Correlació i regressió lineals

Per començar aquestes activitats, suposarem que ja s'han representat gràficament núvols de punts a partir de dues variables numèriques i que s'han vist quina forma tenen aquests núvols segons el tipus de correlació lineal. Si fos necessari, es podrien fer servir les construccions de GeoGebra aquí descrites per practicar més aquests aspectes gràfics.

Mesura de la correlació lineal

Fins ara, has vist que segons la forma del núvol de punts, pots intuir el tipus de correlació entre dues variables. Però aquest mètode només és aproximatiu i depèn de l'escala del gràfic. Necessitem **quantificar** el grau de correlació lineal per tal de comparar millor entre diferents situacions.

Un exemple de la insuficiència del mètode gràfic pot ser el següent:

Exemple 1

En una taula tenim els següents valors de dues variables i volem estudiar la seva correlació:

X	9	3	6	7	8	11	10
Y	6	3	2	4	2	3	5

Accedeix a http://geogebra.pepbujosa.info/estadistica/CorrelaRegre.htm

Fes servir la primera aplicació per representar aquestes variables.

- Entra les dades de la taula anterior a les dues primeres columnes del full de càlcul que apareix a la dreta.

Fixa't com han quedat els punts. Per poder observar-los millor,

- Clica amb el botó dret del ratolí a l'interior de la finestra gràfica i tria l'opció **Mostra tots els objectes**.
- Desplaça la zona gràfica amb l'eina corresponent.

Tot seguit, canvia l'escala de l'eix d'ordenades.

- Tria, prement el botó dret del ratolí, l'opció **EixX:EixY | 1:2** i després **1:5**.

Observa com canvia l'aspecte del núvol de punts amb un sol canvi d'escala. Prova-ho amb més canvis d'escala.

Activitat 1

Comenta les diferents tipus de correlació que pots intuir segons quina escala fas servir.

Sembla clar que l'aspecte gràfic del núvol de punts és insuficient per valorar la correlació entre dues variables.

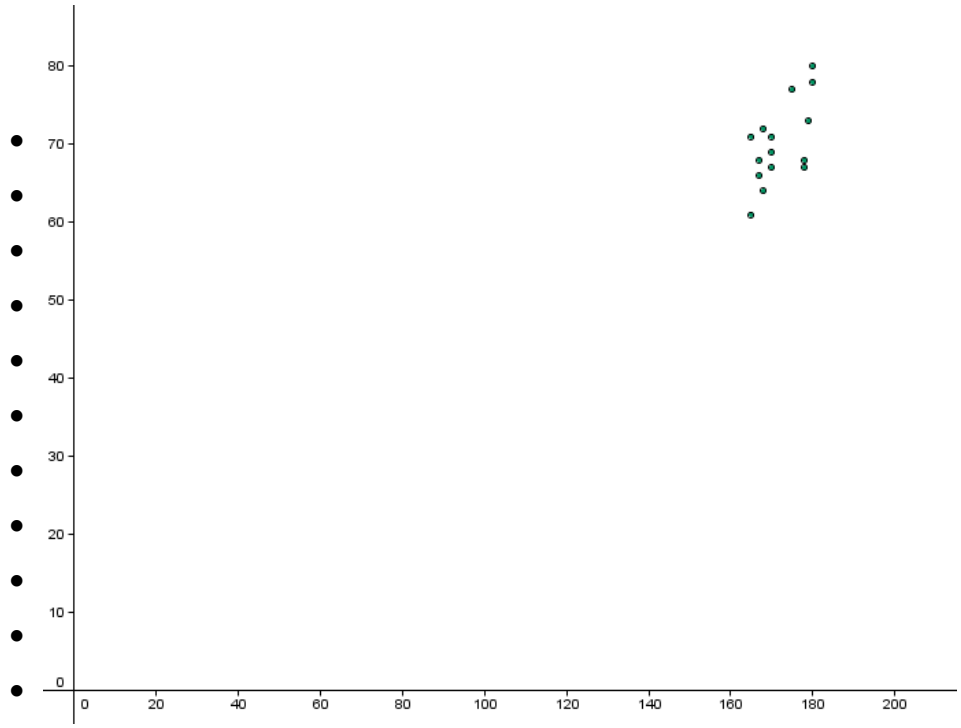
Exemple 2

Les mesures conjuntes de les variables alçada, en cm, i pes, en kg, d'un col·lectiu de 15 persones, són:

Alçada	167	179	165	170	180	170	168	180	165	168	167	178	175	170	178
Pes	66	73	61	67	78	69	64	80	71	72	68	67	77	71	68

- Prem el botó **Inici** per tornar a començar.

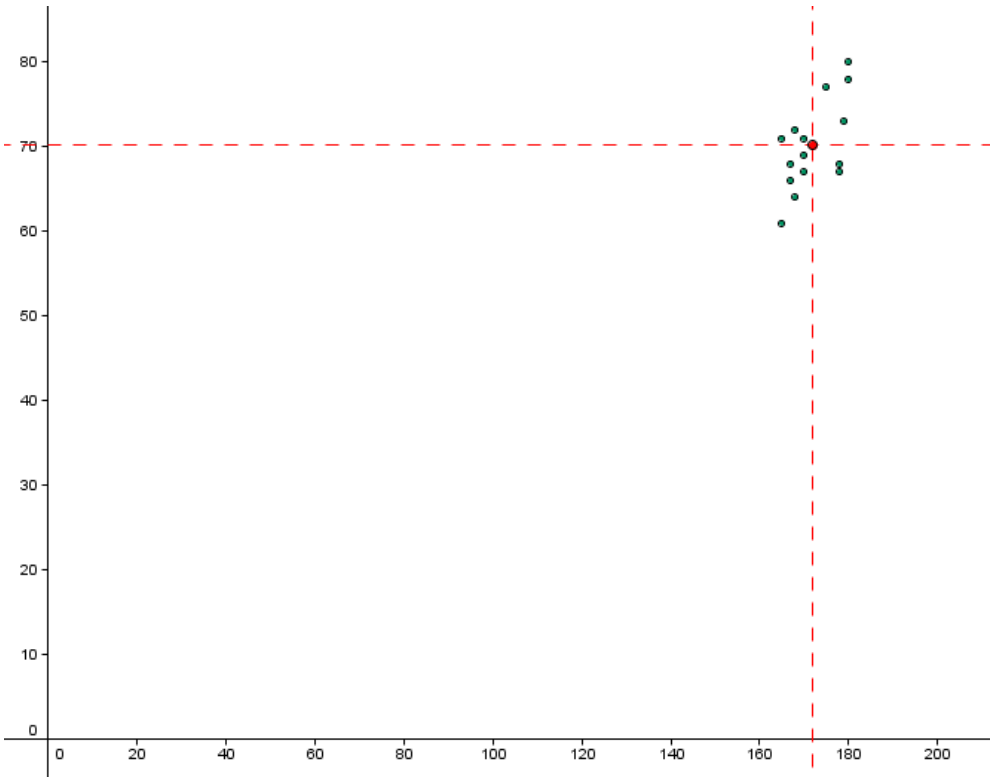
  Atenció! Cada vegada que premis aquest botó, s'esborraran totes les dades introduïdes. Ves en compte!
- Copia aquestes dades que trobaràs a partir de la columna H del full de càlcul.
- Enganxa-les a partir de la cel·la A2.
- Fes clic a la zona gràfica i, amb el botó dret, tria l'opció **Mostra tots els objectes**.
- Juga amb les eines de **Zoom** i de **Desplaçament de la zona gràfica** de la part superior de la finestra fins que et quedi un gràfic similar a:



- Activa la casella **Punt mitjà** i **eixos** . Els eixos que han aparegut en el gràfic es tallen en el **punt mitjà**, les coordenades del qual són les mitjanes de cada una de les variables.

