

Modul 3: Lineær regression

Om at opstille modeller fra data

Et væsentligt spørgsmål om matematiske modeller er hvor de kommer fra. En af de vigtige kilder til at opstille en matematisk model er *data*.

Hvis vi i et NV-forsøg har taget et måleglas og fyldt 50 ml sprit i det og efterfølgende vejer det på en vægt til 78 gram, har vi produceret et datapunkt. De to målinger siges at være *samhørende*. Punktet kunne hedde (50, 78).

Herefter gentager vi forsøget med 100 ml, 150 ml, 200 ml, 250ml og 300 ml. Vi kan da sammenfatte måleresultaterne i en tabel som nedenstående:

Volumen (ml)	50	100	150	200	250	300
Masse (g)	78	132	169	195	231	285

Tabellen repræsenterer altså seks punkter: (50, 78), (100,132), (150,169), (200,195), (250,231) og (300,285). Disse seks punkter kan indtegnes i et koordinatsystem. Det kaldes et *punktplot* over data.

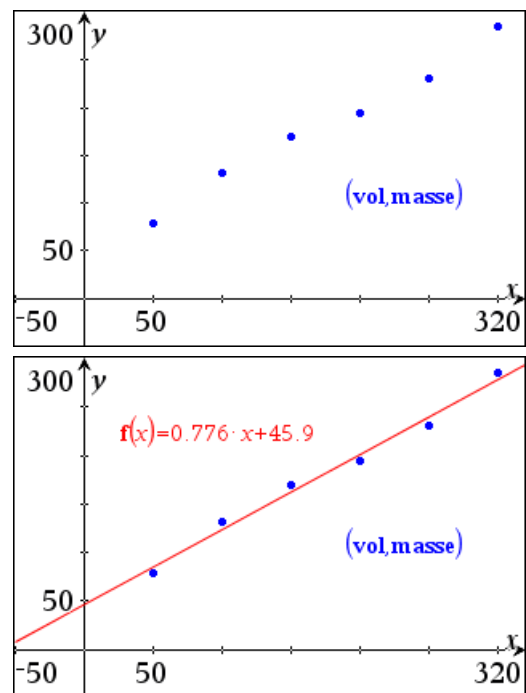
Som det ses på figuren til højre ligger punkterne omtrent på en ret linje, men det kan også ses at de ikke ligger helt præcist på en linje.

Ved hjælp af et værktøj der kaldes *lineær regression* kan vi finde den *bedste rette linje* i forhold til punkterne. Som det ses af figuren til højre er den lineære funktion som bedst passer med punkterne givet ved

$$f(x) = 0,776 \cdot x + 45,9$$

Vi har hermed ud fra data opstillet en lineær model som beskriver sammenhængen mellem massen $f(x)$ (målt i gram) af målebægeret når det indeholder et volumen sprit på x (målt i ml).

Tallet $a = 0,776$ viser at massen vokser ca. 0,78 gram hver gang volumenet vokser med 1ml, mens tallet $b = 45,9$ viser at massen er 45,9 gram når der er 0 ml sprit i måleglasset.



Lineær regression med Nspire

I en tabel har vi opgjort den samlede danske eksport til Island for hvert år i perioden 2010-2018:

Årstal	2010	2011	2012	2013	2014	2015	2016	2017	2018
Eksport (mia. kr.)	1,88	2,06	2,37	2,40	2,68	2,95	3,28	3,54	3,70

Vi vil gerne opstille en lineær model $f(x) = a \cdot x + b$ som beskriver den samlede danske eksport til Island (målt i milliarder kroner) til tidspunktet x (målt i antal år efter 2010).

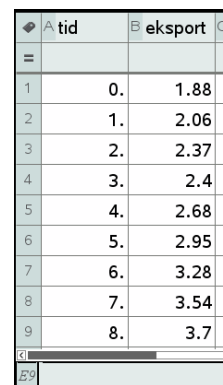
I Nspire oprettes et "Lister og regneark"-vindue. Data fra tabellen tages ind i regnearket så tiden står i én søjle og eksporten i en anden. Da tiden skal måles i "antal år efter 2010" laves årstallene om til værdierne 0, 1, 2, 3, 4, 5, 6, 7 og 8, da f.eks. 2013 er 3 år efter 2010 og 2010 selv er 0 år efter 2010.

Det bør se ud som vist til højre. Husk overskrifter i øverste linje.

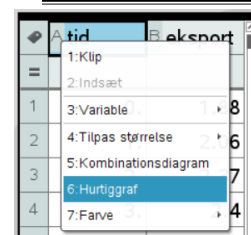
Da vi skal have *tid* på førsteaksen højreklikker vi på overskriften i denne kolonne og vælger "hurtiggraf" i den fremkomne menu (se figuren til højre).

