Statistiska samband: regression och korrelation

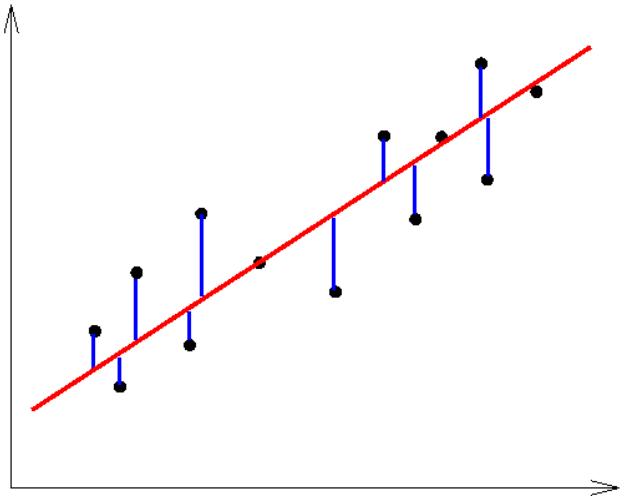
Vi ska nu gå igenom något som kallas *regressionsanalys* och som innebär att man identifierar sambandet mellan en beroende variabel (*x*) och en oberoende variabel (*y*). Olika tester utnyttjas sedan för att avgöra hur pass bra modellen är. Om modellen anses tillfredsställande, kan den s.k. regressionsekvationen användas för att förutsäga värdet på den beroende variabeln för olika värden för den oberoende variabeln

Linjär regressionsmodell

I en enkel *linjär* regressionsmodell kan sambandet mellan den beroende (*y*) och oberoende variabeln (*x*) skrivas som

$y = ax + b$

För att uppskatta värdena på parametrarna *a* och *b* använder man en metod som kallas *minstakvadratmetoden*. Titta på figuren nedan där vi i blått markerat avståndet i vertikal led mellan data och linjen $y = ax + b$. Om vi kallar avstånden d_1, d_2 osv så ska summan $d_1^2 + d_2^2 \ldots$ bli så liten som möjligt.



Vi går inte igenom i detalj hur beräkningarna går till utan visar bara hur värdena på *a* och *b* kan beräknas. *a* motsvarar ju linjens lutning och brukar betecknas med bokstaven *k* och *b* är skärningen med *y*-axeln och betecknas hos oss med bokstaven *m*.

Minstakvadrat-metoden

Om vi har *n* punkter har med koordinater (x_1, y_1) till (x_n, y_n) .och $\bar{x}$ och $\bar{y}$ är medelvärdena av *x*- respektive *y*-koordinaterna så går regressionslinjen genom punkten $(\bar{x}, \bar{y})$ med lutningen eller *k*-värdet

$$k = \frac{x_1 y_1 + \ldots + x_n y_n - n\bar{x}\,\bar{y}}{x_1^{\,2} + \ldots + x_n^{\,2} - n\bar{x}^2}$$

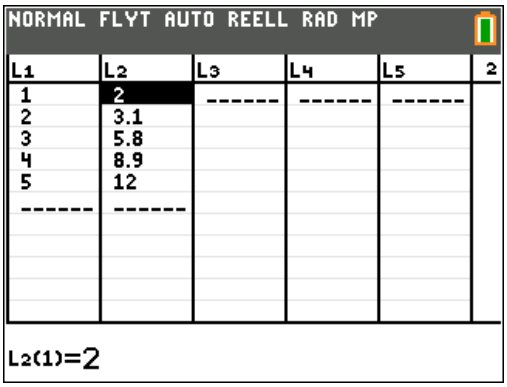
Eftersom vi vet att linjen går igenom $(\bar{x}, \bar{y})$ så kan vi skriva ekvationen som $\bar{y} = k\bar{x} + m$ vilket ger $m = \bar{y} - k\bar{x}$

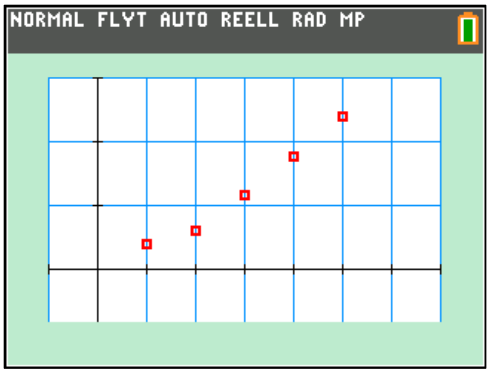
På räknaren finns en inbyggd funktion för att göra dessa beräkningar men vi ska nu kontrollera mot de uttryck för *k* och *m* som vi har ovan. Dags för ett exempel alltså.