Punt mitjà = $(\overline{x}, \overline{y})$

El gràfic ha de ser semblant a :



Atenció! ***Es convenient que feu una còpia de tota aquesta informació. Per fer-ho:***

- ***Fes doble clic a l'interior de la finestra gràfica. T'apareixerà una finestra de GeoGebra amb les dades i gràfics anteriors.***
- ***Accedeix al menú Fitxer | Anomena i desa...***
- ***Dóna un nom i tria un lloc on desar-ho.***
- ***Tanca la finestra de GeoGebra.***

Activitat 2

a) La forma del núvol de punts, indica algun tipus de correlació?

b) Quines són les coordenades del punt mitjà? (*Activa l'opció Paràmetres 1*)

c) En quins quadrants, respecte als eixos que passen pel punt mitjà, estaran situats la majoria de punts d'un núvol si la correlació lineal és directa?

I si és inversa?

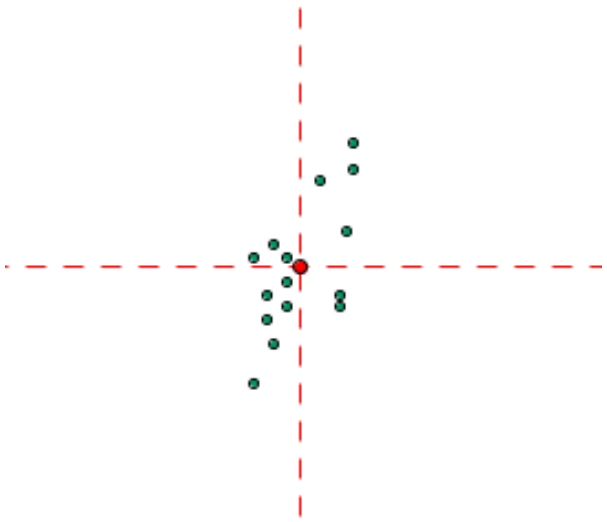
d) Fixa't que tots els punts situats en el primer quadrant corresponen a persones que tenen un alçada i un pes superiors a les respectives mitjanes. Analitza, per als altres quadrants, com són l'alçada i el pes de les persones corresponents, respecte a les mitjanes de cada variable.

e) Si calculem el producte $(x_i - \bar{x})(y_i - \bar{y})$ per a cada punt, quin signe tindrà en cada quadrant?

Amb el producte de desviacions de l'apartat anterior, comencem a caracteritzar numèricament els punts del núvol segons la seva situació respecte a les mitjanes de les variables.

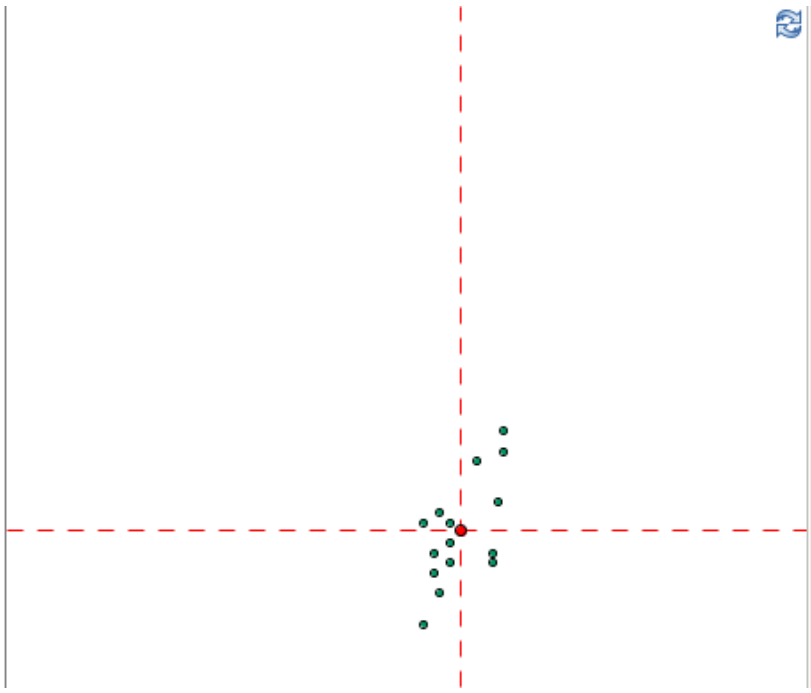
Activitat 3

La taula i el gràfic següent fan referència a l'exemple anterior. A continuació, completaràs la taula amb els productes de les desviacions respecte a les mitjanes i la seva suma. Per fer-ho, pots aprofitar el full de càlcul de GeoGebra



