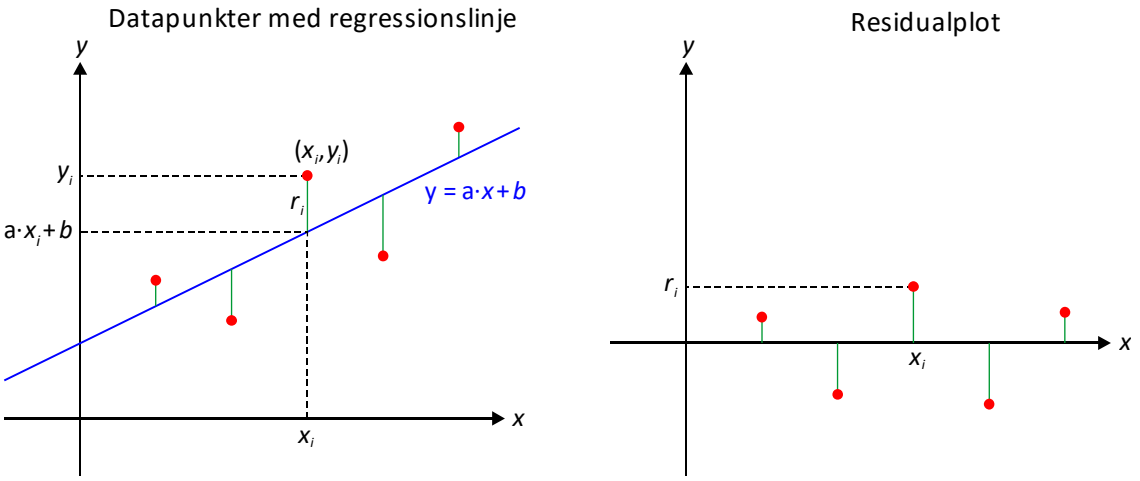
Residualer i grundforløbet

I dette lille tillæg til grundforløbet, skal vi kigge på begreberne *residualer*, *residualplot* samt *residualspredning*. Vi vil se, hvad det handler om og hvordan vi hurtigt kan få tingene udregnet og plottet i Maple. En lidt mere grundig omtale venter i Mat B/Mat A.

Residualer og residualplot

Vi har tidligere studeret begrebet *lineær regression*, hvor man finder den i en bestemt forstand bedste rette linje, som tilnærmer n datapunkter $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. I den forbindelse blev *forklaringsgraden* R^2 introduceret. Løst sagt fortæller tallet, hvor tæt datapunkterne ligger på en ret linje. Er forklaringsgraden tæt på 1, er det et udtryk for, at datapunkter ligger tæt på regressionslinjen. Men som det vil fremgå af dette tillæg, skal man være varsom med at lægge for meget i dette tal. Det ville vel også være underligt, hvis et enkelt tal skulle kunne indeholde al information om regressionen! Forklaringsgraden kan for eksempel ikke afsløre *systematiske afvigelser* fra en lineær sammenhæng. Det er her begrebet *residualer* kommer ind.



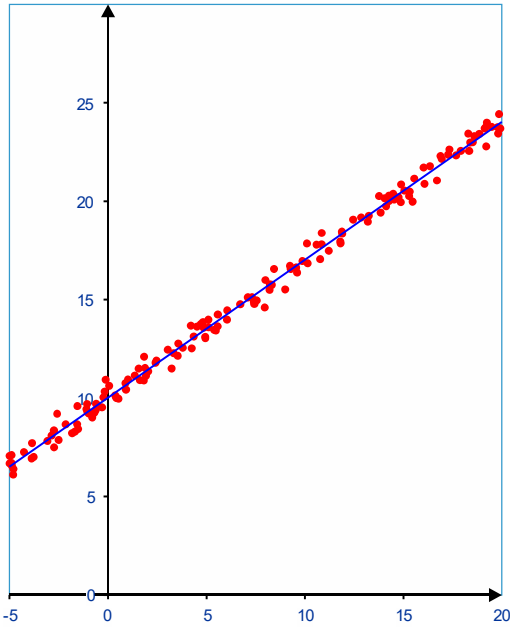
Figuren foroven til venstre viser datapunkter og den tilhørende regressionslinje. *Residualet* for det *i*'te datapunkt defineres som forskellen mellem datapunktets y-værdi og den værdi, regressionslinjen *forudsiger*. Vi kalder den differens for r_i. Vi har altså:

$$r_i = y_i - (a \cdot x_i + b) \quad \Leftrightarrow \quad y_i = a \cdot x_i + b + r_i \tag{1}$$

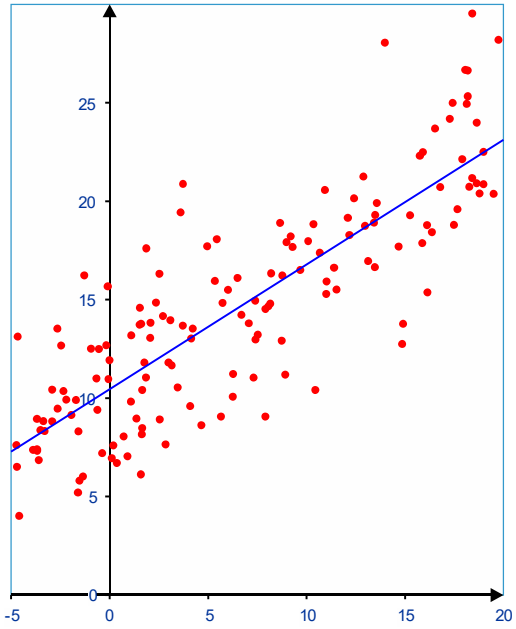
Det *i*'te residual er med andre ord den lodrette afstand fra det *i*'te punkt til linjen, regnet med fortegn. Ligger punktet over linjen, er residualet positivt. Ligger punktet derimod under linjen, er residualet negativt. På figuren er de lodrette afstande markeret med en tynd grøn linje. På figuren til højre er residualet afbildet for hvert datapunkt. Det kaldes et *residualplot* (bortset fra, at de lodrette grønne linjer normalt udelades). Det er nemmere at se punkternes variation i forhold til regressionslinjen, når man anvender residualplottet fremfor bare at kigge på det oprindelige plot.

Residualspredning

Et andet begreb er *residualspredningen*. Som navnet antyder, har dette tal noget at gøre med, hvor meget punkterne spreder sig omkring regressionslinjen. Nedenfor er der vist et eksempel på data med henholdsvis en lille residualspredning og et eksempel på data med en større residualspredning. I begge tilfælde vil man muligvis godtage, at der er tale om en lineær sammenhæng. Sammenhængen er blot mere diffus i tilfældet på figuren til højre. I de naturvidenskabelige uddannelser vil han ofte se data med mindre spredning (fx som resultat af en lille usikkerhed på målingerne), mens man i de samfundsfaglige uddannelser ofte vil se meget større spredning i data, fordi mange andre forhold i samfundet kan spille ind, når man sætter to størrelser op mod hinanden.



Lille residualspredning: s = 0.38907



Stor residualspredning: s = 3.0547

Residualspredningen s er mere præcist et skøn eller estimat over spredningen i den *simple lineære regressionsmodel*. Den udregnes ved at udregne alle residualerne, opløfte hver af dem til 2. potens og lægge sammen. Derefter dividerer man med $n-2$ og tager kvadratroden, hvor n er antallet af datapunkter:

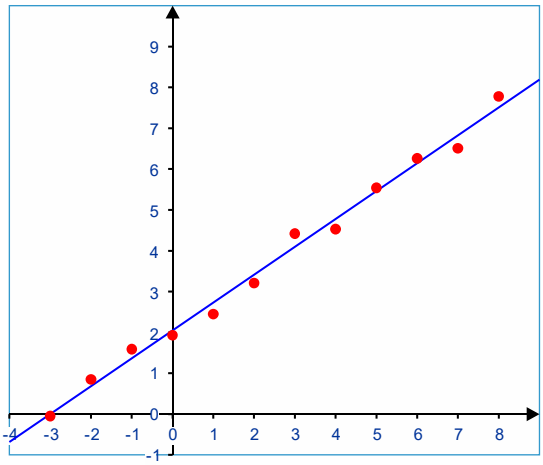
(2) $$s = \sqrt{\frac{r_1^2 + r_2^2 + \ldots + r_n^2}{n-2}}$$

(På engelsk kaldes residualspredningen også for *Residual Standard Deviation*). Hvis alle datapunkter ligger eksakt på en linje, så er alle residualer 0 og residualspredningen er dermed også 0. I alle andre tilfælde giver residualspredningen et positivt tal.

To eksempler i Maple

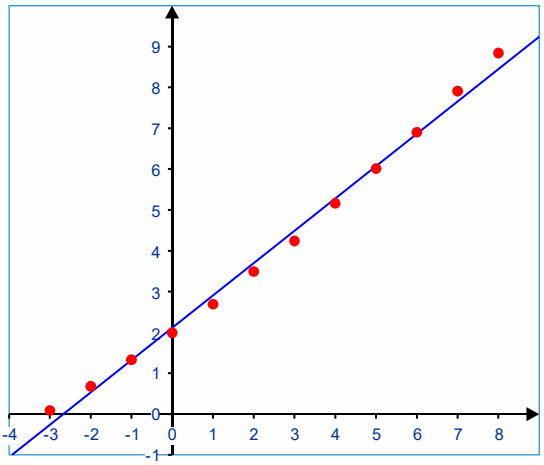
På figurerne nedenfor er der udført lineær regression på to forskellige datasæt. Umiddelbart skulle man tro, at sammenhængen ser mest lineær ud på figuren til højre. Dens forklaringsgrad er da også lidt tættere på 1. Ved at lave residualplot skal vi imidlertid så tvivl, om der overhovedet er tale om en lineær sammenhæng i situationen til højre.

Regression uden systematiske afvigelser



$R^2 = 0.9914$, $s = 0.2398$

Regression med systematiske afvigelser



$R^2 = 0.9941$, $s = 0.2322$

Udført i Maple:

Eksempel 1

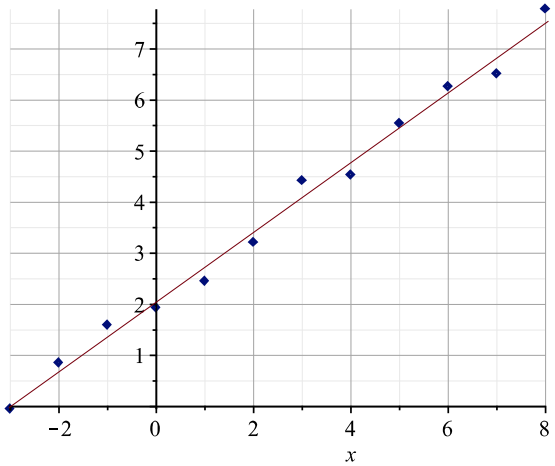
I det følgende udfører vi lineær regression på en række datapunkter. Punkternes x-koordinater anbringes i en liste, som vi kalder X, mens y-koordinaterne anbringes i en liste, som vi kalder Y.

```
restart
with(Gym) :
X := [-3,-2,-1, 0, 1, 2, 3, 4, 5, 6, 7, 8] :
Y := [-0.05, 0.85, 1.59, 1.93, 2.45, 3.21, 4.42, 4.53, 5.54, 6.26, 6.51, 7.78] :
```

Lineær regression

```
LinReg(X, Y)
```

Lineær regression
y =0.68273 x + 2.0448.
Forklaringsgrad R^2 = 0.991448657803292



Vi aflæser forskriften for den lineære sammenhæng: $f(x) = 0.68273\,x + 2.0448$.
Forklaringsgraden er 0.9914.