Här är nu våra datapunkter

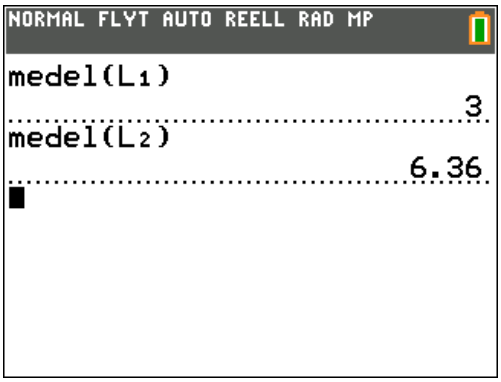
x	y
1	2,0
2	3,1
3	5,8
4	8,9
5	12,0

Vi skriver in den i räknarens statistikeditor och plottar dem i ett spridningsdiagram:





Vi kan först beräkna medelvärdena $\bar{x}$ och $\bar{y}$ i räknarens grundfönster. Tryck på [2nd][list] och välj alternativet MA. I menyn väljer man sedan 3:medel(och trycker på [enter], Då inkopieras instruktionen till grundfönstret. Punkten (3, 6,36) kommer att ligga på linjen.



Nu sätter vi igång att räkna ut *k*-värdet enligt formeln i vänstra spalten.

Vi upprepar formeln:

$$\frac{x_1y_1+\ldots+x_ny_n-n\bar{x}\,\bar{y}}{x_1^{\,2}+\ldots+x_n^{\,2}-n\bar{x}^2}$$

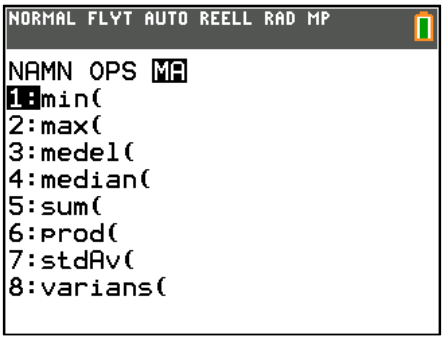
I täljaren: Vi ska först räkna ut produkten $x\cdot y$ för de fem talparen och summera dessa produkter:

$x_1y_1+\ldots+x_ny_n$

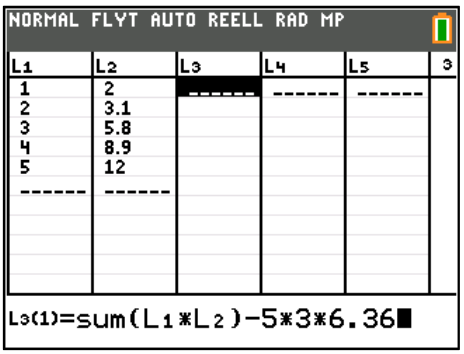
Sedan ska vi subtrahera det värdet med

$n\bar{x}\,\bar{y}$: $5\cdot3\cdot6{,}36$.

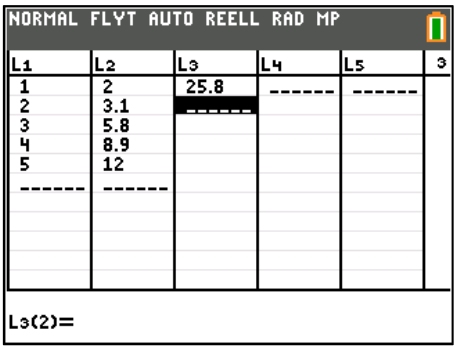
Vi gör nu detta på räknaren. Instruktionen **sum** inkopi-eras på inmatningsraden om du trycker på [2nd][list] och väljer MA i menyn högst upp. Här finns nu ett antal beräkningsverktyg som du kan använda på data i listor.



