Lineære sammenhænge, residualplot og regression

Opgave 1: Er der en bagvedliggende lineær sammenhæng?

I mange sammenhænge indsamler man data som man ønsker at undersøge og afdække eventuelle sammenhænge mellem de forskellige størrelser man har data fra. Data kan f.eks. fås ved en spørgeskemaundersøgelse, et fysisk eksperiment eller fra en statistikbank, men lige meget hvorfra de kommer kan der være "støj" på data. Støjen kan f.eks. skyldes måleusikkerheder, fejl i forsøgets udførelse eller tilfældighed i de adspurgtes svar i forbindelse med spørgeskemaer. Støjen gør at data ikke er helt som man vil forvente i forhold til den bagvedliggende teoretiske sammenhæng.

x	y
0	1,4
1	4,4
2	5,5
3	9,8
4	9,4
5	13,5

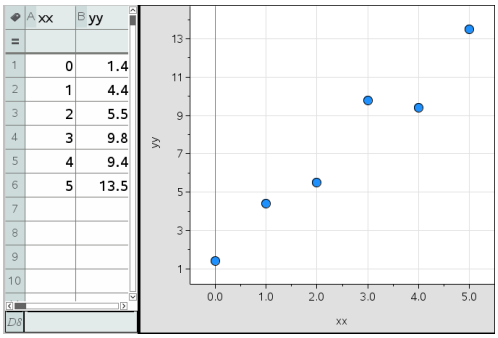
Figur 1: Data til de næste opgaver. Bemærk at her er brugt dansk decimalkomma. I Nspire skal man bruge amerikansk komma dvs. punktum.

Vi skal her undersøge hvordan man kan afdække en sådan sammenhæng.

a) **Lav et punktplot af data i tabellen til højre.**

<u>Hjælp til Nspire:</u>

- Åben et "liste og regneark" værksted, indtast data i hver sin søjle. Navngiv hver søjle *xx* og *yy* (se figur 2). Det er en dårlig vane at bruge *x* som navn da det bruges til mange andre ting, derfor vælges navnene *xx* og *yy*.
- Husk at bruge punktum som decimalkomma.
- Åben et "diagram og statistik" værksted hvor man vælger variablen *xx* på første aksen og *yy* på anden aksen.



Figur 2: Til højre data i "liste og regneark" til venstre data som punkter i "diagram og statistik" værksted.

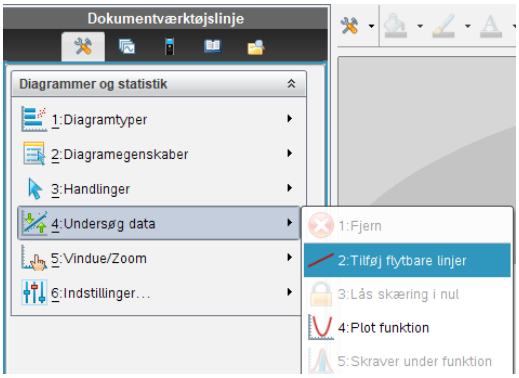
b) Beskriv med ord hvordan data ligger.

Der ser ud til at være en lineær sammenhæng mellem x og y, men støjen gør at data ikke ligger perfekt på en ret linje. Man kan forestille sig at der er nogle ideelle bagvedliggende y-værdier som ligger på en ret linje der ikke er påvirket af støjen.

c) **Tilføj en flytbar linje til grafen med punkterne**.

Hjælp til Nspire:

- Når du er i et diagram og statistik værksted, kan du finde menupunktet "tilføj flytbar linje" i menuen som vist på billedet til højre.
- Du flytter linjen op/ned ved at trække midt på den, og du kan ændre linjens hældning ved trække i dens ender.
- **Bemærk** at Nspire automatisk skriver linjens forskrift og definerer den som $m1(x)$.
- **Bemærk** man kan komme til at trække i punkterne på grafen, hvorved man også ændrer data i regnearket. I disse øvelser skal man udgå det, og hvis det sker trykker man fortryd/undo (PC: CTRL+z MAC: CMD+z).



Figur 3: Tilføj flytbarlinje til data.

d) Træk og drej linjen så den beskriver punkterne bedste muligt.

e) Konkurrer med din sidemand: Hvem kan finde den linje der bedst beskriver punkterne? Diskuter hvordan man afgør hvilken linje der bedst beskriver punkterne.