Residualer

Da der er over 10 datapunkter, fortæller vi først Maple, at den skal vise op til 20 datapunkter. Det gøres med kommandoen lige herunder.

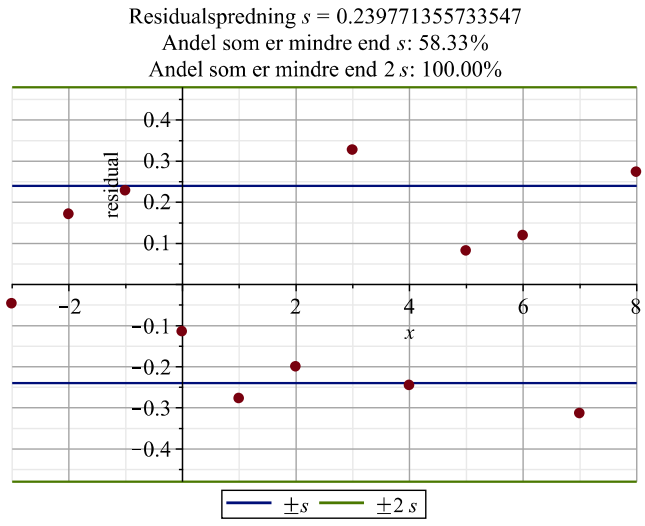
visMatrix(20) :
residualer(*X*, *Y*, *LinReg*)

$$\begin{bmatrix} -3 & -0.0466666666666657 \\ -2 & 0.170606060606061 \\ -1 & 0.227878787878788 \\ 0 & -0.114848484848485 \\ 1 & -0.277575757575757 \\ 2 & -0.200303030303030 \\ 3 & 0.326969696969696 \\ 4 & -0.245757575757575 \\ 5 & 0.0815151515151511 \\ 6 & 0.118787878787877 \\ 7 & -0.313939393939394 \\ 8 & 0.273333333333333 \end{bmatrix} \quad \textbf{(1)}$$

I første søjle står datapunkternes x-værdier, mens de tilhørende residualer står i 2. søjle.

Residualplot

plotResidualer(*X*, *Y*, *LinReg*)



Residualerne ser ud til at være meget tilfældigt fordelte, hvilket kan tale for, at der virkelig er tale om en lineær sammenhæng.

Residualspredning

residualspredning(*X*, *Y*, *LinReg*)

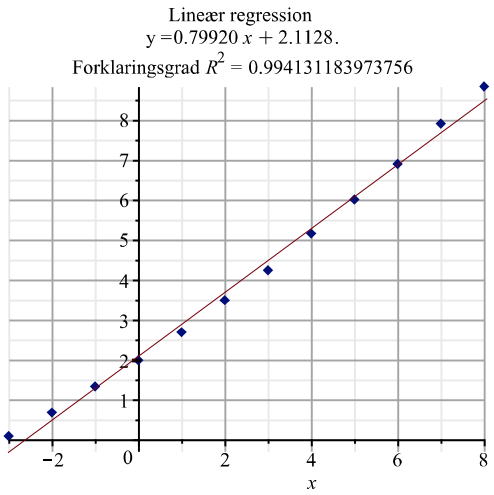
$$0.239771355726490 \quad \textbf{(2)}$$

Eksempel 2

```
restart
with(Gym) :
X := [-3, -2, -1, 0, 1, 2, 3, 4, 5, 6, 7, 8] :
Y := [0.09, 0.68, 1.33, 1.99, 2.69, 3.49, 4.24, 5.16, 6.01, 6.90, 7.91, 8.84] :
```

Lineær regression

```
LinReg(X, Y)
```



Vi aflæser forskriften for den lineære sammenhæng: $f(x) = 0.79920\,x + 2.1128$.
Forklaringsgraden er 0.9941.