Här har vi hela beräkningsuttrycket på inmatningsraden.

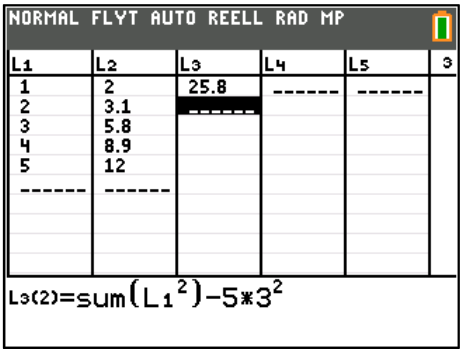


Tryck på [enter] för att slutföra beräkningen.

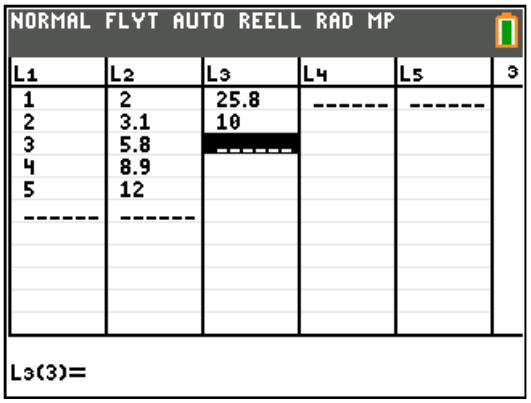


Vi får värdet 25.8

Vi gör nu på liknande sätt för uttrycket i nämnaren:



Vi trycker på [enter]:



k-värdet är nu kvoten mellan dessa tal, alltså 2,58.

För att beräkna *m*-värdet, dvs. skärningen med *y*-axeln, så kan vi nu bara sätta in värden vi känner till i uttrycket för en rät linje: $y = kx + m$.

Vi vet ju att punkten (3, 6,36) ligger på linjen och att *k*-värdet är 2,58. Vi får

$6{,}36 = 2{,}58 \cdot 3 + m$

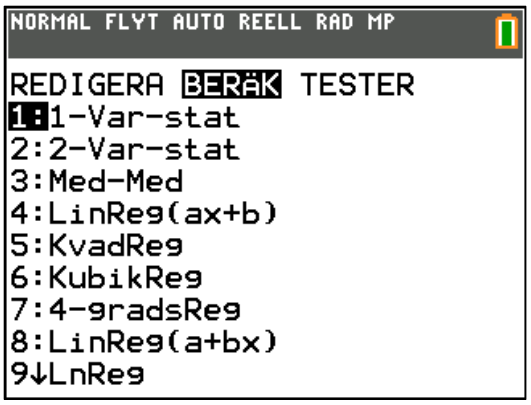
Vi får $m = 6{,}36 - 2{,}58 \cdot 3 = -1.38$

Den bäst anpassade räta linjen är alltså:

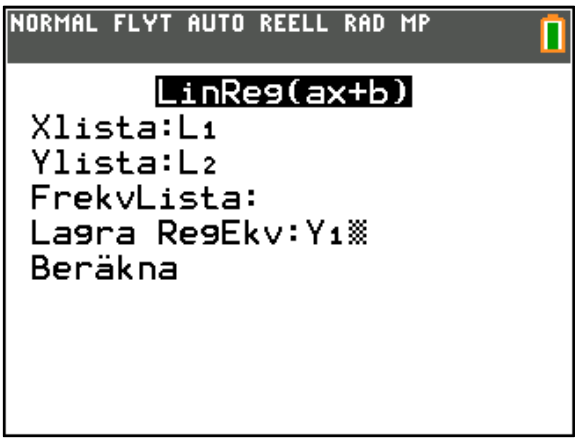
$y = 2{,}58x - 1{,}38$

Vi testar nu om det stämmer med räknarens inbyggda verktyg för linjär regression.