Opgave 2: Residualer - hvilken linje beskriver data bedst?

Som vi så i opgave 1 har vi brug for en måde, der kan afgøre om en linje beskriver en mænge punkter bedre end en anden linje. Vi skal i denne opgave undersøge hvordan man bedst gør det.

Vi forvente at den rette linje beskriver en bagvedliggende teoretisk sammenhæng, og vi vil derfor kalde linjens forskrift for den teoretiske model, og den kan altså skrives på formen

$$f(x) = ax + b \qquad \text{(den teoretiske model).}$$

Årsagen til at punkterne (for "rådata") ikke ligger præcis på linjen er at de er påvirket af "støj", så et godt sted at starte vores undersøgelse er at se på hvilke y-værdier vi ville forvente ud fra den teoretiske model. Vi fortsætter med at se på data fra opgave 1. Fortsæt bare i samme Nspire opgave.

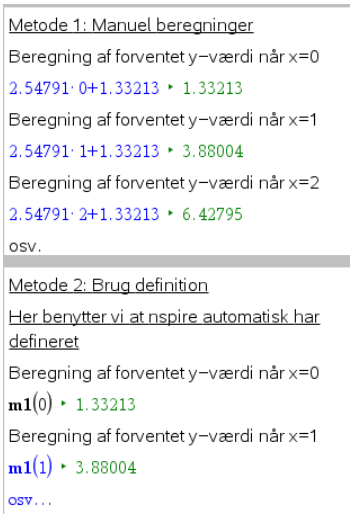
a) Beregn de forventede y-værdier ud fra forskriften for din rette linje

Hjælp til Nspire

Metode 1: Simpel beregning (øverst figur 4)

- I et note felt beregnes den forventede værdi for y når x=0 ved at bruge formlen der står ved den rette linje (opret et matematik felt og udregn)
- Foretag samme beregninger for alle x-værdierne.

Metode 2: Brug funktionens definition (nederst figur 4)



*Figur 4: Eksempel på beregninger hvor linjens ligning er y=2,54791*x+1,33213*

- Som vi bemærkede i opgave 1 gemte Nspire automatisk linjens forskrift som $m1(x)$, prøv i et matematikfelt at skrive m1(0). Kan du finde ud at lave de øvrige beregninger på denne måde?

b) Indfør de forventede y-værdier i en tabel ved siden af data. Navngiv listen y_model.

Hjælp til Nspire:

Det kan gøres utroligt let i Nspire fordi forskriften for den tilføjede linje automatisk gemmes af Nspire som $m1(x)$. Se på figur 5 hvor der er indtastet $m1(xx)$, gør det samme og se hvad der sker… Smart ikke😊

	A xx	B yy	C y_model
=			=m1(xx)
1	0	1.4	1.33213
2	1	4.4	3.88004
3	2	5.5	6.42795
4	3	9.8	8.97586
5	4	9.4	11.5238
6	5	13.5	14.0717

Figur 5: Smart beregning af de forventede værdier.

RESIDUALER: Forskellen på datas y-værdier og de y-værdier som vi forventer ud fra modellen kaldes residualerne. Man kunne også kalde residualerne afvigelsen.

c) Beregn residualerne for data i opgaven.
Vi skal altså beregn hvor meget de forventede data (y_model) afviger fra de korrekte data (yy).

Hjælp til Nspire:

Opret en ny søjle kaldet "residualer". Vi kan let få Nspire til at trække modellens y-værdier fra de y-værdierne fra "rådata", ved at indtaste det i feltet under navnefeltet hvor vi navngav søjlen "residualer". Se på figur 6 hvordan man gør.