Alçada	Pes	Productes
167	66	
179	73	
165	61	
170	67	
180	78	
170	69	
168	64	
180	80	
165	71	
168	72	
167	68	
178	67	
175	77	
170	71	
178	68	
$\sum(x_i-\bar{x})(y_i-\bar{y})$=		

- Fes que es vegin més columnes del full de càlcul, Fent més gran la finestra que el conté.
- Entra a la cel·la D1 la paraula **Productes**.
- Entra a la cel·la D2 la fórmula **=(A2-xx)*(B2-yy)**. Amb aquesta fórmula, estàs calculant $(x_1-\bar{x})(y_1-\bar{y})$ ja que *xx* i *yy* són les variables utilitzades per a la mitjana de x i la mitjana de y, respectivament.
- Copia aquesta fórmula a la resta de la columna fins a la fila 16.
- Entra a la cel·la E1 la paraula **Suma**.
- Entra a la cel·la E2 la fórmula **=Suma[D2:D16]**. Així haureu calculat $\sum(x_i-\bar{x})(y_i-\bar{y})$



	A	B	C	D	E
1	X	Y	Punts	Productes	Suma
2	167	66	(167, 66)	20.67	264
3	179	73	(179, 73)	20.07	
4	165	61	(165, 61)	63.93	
5	170	67	(170, 67)	6.27	
6	180	78	(180, 78)	62.93	
7	170	69	(170, 69)	2.27	
8	168	64	(168, 64)	24.53	
9	180	80	(180, 80)	78.93	
10	165	71	(165, 71)	-6.07	
11	168	72	(168, 72)	-7.47	
12	167	68	(167, 68)	10.67	
13	178	67	(178, 67)	-18.8	
14	175	77	(175, 77)	20.6	
15	170	71	(170, 71)	-1.73	
16	178	68	(178, 68)	-12.8	
17	?	?	(?, ?)		

Si ara passes la fletxa del cursor per sobre dels punts del gràfic, veuràs com s'il·luminen al full de càlcul les coordenades dels punts i pots observar el signe del producte corresponent.

Per a quins punts el producte és negatiu? Per què?

Per a quins punts el producte és sempre positiu? Per què?

Observa el núvol de punts i digues si, més aviat, la correlació és directa o inversa.

Això significa que a altures més grans corresponen pesos ..... grans

Quina relació creus que hi ha entre el signe de la suma $\sum(x_i - \bar{x})(y_i - \bar{y})$ i el tipus de correlació (directa o inversa)? Per què?

Desplaça alguns punts i observa com canvia la suma. Pots aconseguir que sigui negativa? Com serà la correlació en aquest cas?

Atenció! Després de desplaçar els punts, si vols tornar a la situació anterior <u>no actualitzis</u> la pàgina del navegador ni premis el botó d'Inici perquè s'esborrarà tot i hauràs de tornar a entrar les dades.

Fes una còpia com a l'apartat anterior.