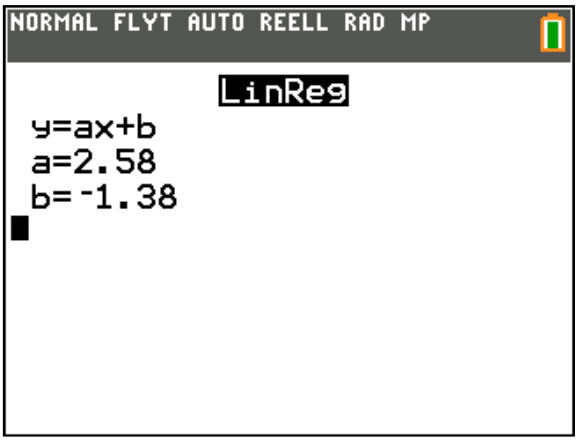
Tryck på [stat] och väljs BERÄK i menyn. Här finns nu att ett antal regressionsverktyg för olika modeller. Välj nu 4:LinReg(ax+b).



Nu fyller vi in vilka listor vi har och var regressionsekvationen ska sparas. För att lägga in Y1 trycker du först på tangenten [vars], väljer Y-VAR och sedan funktion. Välj sedan Y1 till exempel.

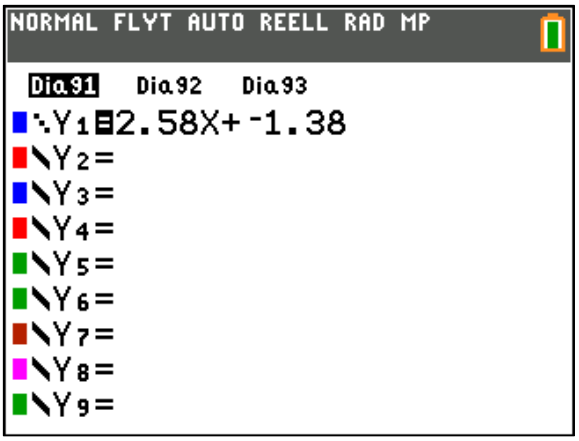


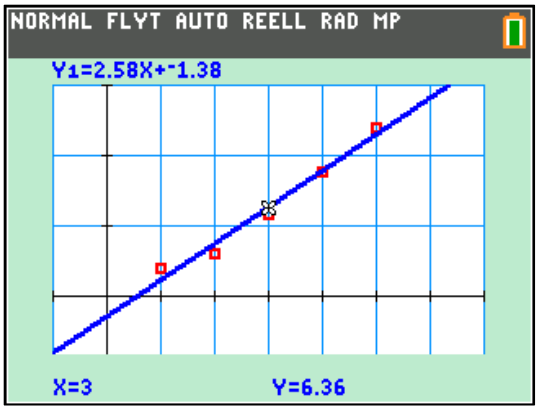
Markera Beräkna och tryck på [enter].



Vi får samma resultat.

Om du tittar i listan med funktioner (tryck på [y=]) så ligger nu vår regressionsekvation där och vi kan plotta punkterna och regressionslinjen i samma diagram.





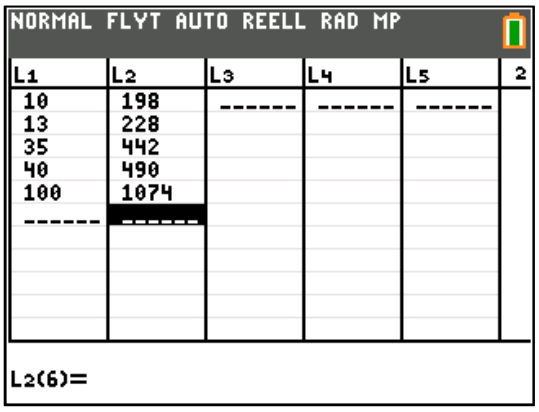
Vi ser att vi har en mycket god anpassning. Punkten (3, 6,36) ligger på linjen om vi spårar.

Vi tar ett exempel till:
Tabellen nedan visar det uppmätta trycket vid olika djup under havsytan.