	A xx	B yy	C y_model	D residualer	E
=			=m1(xx)	r:=yy−y_model	
1	0	1.4	1.33213		
2	1	4.4	3.88004		
3	2	5.5	6.42795		
4	3	9.8	8.97586		
5	4	9.4	11.5238		
6	5	13.5	14.0717		
7					
8					
9					

Figur 6: Beregning af residualer

D residualer:=yy−y_model

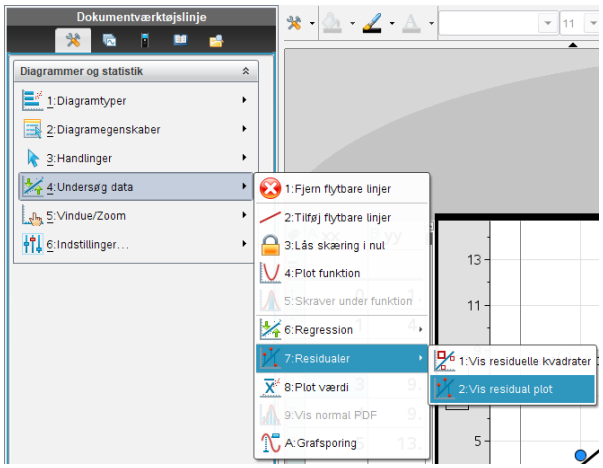
Med residualerne kan man lave et *residualplot* der giver et godt overblik over hvor godt modellens data passer med de observerede data.

RESIDUALPLOT: Et residualplot er en punktplot der består af x-værdier fra de oprindelige data og hvor y-værdierne er residualerne.

d) Lav et residualplot i Nspire til den første graf hvor vi lagde den rette linje ind.

<u>Hjælp til Nspire:</u>

- I det oprindelige "Diagram og statistik" værksted, hvor den rette linje blev indlagt finder man i menuen "undersøg data" muligheden "vis residual plot" som vist på figur.
- Bemærk man kan kun indsætte et residualplot hvis der er indlagt en linje sammen med punktplottet.
- Bemærk at det ikke er nødvendigt at beregne afvigelserne for at lave residualplottet (vi gjorde det for at forstå hvad et residualplot er).



Figur 7: Indsæt residualplot

Nu skulle man gerne have en graf som ses på figur 8 nedenfor.

Nu vender vi tilbage til det oprindelige spørgsmål: *Hvordan afgør man hvilken ret linje der bedst beskriver data?* Dette viser sig at en god måde at "måle" dette på, er ved at beregne summen af residualerne (altså at man lægger alle residualerne sammen).

e) Beregn summen af residualerne i opgaven.

<u>Hjælp til Nspire:</u>

- Da vi har navngivet søjlen hvor vi har beregnet residualerne er det let i Nspire vha. kommandoen sum. Skriv i et matematikfelt i et noteværksted denne kommando:

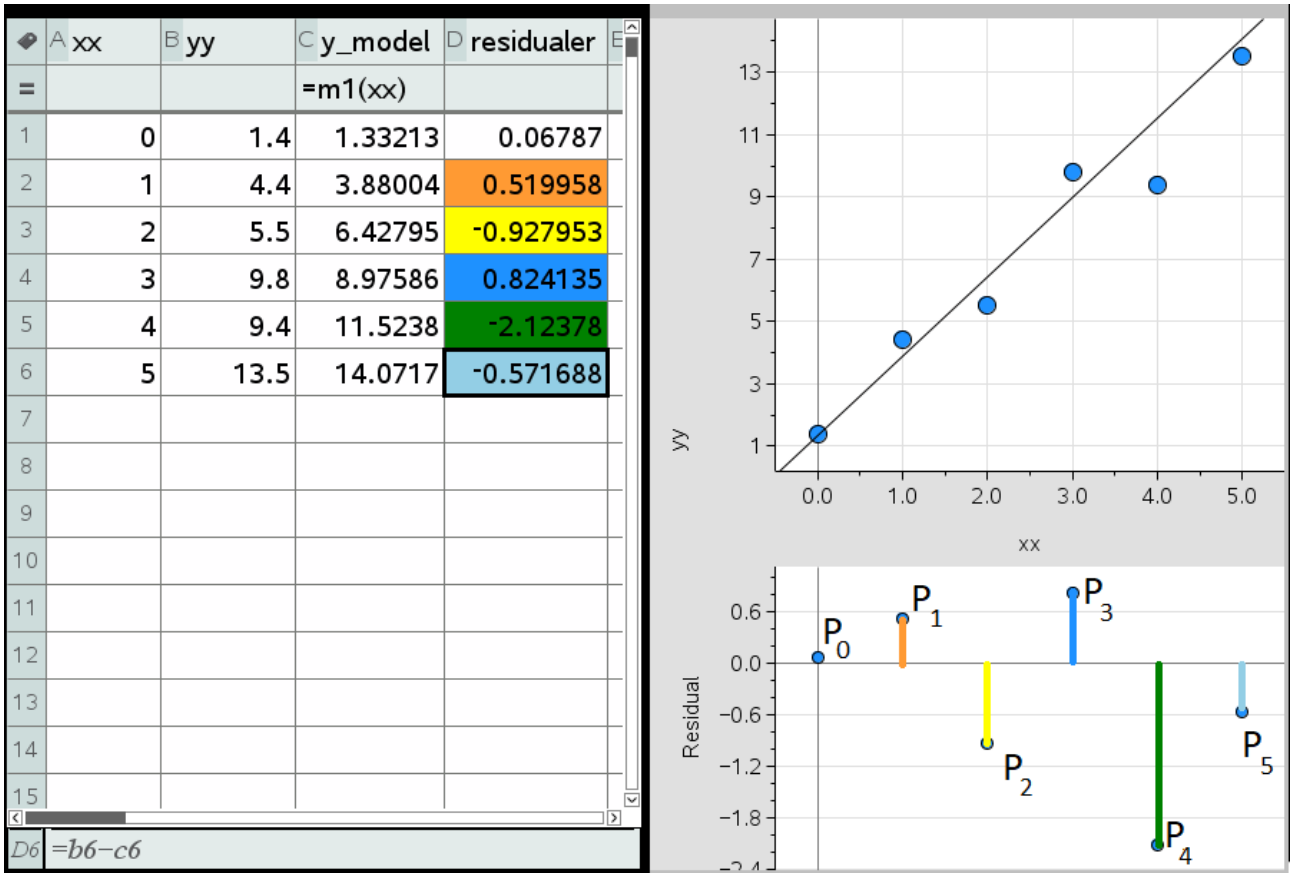
  sum(**residualer**) ▸ -2.21145

Her får alle naturligvis forskellige værdier, da de har lagt forskellige linjer ind. Dem der opnår en sum tættest på nul har altså fundet den linje der beskriver data bedst (dette er ikke helt korrekt som vi skal komme tilbage til senere).

f) Afgør ud fra denne sum om du eller den sidekammerat var bedst til at lægge en ret linje ind.

Opgave 3: Fortolkning af et residualplot

I de foregående opgaver er vi kommet frem til et residualplot som på figur 8 nedenfor (uden farver). Diskuter i gruppen følgende spørgsmål.



	A xx	B yy	C y_model	D residualer
=			=m1(xx)	
1	0	1.4	1.33213	0.06787
2	1	4.4	3.88004	0.519958
3	2	5.5	6.42795	-0.927953
4	3	9.8	8.97586	0.824135
5	4	9.4	11.5238	-2.12378
6	5	13.5	14.0717	-0.571688

D6 =b6−c6

Figur 8: Residualplot

a) **På figuren er residualerne farvet med forskellige farver, de samme farver findes på stregerne på residualplottet. Forklar hvad sammenhængen er mellem farverne på de beregnede afvigelser og farverne på residualplottet.**
b) **Hvordan kan man på residualplottet se om den beregnede afvigelse er positiv eller negativ?**
c) **Residualplot er egentlig et punktplot med samme x-akse som på grafen ovenfor. Hvad er koordinaterne for punkterne $P_0, P_1, \dots P_5$?**

BEMÆRK: Residualplottet vil ofte bruges til at undersøge om der er systematik i datas afvigelse fra modellens forudsigelser. I dette tilfælde er der ingen systematisk afvigelse (senere skal vi se på systematiske afvigelser).

Opgave 4: Anvendelse af residualplot

x	y
0	4.55
1	5.45
2	8.55
3	9.45
4	12.55
5	13.45

Som vi så i opgave 2 kan man altså vurdere hvor godt en linje beskriver data ved hjælp af residualplot, dette skal vi i denne opgave nu anvende til at undersøge et datasæt.

