Regression och korrelation på djupet

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.

Regressionsanalys innebär ett enkelt sätt att få en kvantitativ bild över hur de två variablerna är relaterade. Om t.ex. x lätt att mäta men inte y så kan det vara värdefullt att känna till sambandet.

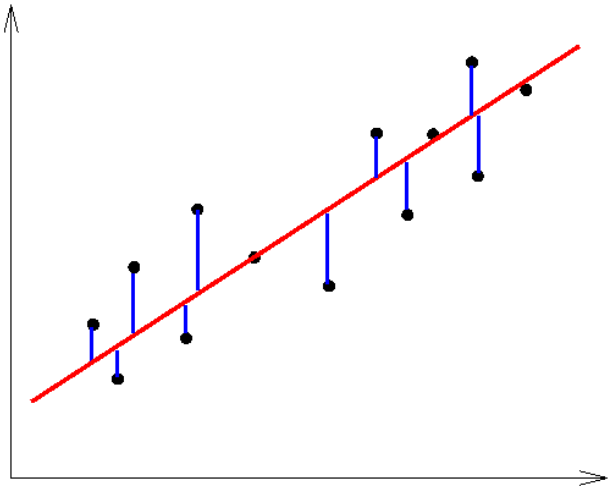
Om man sedan kanske har en idé om att det bör finnas ett samband mellan x och y så kan det vara bra att göra en analys för att bekräfta, visa och få mer klarhet om hur det ligger till.

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 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

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

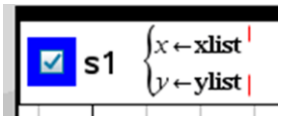
Eftersom vi vet att linjen går igenom $(\bar{x}, \bar{y})$ så kan vi skriva ekvationen som $\bar{y} = k\bar{x} + m$ vilket ger $m = \bar{y} - k\bar{x}$.

TI-Nspire har en inbyggd funktion för att göra dessa beräkningar direkt men vi ska nu kontrollera mot de uttryck för *k* och *m* som vi har ovan. Dags för ett exempel.

Här är nu våra datapunkter i appen Listor&kalkylblad. Data finns i kolumnerna B och C. I kolumnerna D och E har vi kvadraterna på våra data och i kolumn F har vi beräknat medelvärdena för x-list och y-list. De beräknas med kommandot *=mean(xlist)* resp. *=mean(ylist)*. Vi har markerat cellerna med gul färg.

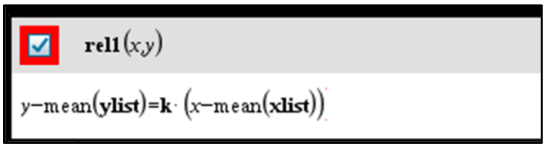
	tarader	B xlist	C ylist	D xkvad	E ykvad	F medelvärde	G
=				=xlist^2	=ylist^2		
1	1	2	12.6	4	158.76	8.5826	
2	2	6.506	28.16	42.328	792.986	35.876	
3	3	3.216	17.36	10.3427	301.37		
4	4	5.62	38.09	31.5844	1450.85		
5	5	2.3	26.93	5.29	725.225		
6	6	9	35.8	81	1281.64		
7	7	9	43.2	81	1866.24		
8	8	16	54	256	2916		
9	9	15.184	55.84	230.554	3118.11		
10	10	17	46.78	289	2188.37		

Vi visar nu först våra 10 datapunkter i ett graffönster. Se nästa sida. Punkterna ska alltså visas i ett spridningsdiagram och du matar då in data så här:

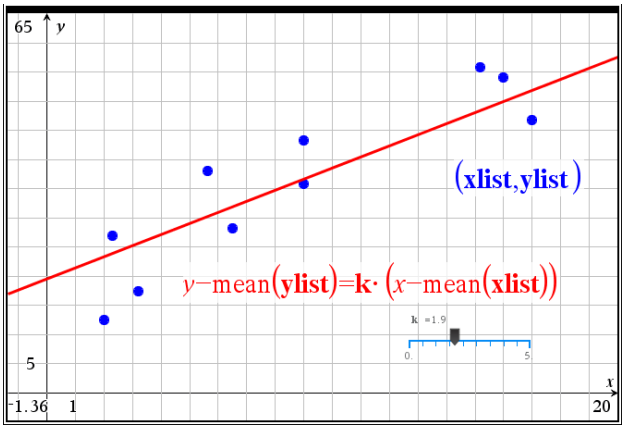


Sedan skriver vi in relationen

y-mean(ylist)=k·(x-mean(xlist))



Man kan skriva in och plotta många andra relationer än funktioner. Denna linje är visserligen en funktion men man har möjligheter att skriva in sitt samband på olika sätt.