Der viser sig nu et "diagrammer og statistik"-vindue til højre, hvor *tid* er på førsteaksen. Klik på "Tilføj variabel" på andenaksen og tilføj "eksport".

Nu ses et *punkplot* over data.



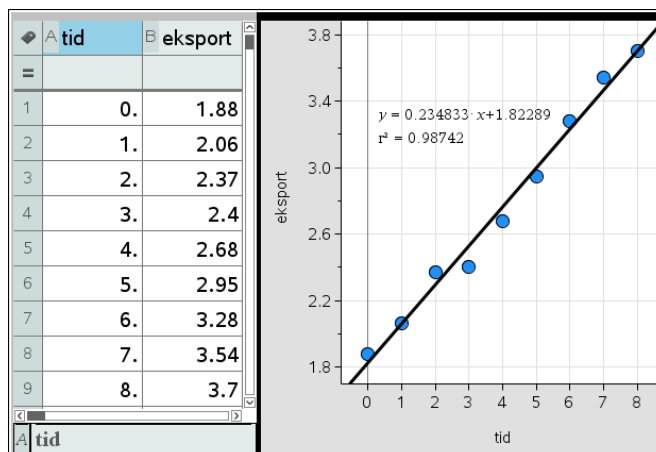
	A tid	B eksport
1	0.	1.88
2	1.	2.06
3	2.	2.37
4	3.	2.4
5	4.	2.68
6	5.	2.95
7	6.	3.28
8	7.	3.54
9	8.	3.7



I menuen til venstre vælges "Undersøg data" og herunder "Regression", hvor der vælges "1. Vis lineær ($mx + b$)".

Det skal herefter se ud som på figuren til højre. Fra diagrammet til højre kan direkte aflæses den fundne model:

$$f(x) = 0,235 \cdot x + 1,823$$



Den fundne model er altså et forsøg på at beskrive udviklingen i Danmarks samlede eksport til Island. Det vil være muligt at bestemme den forventede eksport i 2013 som $f(3)$ og give en forudsigtelse af eksporten i 2020 ved at bestemme $f(10)$.

Tallet $a = 0,235$ kan fortolkes som den årlige stigning i eksporten til Island. Det vil sige at ifølge modellen vokser Danmarks eksport til Island med ca. 0,235 milliarder kroner om året.

Tallet $b = 1,823$ kan fortolkes som Danmarks eksport til Island for $x = 0$, dvs. i 2010.

Vurdering af en model

Bare fordi en linje er den ”bedste rette linje” er det ikke sikkert at det er en god ret linje og dermed en god model til at beskrive datapunkterne. På figuren med regressionen på forrige side, kunne man aflæse tallet $r^2 = 0,987$ (udtales ” r i anden”).

Hvis tallet ikke kan ses, skal man vælge ”Indstillinger”-menuen til venstre, sætte flueben i ”Diagnosticering” og klikke ”Gør til standard”.



Tallet viser hvor godt linjen passer til punkterne. Tallet er altid mellem 0 og 1, dvs.: $0 \leq r^2 \leq 1$. Jo tættere r^2 er på 1, jo bedre passer linjen og punkterne til hinanden. I eksemplet her er tallet meget tæt på 1, så vi siger at r^2 -værdien taler for at det er en god model.

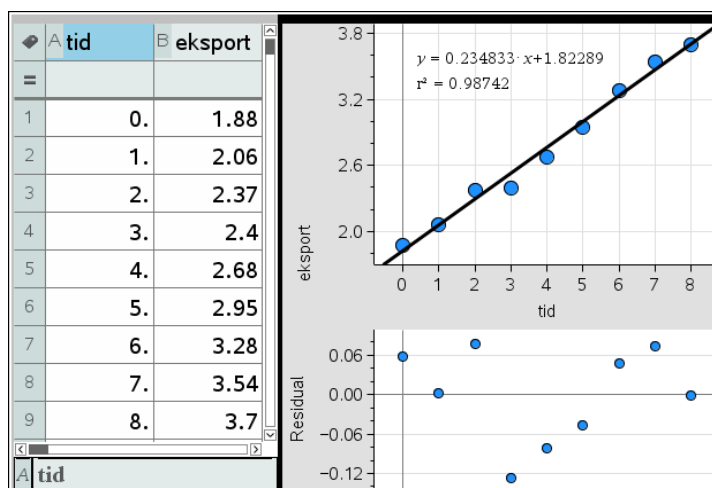
Residualplot

Et andet værktøj vi anvender er et såkaldt *residualplot*. Et residual er den absolutte afvigelse mellem en *observeret* og *forventet* værdi. I tabellen ses en stribe *observerede værdier* og for hver af disse kan nu beregnes den forventede værdi som $f(0)$, $f(1)$, $f(2)$, osv. Og for hvert sæt af observeret og forventet værdi kan beregnes et residual: $r = \text{observeret} - \text{forventet}$

