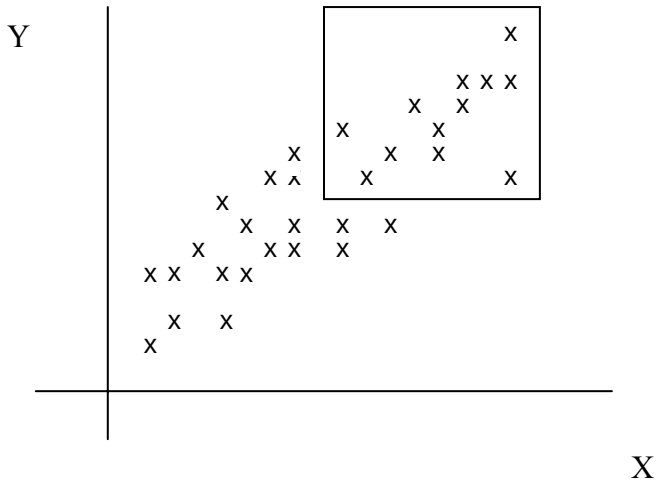
— *Ens surti que hi ha correlació, però en realitat aquesta és deguda a l'atzar*, de manera que, si agaféssim una mostra més gran, veuríem que realment no n'hi ha, de correlació. Al gràfic 7, si ens fixem en els punts que hi ha dins del requadre, veurem que sembla que hi ha correlació. Si féssim l'anàlisi només amb aquests valors, és probable que ens sortís un índex elevat. No obstant això, si agafem una mostra molt més gran, veiem que, en realitat, és molt poc probable que hi hagi correlació entre les dues variables.

Gràfic 3.7



— *Ens surti que no hi ha correlació, però, en realitat, sí que n'hi ha.* Si agafem una mostra més gran, podem veure que el núvol de punts que semblava molt dispers no ho és tant. El gràfic 8 representa aquesta situació.

Gràfic 3.8.



3.2.1. El coeficient de correlació de Pearson

El coeficient de correlació de Pearson ens indica si hi ha correlació lineal entre dues variables. Es representa amb una *r*.
Per poder aplicar el coeficient de Pearson cal que, com a mínim, es compleixin les condicions següents:

1. Les variables han d'estar mesurades en una escala d'interval.
2. Hem de tenir un mínim de 30 casos per analitzar, per tant, 30 parells de valors.
3. Es parteix de la base que la distribució de cada variable segueix el model de la corba normal. Teòricament caldria analitzar si la distribució de valors segueix aquesta distribució abans de decidir aplicar el coeficient de Pearson. Per raons pràctiques, en els exercicis donarem per fet que es compleix aquesta condició.

També cal tenir en compte que el coeficient de Pearson es pot buscar siguin quines siguin les unitats del sistema de mesura que s'hagi utilitzat (sempre que es tracti d'una escala d'interval). És a dir, podem correlacionar una variable mesurada amb metres amb una de mesurada en grams. Podem buscar si hi ha correlació entre els resultats de dues proves encara que una hagi estat puntuada de l'1 al 10 i l'altra de l'1 al 100.

La fórmula és la següent:

$$r_{xy} = \frac{N\Sigma XY - \Sigma X \Sigma Y}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]}} \qquad (8)$$

N = nombre de parelles de valors

Interpretació

Ja hem dit que el coeficient de correlació oscil·la entre els valors -1 i +1, sent menor com més s'acosta a 0 i major com més s'acosta a -1 o a +1. Amb tot, per interpretar si la correlació és significativa, hem de recórrer a les taules de la *r* de Pearson, que trobareu a l'annex 1 d'aquest document.

Per interpretar la *r* de Pearson cal buscar el valor crític de *r*. Això es fa tenint en compte el marge d'error (normalment utilitzarem un marge d'error del 0,05 o del 0,01, mirant la significació bilateral) i els graus de llibertat, que són N - 2.

A continuació comparem la *r* que ens ha sortit a nosaltres utilitzant la fórmula amb la *r* que ens surt a les taules. En aquest cas sempre es comparen els valors absoluts, és a dir, no es té en compte el signe que surt a l'índex de correlació. Cal tenir present que el valor de les taules ens indica quin ha de ser el valor mínim de *r* perquè puguem considerar que la correlació és significativa. Per tant:

- Si la *r observada* és més gran que la *r de la taula*, direm que hi ha correlació significativa.

- Si la *r observada* és més petita que la *r de la taula*, direm que no hi ha correlació significativa.

Així, per exemple, imaginem-nos que busquem el coeficient de correlació de Pearson entre l'edat i l'alçada d'una mostra de 37 nens d'entre 1 i 10 anys. Ens surt que $r = 0{,}825$. Podem pensar que la correlació serà alta, però, per estar-ne segurs i veure si la correlació és realment significativa, hem de recórrer a les taules que tenim a l'annex 1. Farem la interpretació per a un marge d'error de 0,05.

Anem a les taules i mirem el valor crític de *r* per a un marge d'error de 0,05 i $N - 2$ graus de llibertat, que en aquest cas són 35. El valor que trobem a les taules és 0,325. Aquest valor ens indica el valor mínim que ha de tenir *r* perquè puguem dir que la correlació és significativa. Com que la *r* observada és més gran que la *r* crítica, direm que hi ha correlació entre l'edat i l'alçada.

$r = 0{,}825$
$r_c = 0{,}325$ $\quad r > r_c$ Per tant, hi ha correlació entre les dues variables.

Un altre exemple
Imaginem-nos que hem recollit informació sobre el nombre d'anys d'estudi d'una mostra de 100 persones de 20 anys i informació sobre el nombre de faltes d'ortografia que han fet en un dictat de 200 paraules. Busquem el coeficient de correlació de Pearson i ens surt que és $r = -0{,}245$. Volem saber si hi ha correlació entre aquestes dues variables per a un nivell de significació del 95,5 % (això és el mateix que dir que es vol saber per a un marge d'error de 0,05).

En aquest cas, veiem que es tracta d'un índex força baix, però, tal com hem fet abans, per saber si la correlació és significativa ens cal mirar les taules.

r_c (α = 0,05 i 98 g. ll.) = 0,205 (No existeix un valor exacte per a 98 graus de llibertat. En aquest cas, podríem decidir agafar el corresponent a 100 o el corresponent a 90. Per norma general, agafem el valor crític que sigui més gran.)

$r > r_c$, és a dir, $|0{,}245| > |0{,}205|$. Per tant, podrem concloure que hi ha correlació significativa a un nivell de significació del 95,5 %. La correlació és negativa. Això vol dir que, segons aquest resultat, com més anys d'estudis, menys faltes d'ortografia han comès.

(Recordeu que comparem els valors absoluts de la *r* observada i de la *r* crítica.)

Tal com expliquen Welkowitz, Ewen i Cohen (1991), si una *r* de Pearson és estadísticament significativa, aquesta significació denota un cert grau de relació lineal entre les dues variables de la població. No obstant això, no indica una relació significativament elevada o significativament forta; només indica la improbabilitat d'una relació nul·la entre ambdues variables de la població. Cal observar que, com més gran és la mida de la mostra, més petit és el valor del coeficient de correlació necessari perquè hi hagi significació estadística. És a dir, cal tenir no només una significació estadística, sinó també un valor absolut de *r* prou alt abans de decidir que probablement la correlació a la població és prou gran per indicar una relació prou forta.