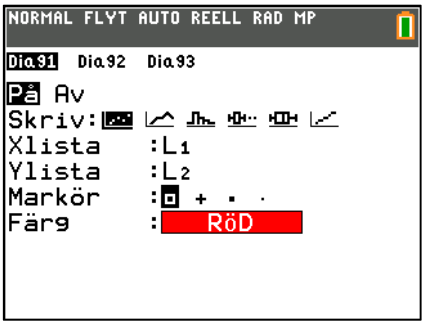
Djup (m)	10	13	35	40	100
Tryck (kPa)	198	228	442	490	1074

- Undersök om trycket är en linjär funktion av djupet.
- Ta reda på trycket på 150 m djup och vid havsytan.
- Till sista ska du ta reda på vid vilket djup trycket är 300 kPa.

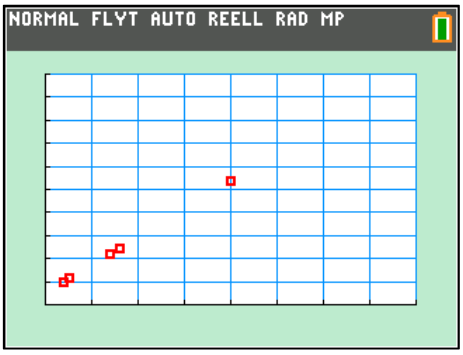
Vi matar först in data i statistikeditorn.



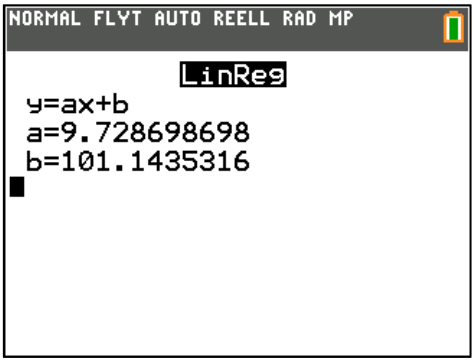
Vi gör nu precis som i första exemplet. Ställer in diagramplottningen:

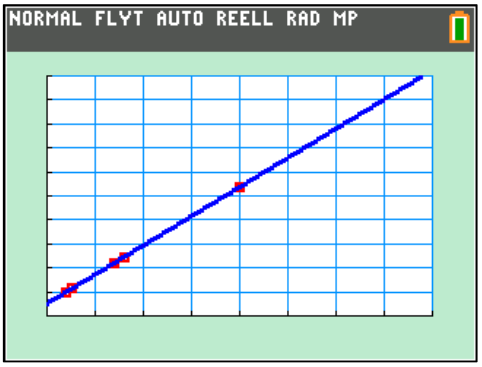


Vi måste också se till att vi har ett bra fönster för att kunna visa alla punkter.



Nu kan vi göra beräkningen och plotta den beräknade linjen.



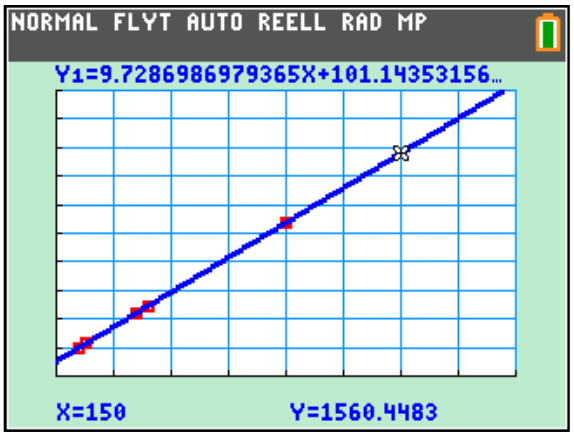


Vi ser att den beräknade räta linjen nästan helt perfekt går igenom datapunkterna.

I uppgiften skulle vi undersöka om trycket var en linjär funktion av djupet. På den frågan kan vi naturligtvis svara ja! Av grafen att döma ligger datapunkterna nästan precis på den beräknade räta linjen.

Sedan skulle vi ta reda på trycket på 150 m djup och vid havsytan, dvs. när djupet är 0 m.