Residualer

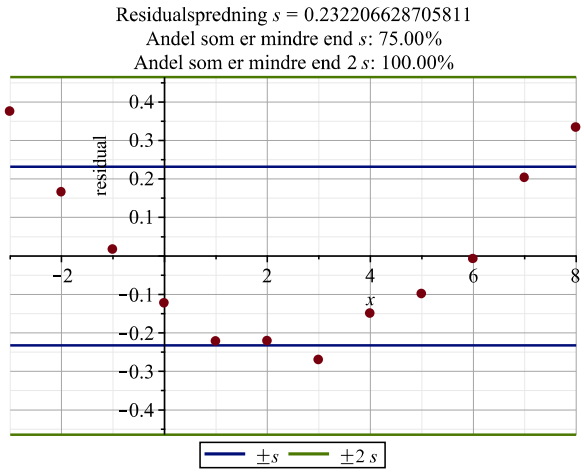
```
visMatrix(20) :
residualer(X, Y, LinReg)
```

$$\begin{bmatrix} -3 & 0.374743589743591 \\ -2 & 0.165547785547787 \\ -1 & 0.0163519813519821 \\ 0 & -0.122843822843822 \\ 1 & -0.222039627039627 \\ 2 & -0.221235431235431 \\ 3 & -0.270431235431236 \\ 4 & -0.149627039627040 \\ 5 & -0.0988228438228447 \\ 6 & -0.00801864801864838 \\ 7 & 0.202785547785547 \\ 8 & 0.333589743589743 \end{bmatrix} \quad \textbf{(1)}$$

I første søjle står datapunkternes x-værdier, mens de tilhørende residualer står i 2. søjle.

Residualplot

plotResidualer(*X*, *Y*, *LinReg*)



Residualerne er positive i enderne af intervallet, mens de er negative i midten. Det kan tyde på en *systematisk afvigelse* fra en lineær sammenhæng, og man bør være skeptisk, om der overhovedet er tale om en lineær sammenhæng mellem x- og y-værdierne! Kigger man på regressionsplottet længere oppe på siden, ser man nu også, at punkterne danner en lidt krum kurve ...

Residualspredning

residualspredning(*X*, *Y*, *LinReg*)

0.232206628698977 **(2)**

I de to eksempler udført i Maple har vi set, at forklaringsgraden er størst i eksempel 2, ligesom residualspredningen er mindst i eksempel 2. Alligevel må en inspektion af residualplottet så tvivl om, der er tale om en lineær sammenhæng i eksempel 2! Det er med andre ord fornuftigt at kigge på *fordelingen* af residualerne, før man drager forhastede konklusioner. Det skal dog pointeres, at man måske nok kunne ønske sig flere datapunkter, hvis man ville være mere sikker i sin sag.

Til sidst skal det lige tilføjes, at man også kan udføre residualplot for andre typer fit, herunder *eksponentiel regression* og *potensregression*.

Opgaver

Nedenstående opgaver skal regnes med dit CAS-værktøj.

Opgave 1

Tid (s)	20	40	60	80	100	120	140	160	180
Temperatur (°C)	28,7	36,5	42,6	50,4	57,6	63,7	71,8	77,7	85,0

Der er udført forsøg med en elkoger, hvor en mængde vand er opvarmet. For hvert 20 sekund er vandets temperatur blevet målt med et termometer. Måledata fremgår af tabellen ovenfor.

a) Foretag lineær regression på data. Angiv a og b i den lineære sammenhæng $y = a \cdot x + b$.

b) Beregn residualerne og lav desuden et residualplot.

c) Kan man med rimelighed sige, at der er tale om en lineær sammenhæng mellem vandets temperatur og den forløbne tid?

d) Angiv forskriften for den lineære funktion, som den lineære regression leverede og giv en sproglig fortolkning af hældningskoefficienten a og konstantleddet b i denne. Husk heri at konkludere med de aktuelle talværdier og enheder!

e) Benyt forskriften fra d) til at forudsige, hvad temperaturen var efter 90 sekunder.

f) Hvornår var temperaturen nået 75°C ifølge modellen?

g) Hvor meget stiger temperatur med for hvert 5. sekund, der går, ifølge modellen?