És important recordar que la *r* de Pearson ens indica <u>només si hi ha correlació lineal</u> entre dues variables. Amb aquest coeficient, doncs, no es pot veure si la relació és curvilínia o d'un altre tipus. Per això calen altres tipus de coeficients, dels quals no parlarem en aquest curs d'introducció a l'estadística.

3.2.2. Correlació i causalitat

Tal com assenyalen Etxeberria i Tejedor (2005), a vegades es confon l'existència de relació entre variables amb la possible existència d'una relació de causa-efecte entre elles, però això són conceptes molt diferents. Entendre correctament el concepte de correlació implica entendre que el coeficient no permet determinar la causa de la relació entre dues variables. Pot haver-hi una correlació alta entre dues variables per tres possibles raons:

a) *Les dues variables varien de manera simultània.* Per exemple, entre el pes i l'alçada de les persones hi ha una relació. Les persones més altes solen pesar més, però no és que el pes sigui la causa de l'alçada o l'alçada del pes; simplement, varien de manera conjunta.

b) Una és la causa i l'altra, l'efecte. Però perquè sigui així cal que:
    a. Hi hagi relació entre la variable causa i la variable efecte.

b. La *variable causa* s'ha de produir abans que la *variable conseqüència*.
c. No ha d'existir una variable que mediatitzi aquesta relació.

Per exemple, podríem trobar que l'autoestima i el rendiment acadèmic estan relacionats, però difícilment podríem dir que una baixa autoestima pot ser una causa del baix rendiment acadèmic, ja que és molt difícil determinar que l'autoestima s'hagi produït abans del rendiment acadèmic. Senzillament, es tracta de dues característiques que estan associades, però no podrem dir que una sigui conseqüència de l'altra.

c) Tant la variable *X* com la variable *Y* són causades per una tercera variable, que és la causa de totes dues i de la relació entre elles. Per exemple, hi ha correlació entre l'alçada dels nens i els resultats en lectoescriptura, és a dir, els nens més alts tenen millors resultats. Això és així perquè hi ha una tercera variable, l'edat, que és la causa que els nens siguin més alts i, per tant, més grans, i llegeixin i escriguin millor que els més baixets, que són més petits.

El problema de l'atribució de la causalitat és un problema lògic o científic, no estadístic. Cal ser molt conscients que l'existència de relació estadística entre variables no és suficient per parlar de relació causa-efecte entre elles. Tal com diuen Etxeberria i Tejedor (2005), l'existència de correlació és una condició necessària però no suficient per establir la causalitat entre dues variables.

3.2.3. Consideracions finals

La utilització del coeficient de correlació de Pearson implica el supòsit que hi ha una relació lineal entre dues variables. Si la relació que existeix entre dues variables no és lineal sinó curvilínia, la *r* de Pearson no la detectarà.

Quan es comprova la significació estadística d'un coeficient de correlació, se suposa que la distribució implicada és normal bivariada, és a dir, les puntuacions *Y* es distribueixen normalment per cada valor de *X* i, de manera recíproca, els valors de *X* es distribueixen normalment per cada valor de *Y*. No obstant això, quan els graus de llibertat són més de 25 o 30, l'incompliment d'aquest supòsit té poca influència sobre la validesa de la prova (Welkowitz, Ewen i Cohen, 1981: 199-200).

Els coeficients de correlació no es poden interpretar com a percentatges. Per exemple, no es pot dir que una correlació de 0,80 és el 80 % d'una relació perfecta o que equival al doble d'una correlació de 0,40. En canvi, el quadrat del coeficient de correlació sí que permet una interpretació en termes de percentatge de la intensitat de relació entre dues variables.

Tal com diu Calvo (1987), els coeficients de correlació no formen una escala quantitativa d'unitat constant; és a dir, entre una $r = 0{,}80$ i una $r = 0{,}70$ no existeix la mateixa diferència que entre una $r = 0{,}40$ i una $r = 0{,}50$. Tampoc no és veritat que $r = 0{,}50$ sigui igual al doble de $r = 0{,}25$.

Com més heterogènia és la població, més força té el coeficient de correlació de Pearson. Si la població és molt homogènia, *r* serà més petita. Per exemple, la correlació entre l'alçada i el salt de longitud és molt alta, però si només agafem les persones que fan més d'1,80 d'alçada, veurem que en aquest grup, pel fet de ser més homogeni, la correlació de l'alçada amb el salt de longitud serà menor (Calvo, 1987: 110).

Cal conèixer la naturalesa de les variables per deduir la importància de la correlació. Així, per exemple, una $r = 0{,}30$ entre el pes i la capacitat intel·lectual és alta (perquè en principi és difícil que pugui haver-hi una correlació entre aquestes dues variables). En canvi, una $r = 0{,}80$ entre la pressió arterial a les 10 del matí i a les 11 del matí és baixa. També cal analitzar si la correlació diu alguna cosa que és real o no.

Per ajudar-nos a interpretar la correlació podem recórrer als valors de Guilford (citat per Calvo, 1987: 110):

	Correlació	**Relació**
$0 < r < 0{,}20$	Petita	Molt poc intensa
$0{,}20 < r < 0{,}40$	Baixa	Petita però apreciable
$0{,}40 < r < 0{,}60$	Regular	Considerable
$0{,}60 < r < 0{,}80$	Alta	Intensa
$0{,}80 < r < 1{,}00$	Molt alta	Molt intensa

Coeficient de determinació r^2

r^2 indica el tant per u de la variació entre les variables. És a dir, indica quina proporció representa la correlació trobada de la correlació perfecta. O, dit d'una altra manera, quina proporció de la variabilitat total de la variable *Y* queda explicada per la variable *X* (Calvo, 1987: 110).

Exercici de correlació

Ens preguntem si hi ha correlació significativa entre les notes obtingudes al final del primer quadrimestre i al final del segon quadrimestre pels estudiants de l'assignatura Bases metodològiques de la investigació educativa. Farem la interpretació amb un nivell de significació del 95,5 %.

Per fer aquest exercici d'exemple, treballarem amb una mostra de 30 estudiants, escollits a l'atzar, del curs 2004-2005. A continuació, tenim la taula amb els valors de les dues variables.

Haurem d'aplicar la fórmula de la *r* de Pearson, que és:

$$r_{xy} = \frac{N\Sigma XY - \Sigma X\Sigma Y}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]}}$$

Per tant, els càlculs que ens caldrà fer són els que apareixen a la taula següent, al costat de les dades.