Vi spårar i den räta linjen med [trace].



Trycker vid havsytan är naturligtvis ca 101kPa.

Till sist skulle vi beräkna vid vilket djup trycket var 300 kPa. Detta ger ekvationen

300 = 9,73*x* + 101. Vi får *x* = 20,5.

I båda dessa exempel har passningen varit nästan perfekt. Man kan bestämma ett mått på hur pass bra passningen är, *korrelationen*, med något som kallas för korrelationskoefficient. Den betecknas med bokstaven *r*.

Korrelation

Korrelationskoefficienten är ett mått på hur *starkt* det linjära sambandet är mellan två variabler. Värden för korrelationskoefficienten är alltid mellan -1 och + 1. Ett positivt värde på korrelationskoefficienten *r* anger att *k*-värdet för linjen är positivt och linjen lutar uppåt. Ett negativt värde på *r* innebär att k-värdet är negativt och linjen lutar nedåt.

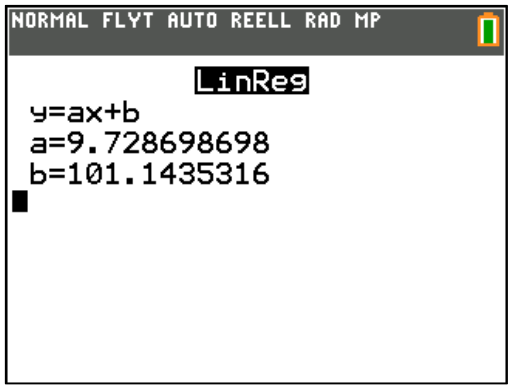
> En korrelation säger ingenting om orsakssamband, eller *kausalitet*. För att ta ett exempel, säg att vi vill uttrycka sambandet mellan *rikedom* och *lycka*, och att vi har lyckats mäta dessa företeelser i en numerisk skala. En stark positiv korrelation, till exempel 0,9, betyder då att ju rikare man är, desto lyckligare är man. Det kan även uttryckas omvänt; ju lyckligare man är, desto rikare är man.
> I det första exemplet ovan säger en stark positiv korrelation alltså inte att man är lycklig *på grund av* att man är rik. Det kan lika gärna vara så att man är rik på grund av att man är lycklig, eller att en tredje variabel (till exempel social bakgrund) orsakar både lycka och rikedom. *(Wikipedia)*

Korrelationskoefficienten mäter alltså bara graden av linjära samband mellan två variabler. Några slutsatser om en relation mellan orsak och verkan kan man inte dra.

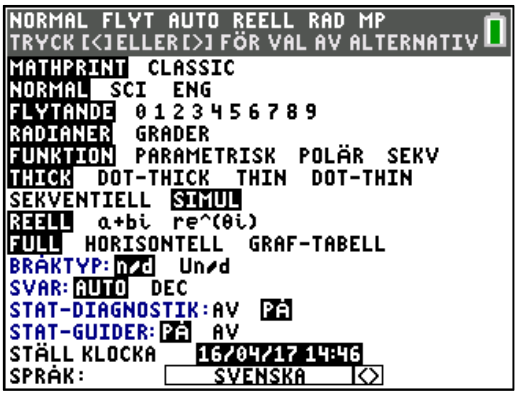
I de exemplet med uppmätt tryck under vattenytan så är det naturligtvis så att ett större vattendjup *orsakar* ett större tryck.

Vi ska nu bestämma hur starkt sambandet är i exemplet med trycket på olika vattendjup.

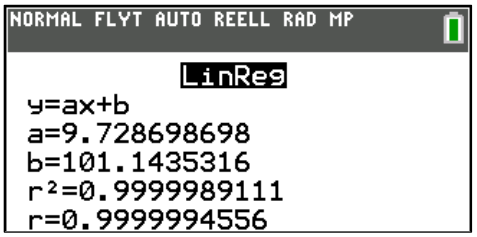
Vi fick ju fram det här resultatet:



Om vi går till räknarens lägesinställningar (tryck på [mode]) så ska du se till att STAT-DIAGNOSTIK är PÅ.