Activitat 4

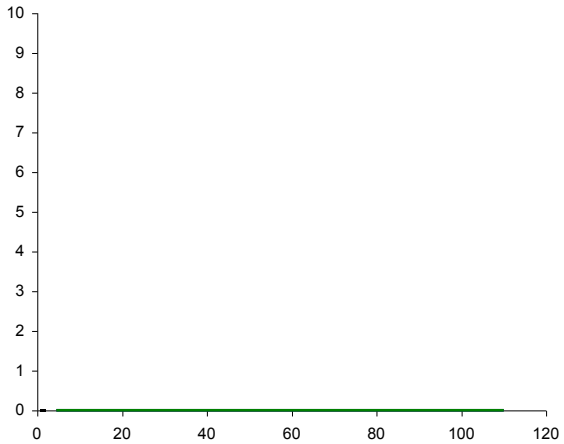
Una persona fa servir diàriament el cotxe per treballar. Cada dia passa per zones urbanes i per trams de carretera, i es pot trobar amb circulació lenta o ràpida. Cada matí surt de casa amb 10 litres de gasolina en el dipòsit i per estudiar si hi algun tipus de correlació entre l'espai recorregut i el nivell de gasolina del dipòsit en acabar cada jornada, ha elaborat la taula següent:

Espai recorregut (Km)	40	70	5	60	100	90	80	45	50
Nivell dipòsit (litres)	5.3	5.5	9.5	5	3.5	3	3.1	5.6	4.5

a) Representa gràficament el núvol de punts, els eixos que passen per al punt mitjà i comenta el tipus de correlació que t'indica. (*Ara sí que has de prémer el botó **Inici***)

b) Fes una taula per als productes de desviacions $(x_i - \bar{x})(y_i - \bar{y})$ i calcula la seva suma $\sum (x_i - \bar{x})(y_i - \bar{y})$.

c) Omple la taula següent i copia el gràfic que t'ha sortit.



Espai	**Nivell**	**Productes**
40	5,3	
70	5,5	
5	9,5	
60	5	
100	3,5	
90	3	
80	3,1	
45	5,6	
50	4,5	
$\sum (x_i - \bar{x})(y_i - \bar{y})$=		

d) Explica la raó per la qual la suma $\sum(x_i - \bar{x})(y_i - \bar{y})$ surt negativa.

Covariància

La suma dels productes de desviacions respecte a la mitjana $\sum(x_i - \bar{x})(y_i - \bar{y})$ sembla que pot ser un bon paràmetre per reconèixer numèricament si hi ha correlació lineal directa o inversa entre dues variables. Si surt positiva hi ha correlació lineal directa i si surt negativa tenim correlació lineal inversa.

Tot i així, aquest paràmetre encara no és prou bo per indicar el grau de correlació lineal. És fàcil veure que depèn del nombre de punts. Per això arribem a un paràmetre que ja serà més útil per mesurar el grau de correlació lineal: la **covariància**, la fórmula del qual és:

$$S_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N}$$

Com pots veure la covariància s'obté dividint la suma dels productes de les desviacions entre el nombre de punts.

Pel que hem vist abans, pots completar les frases següents:

Si $S_{xy} > 0 \Rightarrow$ Correlació lineal ……

Si $S_{xy} < 0 \Rightarrow$ Correlació lineal ……

Activitat 5

Utilitza la fórmula i calcula les covariàncies de les variables de les activitats 3 i 4.

L'aplicació de GeoGebra que has fet servir calcula directament la covariància i la podràs veure activant la casella **Paràmetres 2**. Així podràs comprovar com varia la covariància segons la distribució del núvol de punts.

Activitat 6

A partir dels valors de la taula següent podràs comprovar com varia la covariància.

X	9	3	6	7	8	11	10
Y	6	3	2	4	2	8	5

- Prem el botó **Inici**.
- Entra aquestes dades en les dues primeres columnes del full de càlcul.
- Modifica les escales dels eixos com ja has fet abans.
- Activa **Punt mitjà** i **Eixos punt mitjà**.
- Activa la casella i apunta la covariància que s'ha calculat a partir d'elles: S_{xy} =
- És coherent el signe de la covariància amb la forma del núvol de punts? Per què?

- Modifica alguns punts per tal que la correlació lineal sigui inversa i observa com canvia la covariància.
- Col·loca, aproximadament, els punts en forma de circumferència. Quina covariància té aquesta distribució? Com ho justifiques?