Opret en ny opgave i Nspire (ikke blot en ny side).

a) **Plot data i tabellen til højre og indlæg en linje der beskriver punkterne godt.**

b) **Lav et residualplot og beskriv hvordan det ser ud (husk det er ikke nødvendigt at beregne residualerne først).**

c) **Brug forskriften for linjen til at beregne de forventede y-værdier som vi gjorde i opgaven tidligere (husk at man kan få Nspire til at beregne det automatisk).**

d) **Beregn residualerne og summen af residualerne på samme måde som tidligere.**

e) **Prøv at gøre summen af residualerne så lille som muligt. Er du bedre end din sidemand – hvem kan få den laveste værdi på 20 sekunder?**

f) **Hvad er årsagen til at residualernes sum kan blive meget tæt på nul selvom punkterne ikke ligger tæt på linjen?**

g) **Er residualernes sum et godt mål for hvor tæt på modellen punkterne er? Uddyb dit svar – hvilke begrænsninger er der ved residualplot.**

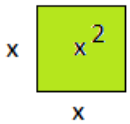
Opgave 5 Residualernes kvadrater

Som vi så i opgave 4 er summen af residualer ikke altid et godt mål for hvor godt en linje beskriver en række punkter. Årsagen var at residualerne regnes med fortegn, og derfor kan positive residualer udlignes af negative residualer. I denne opgave skal vi se hvordan vi let kan komme uden om dette problem. Vi arbejder her videre med datasættet fra opgave 4.

Vi ønsker altså at finde en metode hvor residualernes fortegn ikke har betydning, og det er derfor oplagt at se på residualerne i anden.

a) **Hvorfor er det en god ide at sætte residualerne i anden?**

Når man sætter et tal i anden siger man at man *kvadrere tallet*. Det skyldes at x^2 kan opfattes som arealet af et kvadrat med en sidelængde på *x* som vist på figuren.



b) **Lav en ny søjle i dit regneark, hvor du beregner kvadratet på residualerne.**

Hjælp til Nspire

- Fra tidligere skulle vi gerne have et regneark som ligner figur 9 nogenlunde. Altså har vi beregnet modellens y-værdier og residualerne.
- Lav en ny søjle der hedder res_2, hvor residualerne i anden skal beregnes.
- I feltet under navnet res_2 tastes residualer^2, som vist på figuren.

	A xx	B yy	C y_model	D residualer	E res_2
=			=m1(xx)	=yy−y_mode	esidualer^2
1	0	4.55	4.06322	0.486779	
2	1	5.45	5.95002	-0.500021	
3	2.225	6.495	8.26135	-1.76635	
4	3	9.45	9.72362	-0.273621	
5	4	12.55	11.6104	0.939579	
6	5	13.45	13.4972	-0.047221	

E res_2:=residualer^2

Figur 9: Beregning af kvadraterne på residualerne

c) Beregn summen af kvadraterne.

Dette gøres let i et note værksted med kommandoen

sum(res_2) ▸ 4.56688

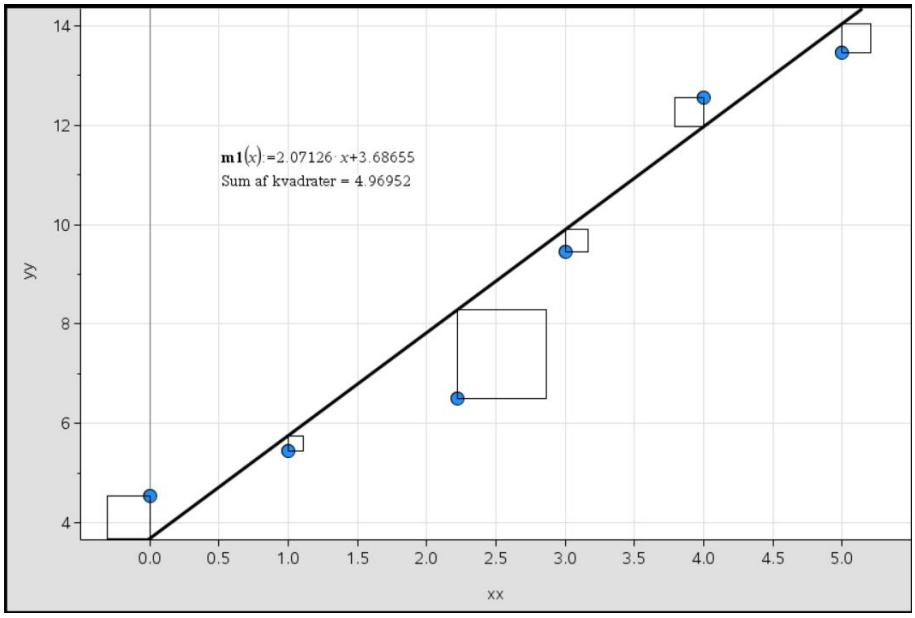
d) Prøv nu at finde den rette linje der bedst beskriver punkterne ved at gøre kvadraternes sum så lille som muligt.