Tid	0	1	2	3	4	5	6	7	8
Observeret	1,88	2,06	2,37	2,40	2,68	2,95	3,28	3,54	3,70
Forventet	1,82	2,06	2,29	2,53	2,76	3,00	3,23	3,47	3,70
Residual	0,06	0,00	0,08	-0,13	-0,08	-0,05	0,05	0,07	0,00

I et residualplot afbilledes x -værdierne på førsteaksen og residualerne på andenaksen. Dermed ses residualerne placeret rundt om x -aksen (hvor $r = 0$).

I Nspire kan man tegne residualplottet ved i menuen ”Undersøg data” at vælge ”Residualer” og derpå vælge ”Vis residualplot”. Det skal se ud som vist til højre.



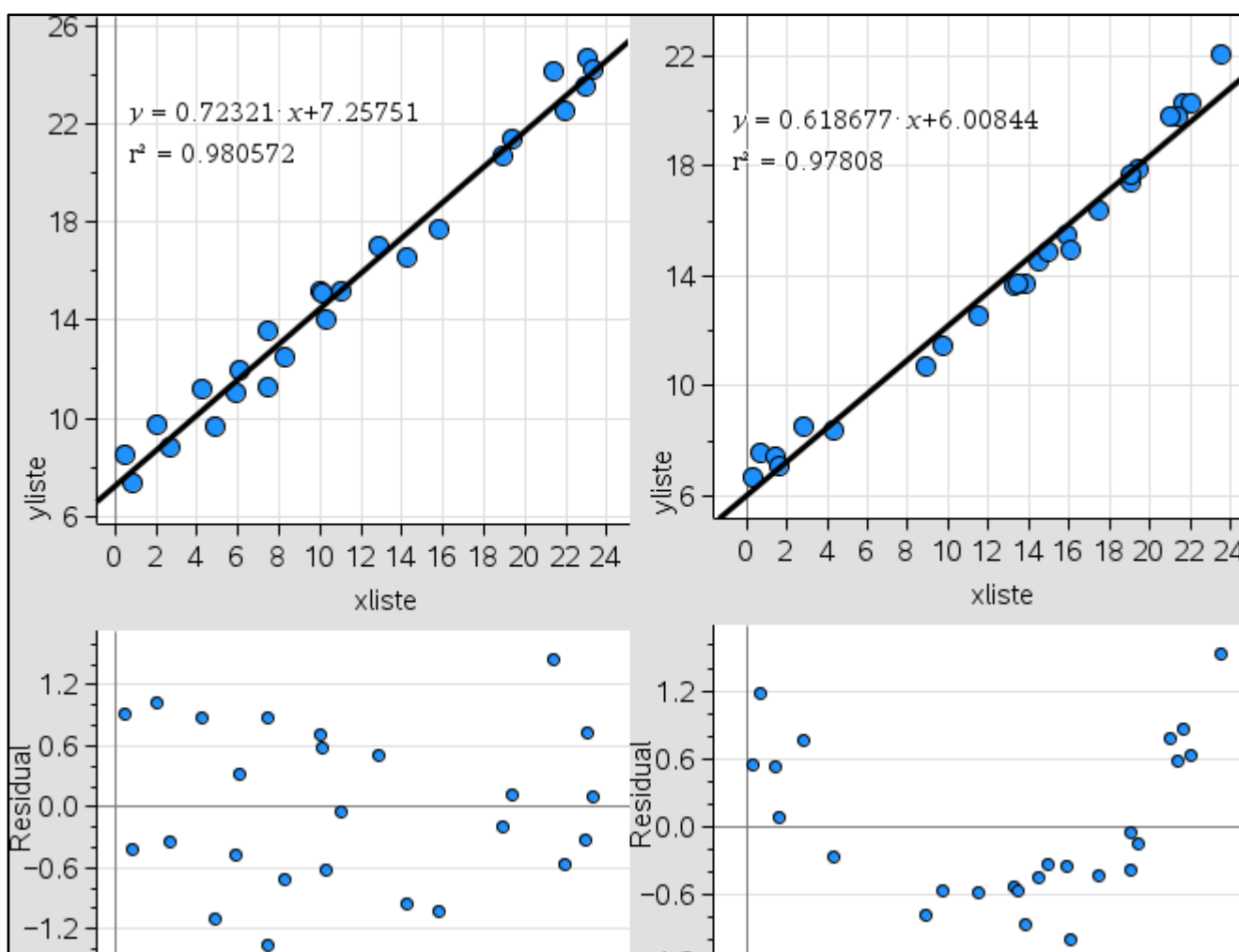
På residualplottet kan vi bl.a. se at den største absolutte afvigelse fås ved $x = 3$, da dette residual ligger længst væk fra x -aksen. Det passer også med tabellen ovenfor.