Coeficient de correlació lineal

Pel que hem vist fins ara, la covariància sembla un bon paràmetre per mesurar el grau i el tipus de correlació lineal entre dues variables. No obstant, en aquest apartat veuràs les seves limitacions i calcularàs un nou paràmetre més precís: el **Coeficient de correlació lineal**.

Activitat 7

Recorda que la covariància de les variables Alçada (en cm) i Pes (en kg) de les activitats 3 i 5 era S_{xy} = 17.6

a) Calcula la covariància per a les mateixes variables però considerant l'alçada en polzades i el pes en lliures. Aquí tens les dades amb aquestes unitats. Les trobaràs a partir de la columna J del full de càlcul.

Alçada (p)	65,7479	70,4723	64,9605	66,929	70,866	66,929	66,1416	70,866	64,9605	66,1416	65,7479	70,0786	68,8975	66,929	70,0786
Pes (l)	145,53	160,965	134,505	147,735	171,99	152,145	141,12	176,4	156,555	158,76	149,94	147,735	169,785	156,555	149,94

$$S_{xy=}$$

b) Observa el núvol de punts i compara'l amb el que sortia amb les variables en cm i kg

Com pots veure, la covariància té el defecte que depèn de les unitats. Per a un mateix núvol de punts pot donar diferents covariàncies si es canvien les unitats en què estan expressades les variables.
Això fa que sigui necessari considerar un nou paràmetre, que ja serà el definitiu, per a quantificar el grau de correlació lineal que hi ha entre dues variables.

Aquest paràmetre s'anomena **Coeficient de correlació lineal** i s'obté dividint la covariància pel producte de les desviacions estàndard:

$$r = \frac{S_{xy}}{s_x \cdot s_y}$$

Si calculem el coeficient de correlació lineal per les variables Alçada - Pes en les diferents unitats que has treballat en les activitats anteriors, surt:

Alçada - Pes expressades en centímetres - quilograms (activitat 3)

$$r = \frac{S_{xy}}{s_x \cdot s_y} = \frac{17.6}{5{,}4772 \times 5{,}0842} = ...$$

Alçada - Pes expressades en polzades - lliures (activitat 7)

$$r = \frac{S_{xy}}{s_x \cdot s_y} = \frac{15{,}2787}{2{,}1564 \times 11{,}2106} = ...$$

En qualsevol cas, pots comprovar que surt aquest resultat si actives la casella **Paràmetres 2**

Aquest coeficient no varia amb els canvis d'unitats i compleix les **propietats** següents:

1. El signe del coeficient de correlació lineal r és igual al de la covariància.

És certa perquè el signe de les desviacions estàndards s_x i s_y sempre és positiu.

Per tant, tot el que ja s'ha dit sobre la relació entre el signe de la covariància i el tipus de correlació lineal és també aplicable al coeficient de correlació lineal.

2. El valor de r està sempre situat entre -1 i 1.

   És a dir que $-1 \leq r \leq 1$. Per a correlacions lineals directes "molt fortes", tindrem una *r* propera a 1. Per a correlacions lineals inverses "molt fortes", tindrem una *r* propera a -1. Per aquells casos en què no hi hagi, pràcticament correlació lineal, *r* tindrà un valor molt proper a 0.

Desplaça punts i comprova aquestes propietats.

Juguem amb el coeficient de correlació lineal

Accedeix ara a la segona aplicació que trobaràs a la mateixa pàgina. Si prems el botó **Un altre full** apareix cada vegada un nou núvol de punts. Cal que entris el valor del coeficient de correlació lineal que creus que es correspon a aquest núvol.

Recta de regressió

A partir de la necessitat de trobar funcions polinòmiques de primer grau (rectes) que relacionin dues variables amb una bona correlació lineal, hem arribat a la recta de regressió:

$$\boldsymbol{y = ax + b} \qquad \text{on} \quad a = \frac{S_{xy}}{s_x^2} \quad \text{i} \quad b = \bar{y} - a\bar{x}$$

Activitat 8

En aquesta activitat trobaràs la recta de regressió de les variables Altura i Pes de 80 persones.