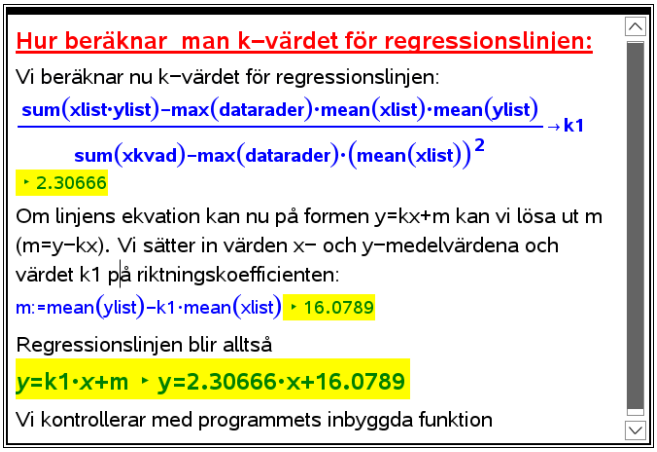
Här har vi alltså skrivit den räta linjen som går igenom punkten som har medelvärdena för x och y som koordinater. När vi skriver in denna relation så skapas samtidigt ett skjutreglage för **k**, där vi sedan kan ställa in minvärde, maxvärde och steglängd.

Om vi nu drar i reglaget ser vi hur linjen *vrider sig runt* en punkt. Denna punkt har koordinater (mean(xlist), mean(ylist). Om vi drar hela vägen från 0 till 5 så har linjen någonstans ett k-värde som är lika med k-värdet för den regressionslinje vi längre fram ska låta TI-Nspire beräkna.

Nu sätter vi igång att räkna ut *k*-värdet enligt formeln i vänstra spalten. Man öppnar en *Anteckningssida* och gör sina beräkningar enligt nedan.

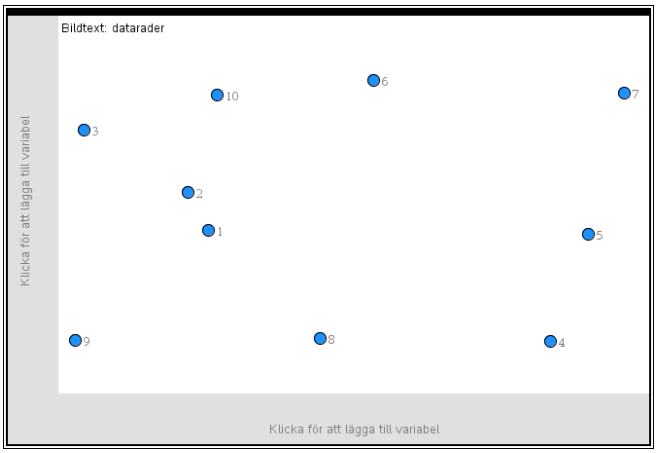
Vi upprepar först formeln för beräkning av k-värdet

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

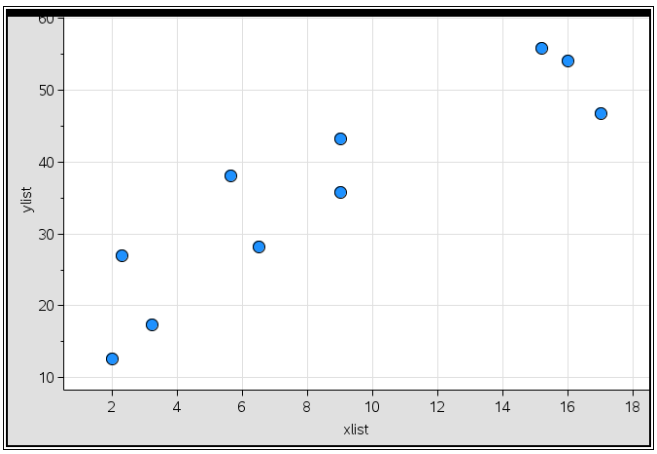


Vi testar nu om det stämmer med räknarens inbyggda verktyg för linjär regression.