	Nota del primer quadrimestre (X)	Nota del segon quadrimestre (Y)	XY	X^2	Y^2
1	5	7,32	36,60	25	53,58
2	6	6,56	39,36	36	43,03
3	4,7	6,79	31,91	22,09	46,10
4	7	7,61	53,27	49	57,91
5	6,7	6,93	46,43	44,89	48,02
6	8,3	8,38	69,55	68,89	70,22
7	6,4	6,55	41,92	40,96	42,90
8	7,8	7,21	56,24	60,84	51,98
9	7,8	8,77	68,41	60,84	76,91
10	7,3	6,49	47,38	53,29	42,12
11	6,7	4,76	31,89	44,89	22,66
12	7,2	7,1	51,12	51,84	50,41
13	9	9,22	82,98	81	85,01
14	7,1	7,67	54,46	50,41	58,83
15	4	7,36	29,44	16	54,17
26	6,2	7,43	46,07	38,44	55,20
17	5,8	6,16	35,73	33,64	37,95
18	4	6,69	26,76	16	44,76
19	6	7,48	44,88	36	55,95
20	5,5	6,77	37,24	30,25	45,83
21	6	5,87	35,22	36	34,46
22	6,8	8,76	59,57	46,24	76,74
23	5,4	7,44	40,18	29,16	55,35
24	4,7	6,77	31,82	22,09	45,83
25	5,9	8,38	49,44	34,81	70,22
26	5,9	7,56	44,60	34,81	57,15
27	5,5	5,25	28,88	30,25	27,56
28	8,2	7,96	65,27	67,24	63,36
29	8,2	7,18	58,88	67,24	51,55
30	7,8	8,88	69,26	60,84	78,85
	ΣX=192,9	**ΣY=217,3**	**ΣXY=1414,74**	**ΣX^2=1288,95**	**ΣY^2=1604,66**

Apliquem la fórmula (8):

$$r_{xy} = \frac{N\Sigma XY - \Sigma X\Sigma Y}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]}} =$$

$$= \frac{30 \cdot 1414{,}74 - (192{,}9 \cdot 217{,}3)}{\sqrt{[30 \cdot 1288{,}95 - (192{,}9)^2][30 \cdot 1604{,}66 - (217{,}3)^2]}} =$$

$$= \frac{42.442{,}2 - (41.917{,}17)}{\sqrt{[38.668{,}5 - (37.210{,}41)][48.139{,}8 - (47.219{,}29)]}} =$$

$$= \frac{525{,}03}{\sqrt{[1458{,}09][920{,}51]}} = \frac{525{,}03}{1.158{,}53} = 0{,}453$$

Per tant, $r_{xy} = 0{,}453$.

Ara, per fer la interpretació ens cal anar a les taules i buscar la *r* crítica. Per *N*-2 graus de llibertat i $\alpha = 0{,}05$ (el marge d'error corresponent al 95,5 % de nivell de significació).

r_c (g. ll. = 28; $\alpha = 0{,}05$) = 0,361

0,361 < 0,453

Per tant, podem dir que, amb un marge d'error del 0,05, hi ha correlació significativa entre les notes del primer quadrimestre i les notes del segon quadrimestre dels estudiants que el curs 2005-2006 van fer l'assignatura Bases metodològiques de la investigació educativa.

Mirant les taules de Guilford veiem que la correlació és regular, cosa que implica una relació considerable.

Si busquem el coeficient de determinació r^2 veiem que és de 0,205. Això vol dir que un 20,5 % de la variabilitat de les notes del segon quadrimestre s'explica per les notes que s'obtenen el primer quadrimestre.

Un altre exemple

Volem saber si hi ha correlació entre les despeses per alumne i curs a educació primària a diferents països de la UE i les despeses en educació (expressades en percentatge del PIB). Les dades apareixen a la taula següent. Fes la interpretació per a un marge d'error tant del 0,05 com del 0,01.

	Despeses per alumne i curs escolar a ed. primària	Despeses d'educació en percentatge del PIB als països de l'euro
Espanya	3635	4,9
Alemanya	3818	5,3
França	4139	6,1
Portugal	3478	5,7
Irlanda	3018	4,6
Itàlia	5354	4,9
Bèlgica	3952	5,5
Àustria	6568	5,7
Finlàndia	4138	5,6
Països Baixos	4162	4,7
Grècia	2176	4

A continuació tens els càlculs fets, només cal que apliquis la fórmula (8):

	Despeses per alumne i curs escolar a ed. primària	Despeses d'educació en percentatge del PIB als països de l'euro	XY	X^2	Y^2
Espanya	3635	4,9	17811,5	13213225	24,01
Alemanya	3818	5,3	20235,4	14577124	28,09
França	4139	6,1	25247,9	17131321	37,21
Portugal	3478	5,7	19824,6	12096484	32,49
Irlanda	3018	4,6	13882,8	9108324	21,16
Itàlia	5354	4,9	26234,6	28665316	24,01
Bèlgica	3952	5,5	21736	15618304	30,25
Àustria	6568	5,7	37437,6	43138624	32,49
Finlàndia	4138	5,6	23172,8	17123044	31,36
Països Baixos	4162	4,7	19561,4	17322244	22,09
Grècia	2176	4	8704	4734976	16
	ΣX =44438	ΣY = 57	ΣXY = 233848,6	ΣX^2 = 192728986	ΣY^2=299,16

Apliquem la fórmula de la *r* de Pearson.

$$r_{xy} = \frac{N\Sigma XY - \Sigma X\Sigma Y}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]}} =$$

$$= \frac{(11 \cdot 233848{,}6) - (44438 \,.\, 57)}{\sqrt{[(11 \cdot 192728986) - (44438)^2][(11 \cdot 299{,}16) - (57)^2]}} = 0{,}505$$

Per fer la interpretació busquem la *r* crítica a les taules:

r_c (g. ll. = 9; α = 0,05) = 0,602

r_c (g. ll. = 9; α = 0,01) = 0,735

Interpretació:

Per als dos marges d'error veiem que la *r* que hem obtingut aplicant la fórmula és menor que la *r* que apareix a les taules. Per tant, podem dir que tant amb un marge d'error de 0,05 com amb un marge d'error de 0,01, no hi ha correlació significativa entre les despeses per alumne i curs escolar i les despeses d'educació en percentatge del PIB. És a dir, no podem afirmar que les despeses per alumne i curs escolar a educació primària tinguin res a veure amb el percentatge del PIB dedicat a educació.

Tot i això, cal fer un advertiment, i és que aquesta vegada hem aplicat el coeficient de Pearson per fer un exercici amb pocs valors. En aquest cas, en què només tenim dades d'onze països, la mostra és excessivament reduïda perquè la informació que extraiem mitjançant la *r* de Pearson sigui realment fiable.

3.3. Predicció i regressió lineal

Tal com dèiem al començament d'aquest capítol, si dues variables estan correlacionades, és a dir, si dues variables varien de forma conjunta, podrem explicar els canvis que es produeixen en una d'aquestes (variable criteri) a partir de les dades de les variables que s'anomenen *variables predictores*.

Per exemple, si la variable criteri que volem analitzar és l'alçada dels nens d'1 a 10 anys, podem tenir diferents variables predictores que poden explicar les variacions d'alçada dels nens d'aquesta edat, entre les quals podem esmentar l'edat, el sexe, el número que calcen, etc. Si treballem només amb una variable predictora, estarem

parlant de regressió simple, i si en tenim diverses, de regressió múltiple (Etxeberria i Tejedor, 2005: 211). Nosaltres només parlarem de la *regressió simple.*

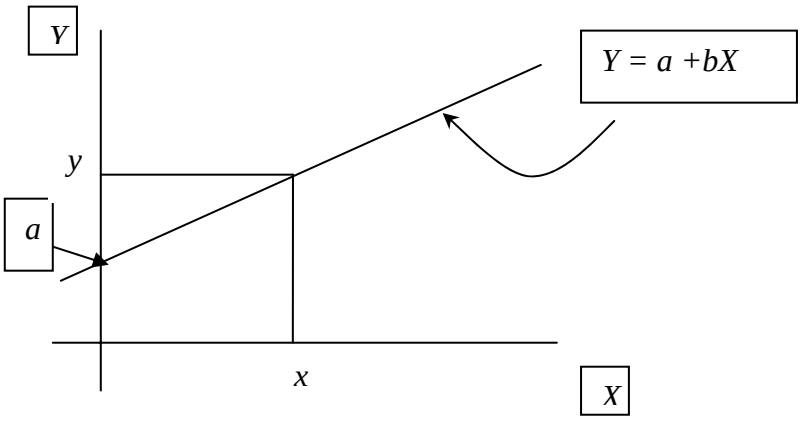
3.3.1. Regressió simple

Per poder fer prediccions d'una variable criteri (Y) a partir d'una variable predictora (X), el primer que cal és que, una vegada establert que efectivament existeix una correlació lineal significativa entre les dues variables, busquem la *recta de regressió.*

En general, se sap que l'equació d'una recta té l'estructura següent:

$Y = a + bX$

Y = el valor predit de la variable criteri
b = pendent de la recta
X = el valor de la variable predictora
a = valor de Y quan $X = 0$



Recordem que la recta o línia de regressió és la recta que representa millor la tendència dels punts en un diagrama de dispersió. Per trobar aquesta recta, cal buscar el coeficient b de l'equació i després buscar la constant a.