- Torna a la primera aplicació.
- Prem el botó **Inici**.
- Visualitza la columna L del full de càlcul. En aquesta part del full trobaràs diferents dades físiques de 80 persones.

- Selecciona les dades de les variables Alçada i Pes (rang L2:M81). Amb el botó dret selecciona **Copia**r.
- Selecciona la cel·la A2 del full i amb el botó dret tria l'opció **Enganxa**.
- Clica sobre la zona gràfica amb el botó dret (potser dos cops) i tria l'opció **Mostra tots els objectes**.
- Fes algun zoom amb les eines corresponents.
- Activa tots les caselles.

En el gràfic t'haurà sortit el núvol de punts de les dues variables Altura i Pes i la recta de regressió que passa pel punt mitjà. Apunta a continuació els valors de les mitjanes, desviacions estàndard, covariància i coeficient de correlació lineal, així com l'expressió de la recta de regressió.

També pots fer una còpia de la finestra de la manera que s'ha explicat abans.

- Desplaça els punts situats en els quadrants 2n i 4t per tal que la correlació lineal sigui molt feble. Fixa't com varien tots els paràmetres que havies apuntat abans.

Aquí pots adonar-te de la importància que poden tenir uns pocs punts "mal col·locats" a l'hora de calcular el coeficient de correlació lineal. Són els anomenat **punts aïllats**.

- Per tornar els punts al lloc inicial, pots anar prement el botó  de la part superior dreta de la pantalla.

Activitat 9

Amb les mateixes dades de l'activitat anterior, a continuació podràs fer prediccions. Per simplificar el gràfic, desactiveu la casella **Eixos punt mitjà**.

- Activa la casella **Prediccions** que ha aparegut en activar la de la **Recta de regressió**.
- Amb l'eina de moure triada, desplaça el punt vermell que està a l'eix d'abscisses. D'aquesta manera, veuràs les imatges calculades a partir de la recta de regressió que són les prediccions.

Quin pes cal esperar que tingui una persona de 165 cm? I de 180 cm?

Fixa't que els pesos que resulten de les prediccions no coincideixen amb els corresponents a persones del núvol de punts que tenen aquestes alçades. Per tant les prediccions no tenen per què coincidir amb dades reals. Són valors de la variable *Y* necessaris per tal que el punt resultant estigui sobre la recta de regressió. Si els punts reals del núvol estan molt lluny de la recta, les prediccions són menys fiables. Això es dóna per a una correlació lineal feble

En definitiva, **com més forta és la correlació lineal, més fiables són les prediccions**.

Activitat 10

a) Fes el gràfic del núvol de punts i de la recta de regressió per a les variables Alçada i Envergadura que trobaràs en el full de càlcul.

b) Escriu, a continuació, el valor del coeficient de correlació lineal i l'expressió de la recta de regressió.

c) Escriu les prediccions resultants per al mateixos valors de l'activitat anterior.

d) Comenta els resultats en funció del diferent valor de *r* en tots dos casos.

Activitat 11

En aquesta activitat simularem un experiment real que va portar a terme Francis Galton (1822-1911) i que va tenir molt a veure amb la paraula "regressió".

- Prem el botó **Inici**.
- Copia a les dues primeres columnes les dades que trobaràs a les columnes S i T del full de càlcul. Es tracta de les alçades mitjanes d'uns pares i les dels seus fills respectius.
- Repeteix els procediments anteriors per veure bé el núvol de punts. Activa totes les caselles.

a) A partir dels resultats que tens en pantalla, es pot dir que els fills, en general són més alts que els seus pares? Per què?

b) Escriu el valor del coeficient de correlació lineal.

Què ens indica en aquest cas?

c) Fes servir la recta de regressió per calcular quina alçada esperem que tingui el fill d'uns pares de 160 cm?

Fes el mateix per a uns pares de 180cm, 190 cm i 200 cm.

d) Aquests resultats contradiuen les afirmacions dels apartats anteriors? Per què?