Da residualerne i anden kan opfattes som et areal får man nu en ny mulighed for at illustrere hvordan punkterne ligger i forhold til linjen, ved at vise de residuelle kvadrater.

e) Få Nspire til at vise "residuelle kvadrater".

Hjælp til Nspire: Samme sted som man fandt "vis residualplot" kan man vælge muligheden "vis residuelle kvadrater" (undersøg data -> residualer -> vis residuelle kvadrater). Bemærk at kvadraternes sum også beregnes automatisk. Din graf skulle gerne ligne figur 10 nu (men med en anden model og en anden sum).

f) Juster den flytbare linje og se om du kan få en mindre kvadratsum end din sidemand.



Figur 10: Residuelle kvadrater

Opgave 6 Lineær regression – mindste kvadraters metode.

Vi har nu fundet en ganske smart måde til at afgøre hvor godt en linje beskriver data, nemlig som summen af kvadraterne på residualerne. Hvis vi skal finde den bedst mulige lineære funktion til at beskrive data, skal vi altså finde den linje der giver den mindste kvadrat sum. Man kalder metoden *"mindste kvadraters metode"*. Det kræver mere avanceret matematik at forstå hvordan metoden virker, men heldigvis er det meget let at få sit værktøj (Nspire) til at gøre arbejdet.[1]

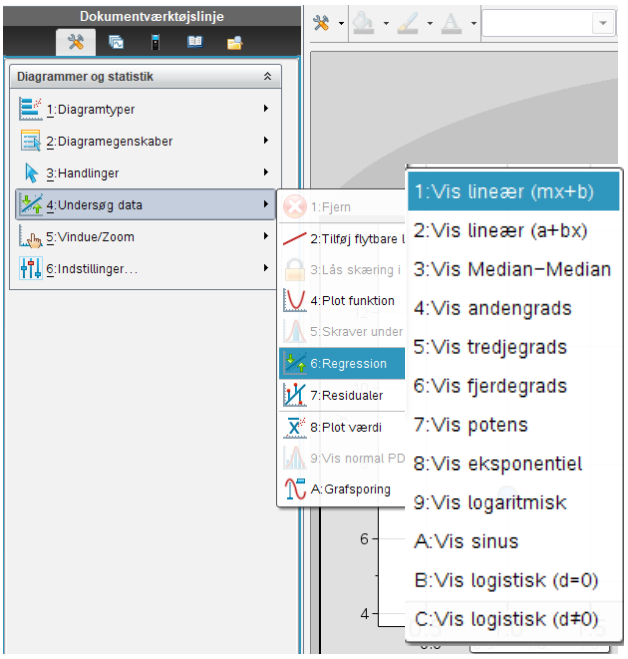
LINEÆR REGRESSION er en metode hvor man finder den linje der passer bedst muligt til et givet datasæt. Den bedste linje findes ved brug af mindste kvadraters metode.

<u>Regression i Nspire</u> finder man i menuen Undersøg data -> regression -> vis lineær som ses på figur 11.

Bemærk: Der er to lineære regressioner, den eneste forskel på dem er hvordan de skriver funktionen op. Den øverste minder mest om den danske skrivemåde

a) **Foretag lineær regression, på data fra den forrige opgave. Var du tæt på at have fundet den bedste rette linje (se på summen af kvadraterne)?**

Hvis man har foretaget lineær regression i et diagram og statistik værksted er det let at få vist resultatet i et note værksted sammen med den tekst der altid skal følge med en matematikopgave. I et matematikfeltet skriver man kommandoen "stat." ("stat punktum"), og får adgang til resultaterne fra regressionen – se figur 12 nedenfor. Prøv f.eks. at skrive "stat.results" i et notefelt når du har lavet regression…



Figur 11: regression i Nspire. Bemærk man skal være i et diagram og statistik værksted for at menuen er tilgængelig.