La fórmula per trobar el coeficient b (que representarem com a b_{yx} per expressar que es tracta de la recta de regressió de Y sobre X) és la següent:

$$b_{yx} = \frac{N\Sigma XY - \Sigma X\Sigma Y}{N\Sigma X^2 - (\Sigma X)^2} \quad (9)$$

Una vegada es té el valor de b_{yx}, es busca el valor de a_{yx} a partir dels valors de la mitjana de la variable *X* i de la mitjana de la variable *Y*:

$$a_{yx} = \overline{Y} - b_{yx}\,\overline{X} \qquad (10)$$

Per exemple, abans hem trobat correlació entre les notes del primer quadrimestre i les del segon quadrimestre de l'assignatura Bases metodològiques de la investigació educativa. Per tant, ara buscarem la recta de regressió.

Primer, cal buscar el coeficient *b*, amb la fórmula (9):

$$b_{yx} = \frac{N\Sigma XY - \Sigma X\Sigma Y}{N\Sigma X^2 - (\Sigma X)^2} = \frac{(30 \cdot 1414{,}74) - (192{,}9 \cdot 217{,}3)}{(30 \cdot 1288{,}95) - 192{,}9^2} =$$

$$= \frac{42442{,}2 - 41917{,}17}{38668{,}5 - 37210{,}41} = \frac{525{,}03}{1458{,}09} = \mathbf{0{,}36}$$

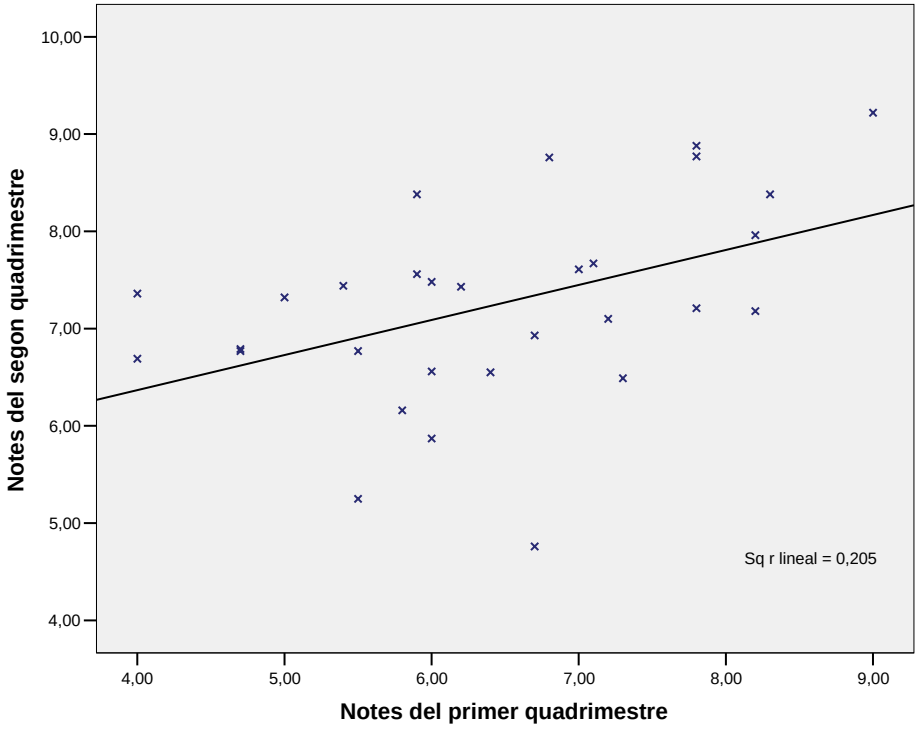
Ara ens cal buscar la *a*, amb la fórmula (10). Per a això, hem de trobar la mitjana de la variable *X* i la mitjana de la variable *Y*.

$\overline{X} = 6{,}43 \qquad \overline{Y} = 7{,}24$

$a_{yx} = \overline{Y} - b_{yx}\,\overline{X} = 7{,}24 - (0{,}36 \cdot 6{,}43) = \mathbf{4{,}93}$

L'equació de la recta és, per tant, $\boldsymbol{Y = 4{,}93 + 0{,}36X}$**.**

Gràfic 3.9: Diagrama de dispersió i recta de regressió de les notes de les variables "notes del primer quadrimestre" i "notes del segon quadrimestre" de la mostra de 30 casos



Sabent quina és la recta de regressió per a aquestes dues variables, se suposa que si coneixem el valor d'una de les variables podrem predir el valor de l'altra. Així, ens podem preguntar quina serà la puntuació del segon quadrimestre per estudiant que obtingui una puntuació de 5 el primer quadrimestre. L'únic que hem de fer és substituir la *X* pel valor de 5 i sortirà Y:

$$Y = 4{,}93 + 0{,}36 \cdot 5 = \mathbf{6{,}73}$$

3.3.2. Error típic de predicció

La puntuació de 6,73 seria la puntuació que es troba exactament a sobre de la recta de regressió per al valor de $X = 5$. Ara bé, cal tenir en compte que quan es fa un pronòstic o predicció, sempre es comet el que es coneix com a *error de predicció*. És a dir, si miréssim realment quina puntuació treu del segon quadrimestre un estudiant que hagi tret una puntuació de 5 al final del primer quadrimestre, difícilment coincidiria exactament amb el valor predit de 6,73. Per a cada valor de *Y* que predim es comet un error de predicció. Els matemàtics ens indiquen com es pot calcular el que es coneix com a *error típic de predicció*, que és aproximadament una mitjana de tots els errors de predicció que es cometen per a cada valor de *X* observat. La fórmula per calcular l'error típic de predicció és la següent:

$\sigma_{y'} = \sigma_y \sqrt{1-r^2}$ σ_y = desviació típica de la variable *Y*
r^2 = quadrat del coeficient de correlació

Aquest error típic de l'estimador de *Y* pot ser interpretat com la dispersió de *Y* al voltant de la línia de regressió.

Per al cas que ens ocupa, l'error típic de predicció serà:

$\sigma_{y'} = \sigma_y \sqrt{1-r^2} = 1{,}03\sqrt{1 - 0{,}453^2} = 0{,}92$

Una predicció serà bona com més petit sigui l'error típic de predicció. En termes generals, si l'error típic de predicció és menor que la desviació típica de la variable *Y*, podem dir que la predicció feta serà bastant bona. Si, en canvi, l'error típic de predicció és més gran que la dispersió de la variable *Y*, llavors la predicció no serà gaire fiable. Tal com diuen Welkowitz, Ewen i Cohen (1981), en aquest segon cas el coneixement d'un valor concret de *X* no permetrà fer una bona predicció sobre del valor que tindrà la *Y* que se li associa. La predicció serà bona en la mesura que el coneixement de *X* redueixi la dispersió en la predicció de *Y*.