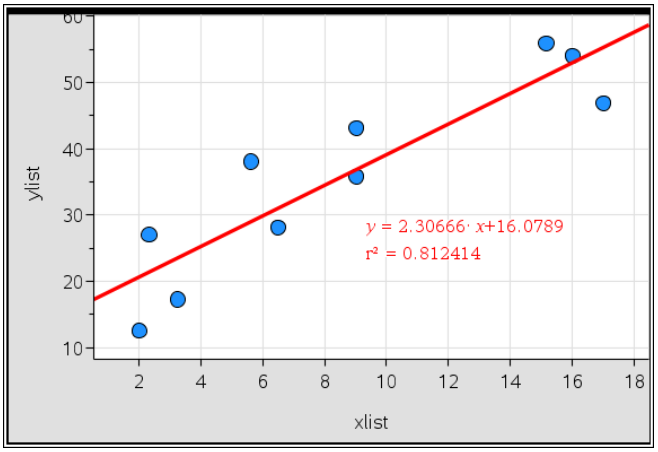
Vi öppnar alltså en Data&Statistik-sida. Då ser de ut så här:



Man väljer sedan variabler genom att klicka nedanför och på vänster sida av själva graffönstret.



Nu väljer vi Regression/Visa linjärt (mx+b) bland analysverktygen. Se nästa sida.



Vi ser att vi får regressionslinjen utritad och vi ser ekvationen i rött. Samma värden som vi fick när vi gjorde beräkningarna på tidigare sidor.

I båda dessa exempel har passningen varit nästan perfekt. Man kan bestämma ett mått på hur pass bra passningen, *korrelationen*, är med något som kallas för korrelationskoefficient. Den betecknas med bokstaven *r*. I diagrammet ovan fick vi se värdet på r^2 på skärmen.

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.

Korrelationskoefficienten definieras så här:

$$r = \sqrt{k_1 \cdot k_2}$$

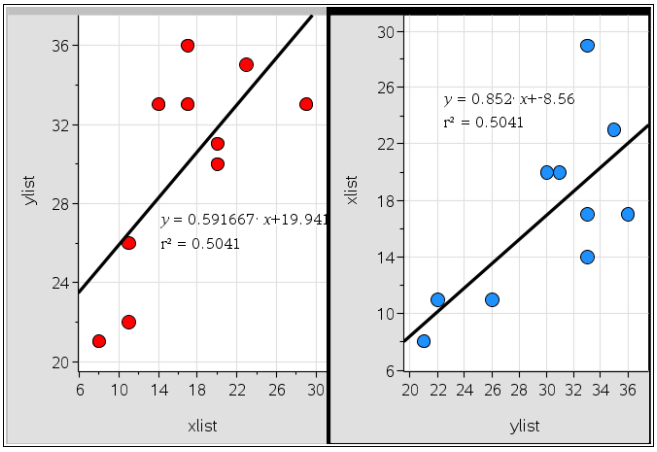
Där k_1 och k_2 är riktningskoefficienterna när data för x-listan är på den vågräta axeln, data för y-listan på den lodräta och *tvärtom*.

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.

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 matematikprov och ett fysikprov för 10 elever. Mata nu in dessa i kalkylarket och beräkna regressionsekvationen och korrelationskoefficienten. Vi gör sedan så att vi ritar två spridningsdiagram där vi vänder på axlarna:



Medelvärdet på matematikprovet var 17 poäng och på fysikprovet 30 poäng. Undersök om medelvärdespunkten (17, 30) ligger på regressionslinjen.

Vi ser att r^2 är 0,5041 vilket ger r =0,71. Ett tämligen starkt samband mellan resultat på matematik- och fysikprovet alltså. Samma resultat får vi om vi använder formeln $r = \sqrt{k_1 \cdot k_2}$:

$$r = \sqrt{0{,}591667 \cdot 0{,}852} \approx 0{,}71$$

Vad är orsak och verkan här?

Till slut: 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 $y = 0{,}5x + 3{,}0$ för *alla* fyra paren
- Korrelationskoefficienten är 0,816 för *alla* fyra paren

Vi analyserar nu detta närmare genom att och plotta alla fyra datauppsättningarna och beräknar regressionsekvationen och korrelationskoefficienten. *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