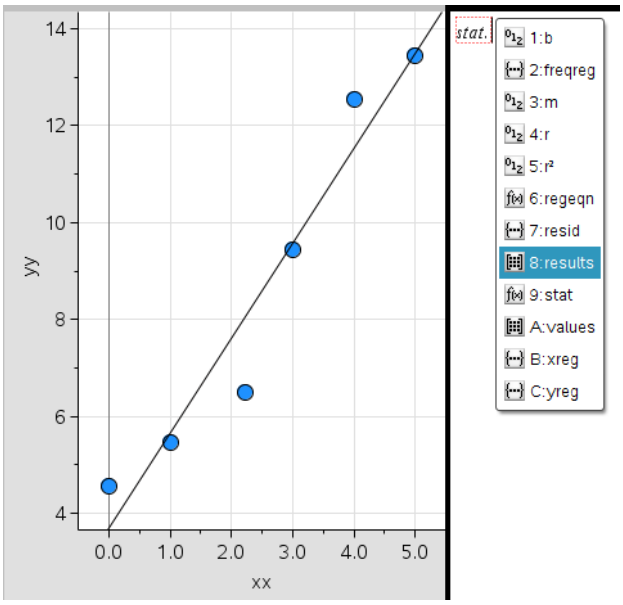
b) **I samme opgave som før prøv at åbne et nyt noteværksted og skriv nogle af de kommandoer vist på figur 12 og 13 på næste side.**

c) **Prøv at undersøge flere af kommandoerne der dukker op når man skriver "stat." i et matematikfelt.**

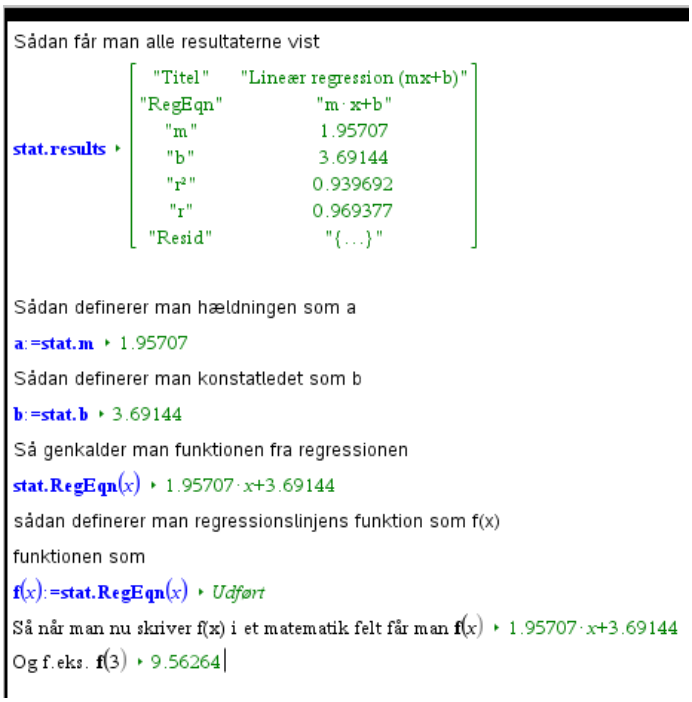
R^2 værdien:

I oversigten på figur 13 kan man bl.a. se den så kaldte R^2-værdi der er 0,9396. Denne værdi angiver hvor godt regressionen beskriver data. Jo tættere på 1 den er des bedre. R^2-værdien kan altså bruges i stedet for kvadraternes sum til at vurdere modellen. I samfundsfag bruges denne størrelse ofte.

[1] Mindste kvadraters metode er en statistisk metode som man evt. kan vende tilbage til i forbindelse med statistik.



Figur 13: Genkald regression resultatet



Figur 12: Nyttige tricks ved regression

Opgave 7 Lineær regression og residualplot – systematisk afvigelse

Lineær regression er et let redskab til at finde den bedste lineære model til at beskrive et datasæt, og sammen med et residualplot er det et stærkt redskab til at analysere data.