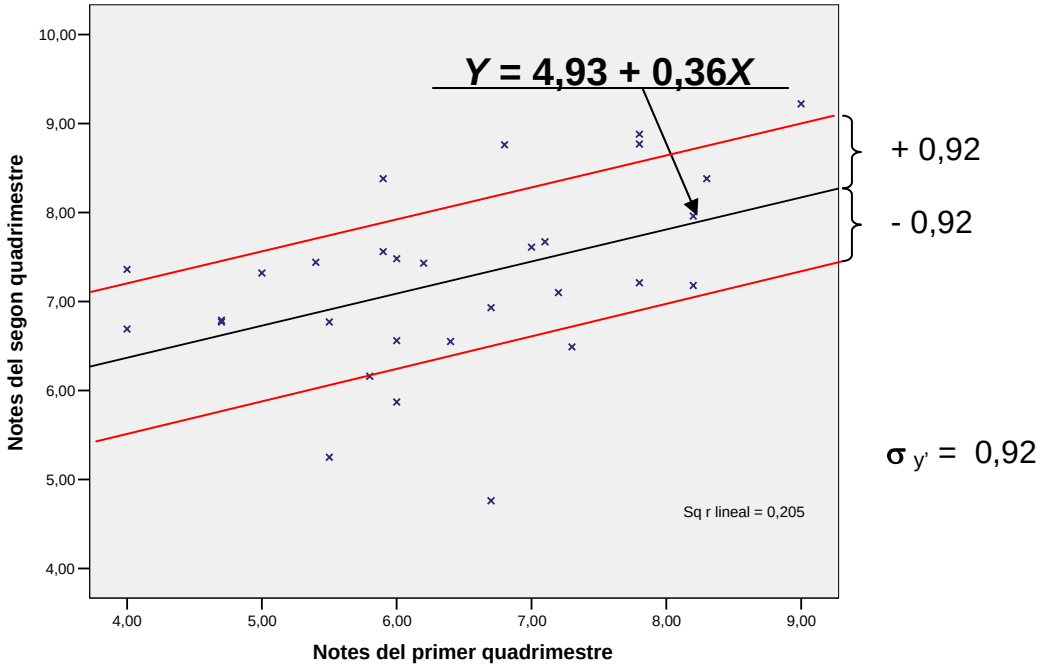
En l'exemple, 0,92 és menor que la desviació típica de la variable *Y*, que és d'1,03, encara que la diferència no és gaire gran.

Si sumem i restem el valor de l'error típic de predicció a la predicció feta per a una $X = 5$, aconseguirem l'interval de valors de *Y* on serà més probable trobar el valor predit. Així, sabent que el valor predit amb la recta de regressió és de 6,73 (aquest seria el valor exacte damunt de la recta), podem sumar i restar l'error de predicció:

6,73 + 0,92 = 7,65
6,73 – 0,92 = 5,81

Per tant, un estudiant que obtingui una puntuació de 5 el primer quadrimestre probablement traurà una puntuació d'entre 5,81 i 7,65 el segon quadrimestre.

Gràfic 3.10: Recta de regressió i error de predicció de les notes de les variables "notes del primer quadrimestre" i "notes del segon quadrimestre" de la mostra de 30 casos

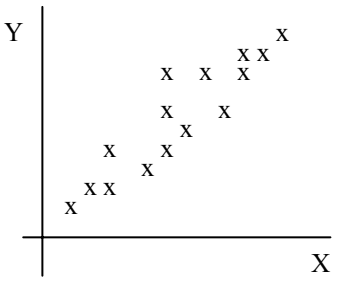


3.4. Exercicis de correlació lineal

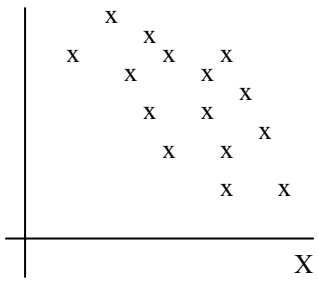
1. Entre els coeficients de correlació següents, escull els més adients als diagrames de dispersió que es presenten:

$r = 0{,}44$ $r = -0{,}12$ $r = -0{,}54$ $r = -0{,}89$ $r = 0{,}92$

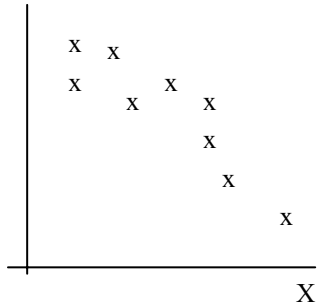
1.1



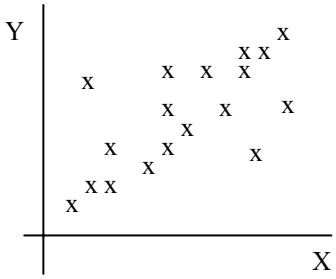
1.2



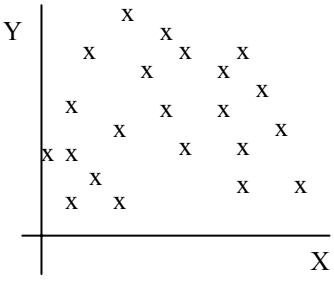
1.3



1.4



1.5



2. A una mostra escollida a l'atzar de 10 alumnes d'una escola se'ls ha passat un test estandarditzat sobre relacions lògiques (test 1) i un altre de càlcul mental (test 2). Els resultats obtinguts són els següents:

Cas	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Test 1	22	20	18	17	16	14	14	12	10	7
Test 2	14	18	12	16	14	12	11	10	9	4

Volem saber si existeix algun tipus de relació entre ambdues proves. Suposem, malgrat el nombre d'elements de la mostra, que es donen les condicions per aplicar proves paramètriques.

2.1. Quina prova aplicaràs? Per què?

2.2. Aquest coeficient és estadísticament significatiu per a un nivell de confiança del 95 %?

2.3. Interpreta'l.

3. Volem saber si hi ha relació entre les habilitats socials i l'acceptació per part del grup d'iguals en nois i noies que realitzen el segon curs d'ESO. Per això es passa un test d'habilitats socials a una mostra de 30 persones escollides a través d'un procediment de mostreig aleatori simple. Per cada alumne també es mesura el

grau d'acceptació del grup d'iguals a partir d'una escala que es passa al grup classe. Contesta les preguntes següents:

3.1. Per saber si hi ha relació entre habilitats socials i acceptació per part del grup d'iguals, quina prova has aplicat?

3.2. És estadísticament significativa? Per què? (Analitza-ho tant per a un marge d'error de $\alpha = 0{,}05$ com per a un marge d'error de $\alpha = 0{,}01$.)

3.3. Interpreta el resultat.

3.4. Es pot buscar la recta de regressió? Si es pot, busca-la i també calcula l'error típic de predicció. (Busca la recta considerant les habilitats socials com a variable predictora (X) i el grau d'acceptació com a variable criteri (Y).)

3.5. Quina predicció podem fer per al cas d'un estudiant que tingui una puntuació de 6,5 en la prova d'habilitats socials?