Opgave 2

Asger er gået i skarp træning hos Team Danmark med henblik på at forberede sig til EM i Mountainbike. Hans træning har været kondition kombineret med styrketræning. Han har fået målt sin fedtprocent hver 10. dag for at kunne følge udviklingen. I nogle tilfælde glemte han dog at notere målingerne ned. Resultaterne af målingerne var følgende:

Dage forløbet	0	10	20	30	50	70	80	90	110	120
Fedtprocent	22,5	20,6	18,9	17,6	15,0	12,7	11,5	10,6	8,6	8,1

a) Foretag lineær regression på data. Hvad er a og b i den lineære model $y = a \cdot x + b$?

b) Bestem residualerne og lav desuden et residualplot.

c) Kan man med rimelighed sige, at hans fedtprocent er aftaget lineært i perioden, eller er der tale om systematiske afvigelser, som kan rokke ved den forestilling?

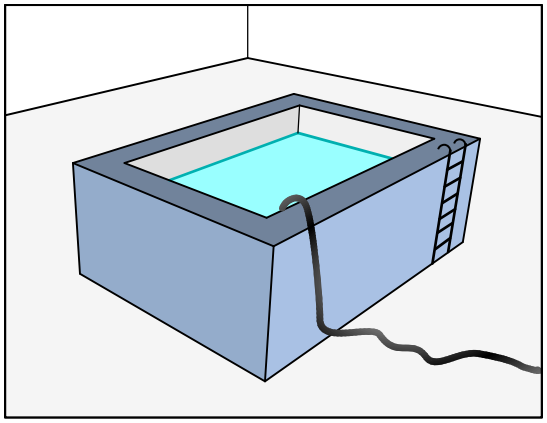
Opgave 3

x	-7	-2	3	5	7	11	15	18	20	23	26	30
y	9,4	17,0	26,0	30,4	31,9	40,5	45,5	53,9	55,1	62,6	66,1	74,8

a) Foretag lineær regression på data ovenfor.
b) Bestem residualerne.
c) Lav et residualplot.
d) Kan man med rimelighed godtage, at der er tale om en lineær sammenhæng, eller er der tegn på systematiske afvigelser?
e) Bestem residualspredningen s.
f) Hvor stor en procentdel af datapunkterne har en lodret afstand til regressionslinjen, som er højst 1 residualspredning, altså opfylder $-s \le r \le s$ for residualet r? Samme spørgsmål for 2 residualspredninger, altså opfylder $-2s \le r \le 2s$.

Opgave 4

En stor rektangulær beholder med vand skal tømmes med en pumpe. Beholderens indvendige dimensioner er: Længde 8 m, bredde 5 m og højde 3 meter. Man er interesseret i at undersøge, om pumpen har en konstant flowrate. Derfor noteres for hver 10 minutter, der går, vandstanden (højden) i beholderen. Den kan aflæses meget præcist på et målebånd, som sidder fast på indersiden af beholderen. Måleresultaterne er følgende:



Tid (min).	10	20	30	40	50	60	70	80	90	100
Højde (cm)	285.6	262.4	252.0	234.6	225.7	207.6	188.4	183.4	166.7	150.4

a) Foretag lineær regression på data. Hvad er a og b i den lineære model $y = a \cdot x + b$?
b) Lav et residualplot. Redegør for, hvorfor man med rimelighed kan antage, at pumpen har en konstant flowrate.
c) Benyt forskriften i den lineære regression til at afgøre, hvor lang tid det tager, før beholderen er tom.
d) Bestem residualspredningen.
e) (Lidt svær). Hvor mange liter vand kan pumpen flytte i minuttet?

Opgave 5

I forbindelse med en lineær regression på ti datapunkter fik man følgende regressionslinje: $y = 0{,}742 \cdot x + 3{,}921$. To af datapunkterne var $(2,6)$ og $(7,8)$. Beregn residualet for hvert af de to punkter. Tegn evt. linjen og de to punkter først i GeoGebra for at få overblik.