0	3.1
1	4.9
2	7.1
3	8.9
4	11.1
5	12.9
6	15.1
7	16.9
8	19.1
9	20.9
10	23.1

a) **Udfør lineær regression på data til højre (copy-paste den ind i Nspire).**
b) **Bestem R^2-værdien.**

Ud fra R^2-værdien at dømme er modellen rigtig god til at beskrive data, men prøv at lave et residualplot…

c) **Hvad viser residualplottet om modellen? Afviger data systematisk fra tabellen – hvordan?**

Opgave 8 Lineær regression og residualplot

Vi har modtaget en række data, som vi forventer kan beskrives med en lineær model på formen $f(x) = a \cdot x + b$

a) **Copy- paste data til højre ind i Nspire og bestem forskriften for den rette linje der bedst beskriver data til højre.**
b) **Hvad er konstanterne a og b i modellen?**
c) **Lav et residualplot og svar på spørgsmålet: Afviger data systematisk fra den fundne model?**
d) **Diskuter i gruppen om den fundne model er god til at beskrive data.**

0	-1.
1	1.31
2	3.59
3	5.81
4	7.95
5	10.
6	11.95
7	13.81
8	15.59
9	17.31
10	19.
11	20.69
12	22.41
13	24.19
14	26.05
15	28.
16	30.05
17	32.19
18	34.41
19	36.69
20	39.
21	41.31
22	43.59
23	45.81
24	47.95
25	50.
26	51.95
27	53.81
28	55.59
29	57.31
30	59.
31	60.69
32	62.41
33	64.19
34	66.05
35	68.
36	70.05
37	72.19
38	74.41

Opgave 9 Den hemmelige besked…

Som vi har set kan residualer vise dybereliggende tendenser i data end den mere overordnede tendens (som en lineær sammenhæng). Nedenfor er en enorm mængde data. Se om du kan finde den hemmelige besked…

1	5.2
1.5	6.2
2	7.2
2.5	8.2
3	9.2
3.5	10.2
4	11.2
4.5	12.2
5	13.2
0.9	4.6
1.4	5.6
1.9	6.6
2.4	7.6
2.9	8.6
3.4	9.6
3.9	10.6
4.4	11.6
4.9	12.6
2.91	8.97
2.93	8.96
2.95	8.95
2.97	8.94
2.99	8.98
3.01	8.97
3.03	8.96
3.05	8.95
8	19.15
8.4	19.97
8.8	20.79
9.2	21.6
9.6	22.4
10.	23.2
10.4	24.
10.8	24.8

11	25.2
11.2	25.59
11.6	26.38
12.	27.17
12.4	27.98
12.8	28.79
13.2	29.6
13.6	30.4
14.	31.2
14.4	32.
14.8	32.8
15.2	33.6
15.6	34.39
16.	35.17
8.1	19.32
8.5	20.09
8.9	20.85
9.3	21.62
9.7	22.39
10.1	23.16
10.5	23.93
10.9	24.7
11.3	25.46
11.7	26.23
12.1	27.
12.5	27.83
12.9	28.66
13.3	29.5
13.7	30.33
14.1	31.16
14.5	31.99
14.9	32.82
15.3	33.65
15.7	34.49
16.1	35.32
16.2	35.55
19	40.8

19.1	41.05
19.2	41.3
19.3	41.55
19.4	41.8
19.5	42.05
19.6	42.3
19.7	42.55
19.8	42.8
19.9	43.
20.5	44.2
21.1	45.37
21.7	46.54
22.3	47.71
22.9	48.88
23.2	49.51
23.5	50.14
23.8	50.77
24.1	51.4
24.6	52.4
24.9	52.95
25.	53.1
25.1	53.25
25.2	53.4
25.3	53.55
25.4	53.7
25.5	53.85
25.6	54.
27	56.8
27.3	57.45
27.6	58.1
27.9	58.75
28.2	59.4
28.5	60.05
28.8	60.7
29.1	61.35
29.4	62.
29.7	62.55

30.	63.1
30.3	63.65
30.6	64.2
30.9	64.75
31.2	65.3
31.5	65.85
31.8	66.4
28.25	59.45
28.75	60.45
29.25	61.45
29.75	62.45
30.25	63.45
32	67.2
32.5	68.2
33.	69.2
33.5	70.2
34.	71.2
34.5	72.2
35.	73.2
35.5	74.2
36.	75.2
36.5	76.2
37.	77.2
34.5	71.8
34.51	71.87
34.52	71.94
34.53	72.01
34.54	72.08
34.55	72.15
34.56	72.22
34.57	72.29
34.58	72.36
39	80.8