Cas	Habilitats socials	Grau d'acceptació
1	8	10
2	7	12
3	6	9
4	5	15
5	9	14
6	3	6
7	4	7
8	5	11
9	7	12
10	8	13
11	8	10
12	4	6
13	9	12
14	7	7
15	8	12
16	6	9
17	5	10
18	5	14
19	3	5
20	2	4
21	4	8
22	5	9
23	8	10
24	6	9
25	8	11
26	9	12
27	10	10
28	9	9
29	8	11
30	3	7

4. A la matriu de la pàgina següent tens les dades de 100 casos. Són 100 persones adultes amb discapacitat psíquica profunda a les quals s'ha passat dues proves: una avalua les destreses socials i comunicatives i l'altra, les destreses de vida personal. Una puntuació alta en aquestes proves significa un nivell més alt de destreses. Es vol saber si hi ha alguna relació entre les puntuacions de destreses socials i comunicatives i les puntuacions de destreses de vida personal. Suposem que la distribució de les dues variables segueix el model de la corba normal. S'aplica una determinada prova mitjançant el programa SPSSx i s'obtenen els resultats que apareixen a continuació.

		Puntuació de destreses socials i comunicatives	Puntuació de destreses de vida personal
Puntuació de destreses socials i comunicatives	Correlació de Pearson	1	,819
Puntuació de destreses de vida personal	Correlació de Pearson	,819	1

4.1. Quina prova s'ha aplicat?

4.2. Quin índex s'ha obtingut? Interpreta'l segons un nivell de confiança del 95 % (signe, intensitat).

4.3. En cas que hi hagi correlació, busca el coeficient de determinació i explica què significa.

4.4. Si hi ha correlació, busca la recta de regressió. (Per fer això, utilitza els càlculs de la taula de la pàgina següent.)

4.5. Una persona que obtingui una puntuació de 20 en la prova de destreses socials i comunicatives, quina puntuació tindrà en la prova de destreses de vida personal?

Destreses socials i comunicatives	Destreses de vida personal	***XY***	X^2	Y^2
18	16	288	324	256
3	22	66	9	484
16	17	272	256	289
21	30	630	441	900
9	10	90	81	100
15	16	240	225	256
17	19	323	289	361
26	39	1014	676	1521
21	33	693	441	1089
12	25	300	144	625
1	13	13	1	169
17	20	340	289	400

6	11	66	36	121
12	29	348	144	841
32	41	1312	1024	1681
24	30	720	576	900
1	2	2	1	4
14	18	252	196	324
14	21	294	196	441
20	28	560	400	784
24	38	912	576	1444
0	13	0	0	169
32	37	1184	1024	1369
31	38	1178	961	1444
41	43	1763	1681	1849
32	35	1120	1024	1225
28	33	924	784	1089
29	30	870	841	900
30	29	870	900	841
35	34	1190	1225	1156
20	29	580	400	841
28	38	1064	784	1444
13	28	364	169	784
18	28	504	324	784
8	16	128	64	256
27	41	1107	729	1681
15	23	345	225	529
8	14	112	64	196
21	31	651	441	961
22	13	286	484	169
11	8	88	121	64
15	24	360	225	576
33	29	957	1089	841
26	35	910	676	1225
24	28	672	576	784
13	34	442	169	1156
42	49	2058	1764	2401
13	16	208	169	256
3	7	21	9	49
3	5	15	9	25
17	26	442	289	676
9	4	36	81	16
41	38	1558	1681	1444
40	40	1600	1600	1600
25	35	875	625	1225
12	21	252	144	441
19	16	304	361	256
19	24	456	361	576
15	17	255	225	289
17	17	289	289	289
12	17	204	144	289
16	24	384	256	576

15	22	330	225	484
31	27	837	961	729
25	37	925	625	1369
32	35	1120	1024	1225
18	21	378	324	441
9	13	117	81	169
30	27	810	900	729
34	33	1122	1156	1089
28	36	1008	784	1296
32	38	1216	1024	1444
21	38	798	441	1444
32	40	1280	1024	1600
22	29	638	484	841
22	18	396	484	324
22	32	704	484	1024
22	22	484	484	484
22	27	594	484	729
8	8	64	64	64
29	24	696	841	576
46	49	2254	2116	2401
8	10	80	64	100
9	10	90	81	100
2	6	12	4	36
1	16	16	1	256
12	30	360	144	900
4	25	100	16	625
13	19	247	169	361
1	7	7	1	49
20	26	520	400	676
32	41	1312	1024	1681
24	37	888	576	1369
4	33	132	16	1089
9	9	81	81	81
37	41	1517	1369	1681
8	26	208	64	676
3	1	3	9	1
13	19	247	169	361
18	29	522	324	841
ΣX = 1904	ΣY = 2506	ΣYX = 57474	ΣX^2 = 47834	ΣY^2 = 75076

5. S'ha passat una prova sobre conducta adaptativa a una mostra de 239 persones adultes amb discapacitat psíquica. Aquesta prova té un bloc compost per les puntuacions següents: destreses motores, destreses socials i comunicatives, destreses de vida personal i destreses de vida en comunitat. Per cada persona s'ha obtingut una puntuació de cada una d'aquestes variables. A continuació hi ha una taula amb l'explicació de cada un d'aquests grups de destreses.

Escala	Aspectes avaluats
Destreses motores	• Destreses de motricitat fina i bàsica. • Destreses relatives a la mobilitat. • Forma física. • Coordinació motora general. • Coordinació visuomotora. • Precisió de moviments.
Destreses socials i comunicatives	• Destreses implicades en la interacció social de diferents entorns. • Comprensió i expressió del llenguatge transmès a través de signes, de forma escrita o de forma oral.
Destreses de vida personal	• Capacitat del subjecte per satisfer les seves necessitats d'autonomia personal, principalment a la llar, però també en altres entorns socials: destreses relacionades amb el menjar i la seva preparació; destreses relacionades amb l'ús del servei; vestit; cura de si mateix; habilitats domèstiques.
Destreses de vida en comunitat	• Habilitats necessàries per a un ús adequat dels recursos i serveis de la societat. • Capacitat per respondre adequadament als requeriments econòmics i socials del món laboral i altres situacions socials: ús del rellotge, capacitat per ser puntual, diners i valor de les coses, destreses relacionades amb l'àmbit laboral, sentit de l'orientació a la llar i a la comunitat.

S'ha aplicat el coeficient de correlació de Pearson i s'ha obtingut la matriu de correlacions següent:

		Puntuació escala destreses motores	Puntuació escala destreses socials	Puntuació escala destreses de vida personal	Puntuació escala destreses de vida en comunitat
Puntuació escala destreses motores	Correlació de Pearson	1	,709	,856	,753
Puntuació escala destreses socials	Correlació de Pearson	**,709**	1	,725	,793
Puntuació escala destreses de vida personal	Correlació de Pearson	**,856**	**,725**	1	,753
Puntuació escala destreses de vida en comunitat	Correlació de Pearson	**,753**	**,793**	**,753**	1

Digues quines correlacions resulten significatives per a un marge d'error del = 0,05.

6. Seguint amb el mateix estudi, es vol analitzar si hi ha relació entre el nivell de conductes adaptatives i els problemes de conducta. Per a això, s'aplica la prova de correlació de Pearson entre les diferents variables de conducta adaptativa (destreses motores, destreses socials, destreses de vida personal, destreses de vida en comunitat) i els índexs de problemes de conducta (índex intern, índex extern, índex asocial i índex general). L'*índex intern* inclou els problemes de conducta autolesiva, hàbits atípics i falta d'atenció; *l'índex extern* inclou la conducta heteroagressiva, la destrucció d'objectes i la conducta disruptiva, i

l'índex asocial inclou la conducta social ofensiva i la conducta no col·laboradora.