Om vi gör samma beräkning en gång till:

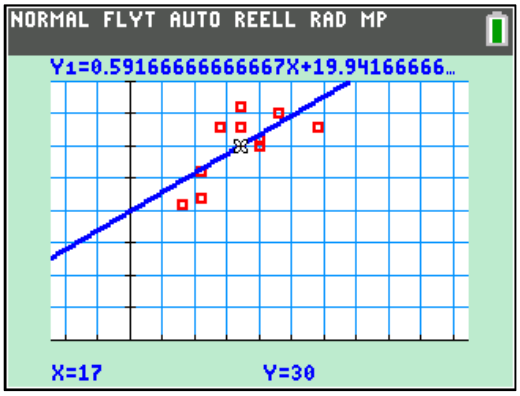


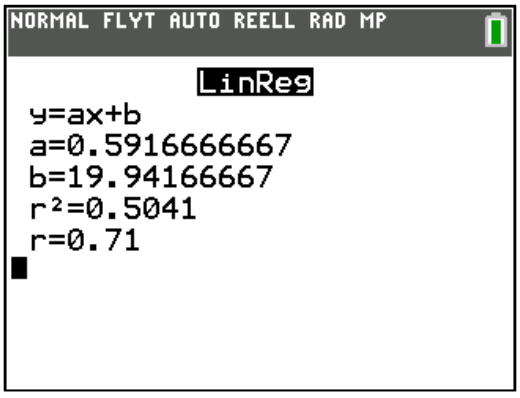
Vi får att *r* blir nästan 1. Ett nästan perfekt samband alltså. Vi tar ännu ett exempel:

Pupil	A	B	C	D	E	F	G	H	I	J
Maths mark (out of 30) *x*	20	23	8	29	14	11	11	20	17	17
Physics mark (out of 40) *y*	30	35	21	33	33	26	22	31	33	36

Det handlar alltså om resultat på ett matematik-och ett fysikprov för ett antal elever. Mata nu in dessa i räknarens statistikeditor och beräkna regressions-ekvationen och korrelationskoefficienten.

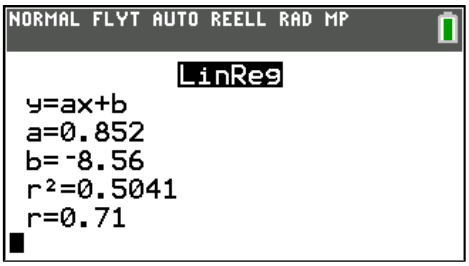
Medelvärdet i matematikprovet var 17 poäng och på fysikprovet 30 poäng. Vi ser att punkten (17, 30) ligger på regressionslinjen.

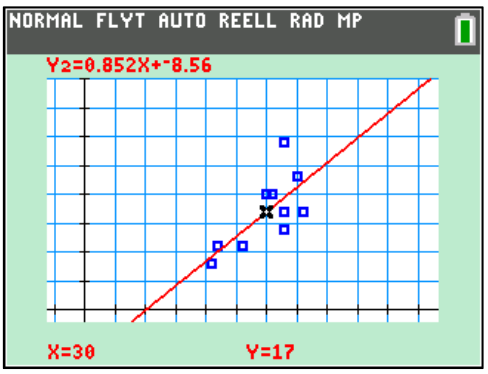




Ett tämligen starkt samband mellan resultat på matematik- och fysikprovet alltså.

Man kan också vända på axlarna och göra samma beräkningar. Ovan hade vi matematikpoäng på *x*-axeln och fysikpoängen på *y*-axeln.





Vad är orsak och verkan här?

Slutligen:
Nedan har vi 4 par av data:

- Medelvärdet för *x* är 9 för *alla* 4 paren
- Medelvärdet för *y* är 7,50 för *alla* 4 paren
- Regressionsekvationen är *y* = 0,500*x* +3,00 för *alla* fyra paren
- Korrelationskoefficienten är 0,816 för *alla* fyra paren

Analysera nu detta närmare genom att mata in och plotta alla fyra datauppsättningarna. *Vad upptäcker du?*

x1	y1	x2	y2	x3	y3	x4	y4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89