També cal tenir molt present, perquè afecta la interpretació de la correlació, que en el cas de les puntuacions de les variables de conducta adaptativa, un valor alt indica un nivell alt de destreses. *En el cas dels índexs de problemes de conducta, els valors més alts indiquen menys problemes de conducta i, doncs, els valors més negatius indiquen més problemes de conducta*. Una vegada puntualitzat això, podem observar en la taula següent la matriu de correlacions:

		Índex intern de problemes de conducta	Índex extern de problemes de conducta	Índex asocial de problemes de conducta
Puntuació escala destreses motores	Correlació de Pearson	,148	-,234	-,169
Puntuació escala destreses socials	Correlació de Pearson	,313	-,163	-,078
Puntuació escala destreses de vida personal	Correlació de Pearson	,112	-,201	-,134
Puntuació escala destreses de vida en comunitat	Correlació de Pearson	,269	-,123	-,061

Interpreta quines resulten significatives a un nivell de confiança del 95 % i digues quin tipus de relació és, així com el seu significat.

3.5. Respostes als exercicis de correlació i regressió

Exercici 1

1.1. 0,92

1.2. -0,54

1.3. -0,89

1.4. 0,44

1.5. -0,12

Exercici 2

2.1. Apliquem la prova de la *r* de Pearson tot i que, de fet, aquesta no seria la prova adequada, ja que tenim massa pocs valors. Aquí es tracta de fer un exercici per comprovar si sabem aplicar la fórmula de la *r* de Pearson, i per això posem pocs valors.

Cas	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	
Test 1	22	20	18	17	16	14	14	12	10	7	$\Sigma X = 150$
Test 2	14	18	12	16	14	12	11	10	9	4	$\Sigma Y = 120$
XY	308	360	216	272	224	168	154	120	90	28	$\Sigma YX = 1940$
$\mathbf{X^2}$	484	400	324	289	256	196	196	144	100	49	$\Sigma X^2 = 2438$
$\mathbf{Y^2}$	196	324	144	256	196	144	121	100	81	16	$\Sigma Y^2 = 1578$

$r = 0{,}869$

2.2. Per saber si és significatiu, ens cal mirar la *r* crítica a les taules de l'annex 1, per a *N* - 2 graus de llibertat i per a un marge d'error del 0,05 (corresponent al 95,5 % de nivell de significació). Mirant les taules veiem que

$$r_c \text{ (g. ll. 8; } \alpha = 0{,}05) = 0{,}632$$
$$0{,}632 < 0{,}869$$

Per tant, podem dir que hi ha correlació significativa entre les dues variables, per a un nivell de confiança del 95,5 %.

2.3. Suposant, doncs, que hàgim aplicat la prova correcta, podem dir que, amb un nivell de confiança del 95,5 %, existeix correlació significativa entre les puntuacions de la prova de relacions lògiques i les puntuacions de la prova de càlcul mental. Com que la correlació és positiva, podem dir que els alumnes que treuen una puntuació elevada en la prova de relacions lògiques tendeixen a treure també una puntuació elevada en la prova de càlcul mental, i que els alumnes que treuen puntuacions baixes en la prova sobre relacions lògiques, treuen normalment puntuacions baixes també en la prova de càlcul mental.

Exercici 3

3.1. Apliquem la prova del coeficient de correlació de Pearson.

3.2.

Cas	Habilitats socials	Grau d'acceptació	XY	X^2	Y^2
1	8	10	80	64	100
2	7	12	84	49	144
3	6	9	54	36	81
4	5	15	75	25	225
5	9	14	126	81	196
6	3	6	18	9	36
7	4	7	28	16	49
8	5	11	55	25	121
9	7	12	84	49	144
10	8	13	104	64	169
11	8	10	80	64	100
12	4	6	24	16	36
13	9	12	108	81	144
14	7	7	49	49	49
15	8	12	96	64	144
16	6	9	54	36	81
17	5	10	50	25	100
18	5	14	70	25	196
19	3	5	15	9	25
20	2	4	8	4	16
21	4	8	32	16	64
22	5	9	45	25	81
23	8	10	80	64	100
24	6	9	54	36	81
25	8	11	88	64	121
26	9	12	108	81	144
27	10	10	100	100	100
28	9	9	81	81	81
29	8	11	88	64	121
30	3	7	21	9	49
	$\Sigma X = 189$	$\Sigma Y = 294$	$\Sigma YX = 1959$	$\Sigma X^2 = 1331$	$\Sigma Y^2 = 3098$

$r = 0{,}612$

r_c (g. ll. 28; $\alpha = 0{,}05$) = 0,361

r_c (g. ll. 28; $\alpha = 0{,}01$) = 0,463

En ambdós casos la *r* que hem observat és més gran que la *r* de les taules. Per tant, podem dir que tant per a un marge d'error de 0,05 com per a un marge d'error de 0,01, hi ha correlació significativa entre aquestes dues variables.

3.3. Per tant, podem dir que, tant amb un nivell de significació del 95 % com amb un nivell de significació del 99 %, hi ha correlació entre les proves d'habilitats socials i d'acceptació per part del grup d'iguals en nois i noies que fan segon d'ESO. El fet que l'índex surti positiu significa que una puntuació alta en habilitats socials implica una puntuació alta en el grau d'acceptació. Generalitzant una mica més, podríem dir que, com més nivell d'habilitats socials, més grau d'acceptació per part del grup d'iguals. I a l'inrevés, com menys nivell d'habilitats socials, menys nivell d'acceptació per part del grup d'iguals.

3.4. Es pot buscar la recta de regressió perquè ja hem vist que hi ha correlació significativa.

$$b_{yx} = \frac{N\Sigma XY - \Sigma X\Sigma Y}{N\Sigma X^2 - (\Sigma X)^2} = \frac{30 \cdot 1959 - (189 \cdot 294)}{(30 \cdot 1331) - (189)^2} =$$

$$= \frac{58770 - 55566}{39930 - 35721} = \frac{3204}{4209} = 0{,}761$$

$a_{yx} = \overline{Y} - b_{yx}\,\overline{X} = 9{,}8 - 0{,}761(6{,}3) = 5{,}01$

La recta de regressió serà $\boldsymbol{Y = 5{,}01 + 0{,}761X}$

Per calcular l'error de predicció necessitem la desviació típica de *Y*, que en aquest cas és 2,73.

$\sigma_{y'} = \sigma_y \sqrt{1-r^2} = 2{,}73 \sqrt{1 - 0{,}612^2} = 2{,}16$

2,16 és menor que la desviació típica de la variable *Y*, per la qual cosa la predicció que podem fer és bastant ajustada.

3.5. Per fer la predicció cal substituir el valor de *X* a la recta de regressió. En aquest cas la recta és:

$\boldsymbol{Y = 5{,}01 + 0{,}761X}$ $\quad Y = 5{,}01 + 0{,}761 \cdot 6{,}5 = 9{,}96$

El valor $Y = 9{,}96$ seria la predicció exacta, és a dir, és el valor que es troba damunt la recta quan $X = 6{,}5$. Però hem de tenir en compte l'error típic de predicció que podem cometre en aquest cas.

9,96 + 2,16 =12,12
9,96 – 2,16 = 7,8

Per tant, en aquest cas, un alumne que hagi tret una puntuació de 6,5 en la prova d'habilitats socials és probable que obtingui una puntuació d'entre 7,8 i 12,12 en la prova de grau d'acceptació per part del grup d'iguals.

Exercici 4

4.1. S'ha aplicat el coeficient de correlació de Pearson.

4.2. S'ha obtingut un índex de $r = 0{,}819$. Per interpretar-lo hem de buscar el valor de r crítica.

r_c (g. ll. 98; a = 0,05) = 0,205
0,205 < 0,819

Per tant, podem concloure que la correlació entre les dues variables és significativa. Es tracta d'una correlació positiva. Si mirem els valors de Guilford, observem que l'índex és molt alt; per tant, la relació és molt intensa entre les dues variables. Dit d'una altra manera, amb un nivell de confiança del 95 % podem afirmar que hi ha una correlació positiva entre les puntuacions de destreses socials i comunicatives i les de destreses de vida personal en les persones adultes amb discapacitat psíquica. Com que la correlació és positiva, una puntuació alta en la prova de destreses socials i comunicatives es correspon amb una puntuació alta en la prova de destreses de vida personal, i a l'inrevés, una puntuació baixa en la prova de destreses socials i comunicatives s'associa a puntuacions baixes de la prova de destreses de vida personal.

4.3. $r^2 = 0{,}671$

Significa que el 67,1 % de la variació de les puntuacions de la prova de destreses de vida personal s'explica per la variable "destreses socials i comunicatives"; per tant, hi ha una elevada dependència entre elles.

4.4. L'equació de la recta de regressió és $\boldsymbol{Y = 9{,}015 + 0{,}843X}$.

4.5. Cal substituir la X de l'equació de la recta pel valor 20.

$Y = 9{,}015 + 0{,}843 \cdot 20 = 25{,}86$

Fixem-nos, però, que el valor 25,86 seria el valor de l'ordenada que es troba sobre la recta de regressió. Cal recordar que per fer prediccions hem de buscar l'error típic de predicció. Per això necessitem la desviació típica de la variable Y (destreses de vida personal), que és 11,14:

$$\sigma_{y'} = \sigma_y\sqrt{1\text{-}r^2} = 11{,}14\sqrt{1 - 0{,}819^2} = 6{,}39$$

25,86 + 6,39 = 32,25
25,86 – 6,39 = 19,47

Per un valor de 25,86 de destreses socials i comunicatives podem trobar valors entre 19,47 i 32,25 en destreses de vida personal.

Exercici 5

Per a un marge d'error del 0,05 i 237 graus de llibertat, el valor de la *r* crítica és 0,195. Mirant la taula observem que totes les correlacions són significatives i de caràcter positiu. Això vol dir que una puntuació alta de destreses socials i comunicatives es relaciona amb puntuacions altes de destreses motores, de destreses de vida personal i de destreses de vida en la comunitat. I així amb totes les relacions. La correlació més elevada és la correlació entre la puntuació de destreses de vida personal i la puntuació de destreses motores ($r = 0,856$). La resta són correlacions intenses. Això ens indica que hi ha una elevada dependència entre aquests tipus de destreses en les persones adultes amb discapacitat psíquica.

Exercici 6

Mirant les taules, per a = 0,05 i 237 graus de llibertat la *r* crítica és 0,195. Ens cal mirar quins valors de *r* són més grans que aquest 0,195. A la taula següent posem els índexs que indiquen una correlació significativa per a un nivell de confiança del 95 %.

Correlacions significatives entre índexs de problemes de conducta i puntuacions de les escales de conducta adaptativa

	Índex intern	**Índex extern**	**Índex asocial**
Destreses motores		$r = -0,234$	
Destreses socials i comunicatives	$r = 0,313$		
Destreses de vida personal		$r = -0,201$	
Destreses de vida en comunitat	$r = 0,269$		

Observem que les correlacions entre les variables referides a conducta adaptativa i l'índex extern de problemes de conducta resulten negatives. A la taula següent apareix la interpretació que podem fer per cada una d'aquestes correlacions.

Hem marcat en negreta les correlacions que resulten més significatives segons els resultats de les proves. Així, pel que fa a la correlació entre les puntuacions en destreses motores i els diferents índexs de problemes de conducta, observem que la relació més significativa es dóna entre les destreses motores i l'índex extern de problemes de conducta, de manera que un nivell alt en destreses motores es correspon amb un nivell alt en els problemes de conducta relacionats amb l'índex extern, que —recordem-ho— engloba l'heteroagressivitat, la destrucció d'objectes i la conducta disruptiva.

Correlació entre índexs de problemes de conducta i puntuacions de les escales de conducta adaptativa. Interpretació.

	Índex intern (c. autolesiva, hàbits atípics i falta d'atenció)	**Índex extern (heteroagressivitat, destrucció d'objectes i c. disruptiva)**	**Índex asocial (c. social ofensiva i conducta no col·laboradora)**
Destreses motores		La relació que s'obté és negativa, és a dir, les puntuacions altes en la variable "destreses motores" es relacionen amb puntuacions baixes en la variable "índex intern". Per tant, un **nivell alt de destreses motores** es correspon amb un **nivell alt també dels problemes de conducta** inclosos en aquest índex**, i un nivell baix de destreses motores** tendeix a correspondre's **amb nivells baixos dels problemes de conducta** inclosos en aquest índex**.**	
Destreses socials i comunicatives	Un **nivell alt de destreses socials i comunicatives** es relaciona amb un **nivell baix de problemes de conducta** d'aquest grup.		
Destreses de vida personal		La *r* és negativa. Per tant, una puntuació alta en destreses de vida personal es correlaciona amb puntuacions baixes de l'índex extern, la qual cosa significa que **nivells alts de destreses de vida personal** es corresponen amb **nivells alts de problemes de conducta** d'aquest grup (ja que una puntuació baixa indica l'existència d'altres problemes de conducta).	
Destreses de vida en comunitat	**Com més destreses de vida en comunitat menys problemes de conducta** d'aquest grup.		

En el cas de la relació entre les puntuacions de destreses socials i comunicatives i els diferents índexs de problemes de conducta, trobem que únicament hi ha relació entre aquests tipus de destreses i l'índex intern. Per tant, un nivell alt en destreses socials i comunicatives es relaciona amb un nivell baix de problemes de conducta vinculats amb l'índex intern (conducta autolesiva, hàbits atípics, retraïment i falta d'atenció).

Quant a les destreses de vida personal, es correlacionen amb l'índex extern de problemes de conducta, la qual cosa significa que nivells alts de destreses de vida personal es corresponen amb nivells alts en els problemes de conducta inclosos en l'índex extern (heteroagressivitat, destrucció d'objectes i conducta disruptiva).

Pel que fa a les destreses de vida en comunitat, l'única correlació significativa la trobem entre les puntuacions d'aquest tipus de destreses i l'índex intern de problemes de conducta. Per tant, un nivell alt de destreses de vida en comunitat es correspon amb una puntuació alta en l'índex intern, i recordem que puntuacions altes en aquest índex signifiquen menys problemes de conducta autolesiva, conductes atípiques i falta d'atenció.

En conjunt observem que els nivells més alts de destreses dels diferents tipus es corresponen amb nivells baixos dels problemes de conducta relacionats amb l'índex intern, és a dir, sembla que tenir un nivell més alt de conducta adaptativa implica menor nivell de problemes de conducta relacionats amb la conducta autolesiva, els hàbits atípics i la falta d'atenció.