Vurdering af residualplot

Herunder ses to eksempler på lineær regression med tilhørende residualplot. Begge har r^2 -værdi tæt på 1 og ser fornuftige ud. Med Nspire er residualplottet tegnet. Vi interesserer os nu for to ting:

- A. Er residualerne relativt store? På y -aksen af dataplottet ses at de observerede y -værdier ligger mellem 6 og 25, mens de fleste residualer ligger mellem -1 og 1. Residualerne er dermed forholdsvis små sammenlignet med y -værdierne. Det taler for at modellen er god.
- B. Er residualerne *systematiske*? På residualplottet til venstre er punkterne placeret i en "sky" omkring x -aksen. De er altså *usystematiske*. Det taler for at modellen er god.

På residualplottet til højre ses at punkterne er over x -aksen i kanterne og under i midten. De danner nærmest en bue under x -aksen. Det opfatter vi som *systematiske afvigelser*. Det taler kraftigt for, at en lineær model ikke er velegnet til at beskrive punkterne med.



Observation og forventning

En matematisk model er altid en idealiseret beskrivelse af virkeligheden. Den *er* ikke virkeligheden. Vi kan altså ikke være sikre på at det modellen forventer svarer til det vi vil observere, hvis vi faktisk undersøger sagen. Vi skelner derfor mellem modellens *forventede værdi* og en *observeret værdi* hentet fra virkeligheden.

En matematisk model beskriver sammenhængen mellem temperaturen og trykket i en beholder ved

$$f(x) = 1,65 \cdot x - 273,$$

hvor $f(x)$ angiver temperaturen (målt i °C) ved trykket x (målt i bar).

Den *forventede* temperatur ved et tryk på 200 bar kan bestemmes: $f(200) = 57$.

Imidlertid kan det ske at vi ved et praktisk eksperiment indstiller trykket i beholderen til 200 bar og derpå måler temperaturen til at være 59°C.

Vi har altså en *forventet* værdi på 57 og en *observeret* værdi på 59.

Vi kan nu beregne den *absolutte afvigelse* (kaldet *residualet*) mellem observeret og forventet værdi:

$$\text{absolut afvigelse} = \text{observeret} - \text{forventet}$$

$$\text{absolut afvigelse} = 59 - 57 = 2$$

Den absolutte afvigelse siger dog ikke i sig selv noget om hvor vidt afvigelsen er stor eller lille. Dette afhænger nemlig af hvor stort et tal den er en afvigelse fra. Vi beregner derfor den *relative afvigelse*, hvor afvigelsen måles i procent af det vi forventede:

$$\text{relativ afvigelse} = \frac{\text{observeret} - \text{forventet}}{\text{forventet}} \cdot 100\%$$

$$\text{relativ afvigelse} = \frac{59 - 57}{57} \cdot 100\% = \frac{2}{57} \cdot 100\% = 0,035 \cdot 100\% = 3,5\%$$

Bemærk at hvis observation er mindre end forventning, så bliver både absolut og relativ afvigelse *negativ*.

I et beregninger-vindue kan såvel absolut som relativ afvigelse nemt beregnes, hvis man definerer den funktion der udgør modellen. Se eksempel på figuren til højre.

Af en god model vil vi forvente at den relative afvigelse ikke bliver større end 5% (eller mindre end -5%).

$f(x) := 1.65 \cdot x - 273$	Udført
$f(200)$	57.
$59 - f(200)$	2.
$\frac{59 - f(200)}{f(200)} \cdot 100$	3.50877

Lineære modeller teoretisk set

En lineær model er en lineær funktion $f(x) = a \cdot x + b$ hvor x og $f(x)$ hver repræsenterer målbare størrelser fra virkeligheden, som modellen påstår en sammenhæng mellem.

Modellen er en påstand om, at til en bestemt x -værdi x_i hører y -værdien $f(x_i)$.

Det vil sige at hvis vi observerer $x = x_i$ så kan vi ifølge modellen *forvente* $y = f(x_i)$.

Hvis vi i virkeligheden observerer at der til $x = x_i$ hører $y = y_i$ kan vi finde den absolutte afvigelse mellem observation og forventning, også kaldet for *residualet*, ved formlen:

$$r_i = y_i - f(x_i)$$

Vi kan endvidere bestemme den relative afvigelse mellem observation og forventning:

$$p = \frac{y_i - f(x_i)}{f(x_i)} \cdot 100\% = \frac{r_i}{f(x_i)} \cdot 100\%$$

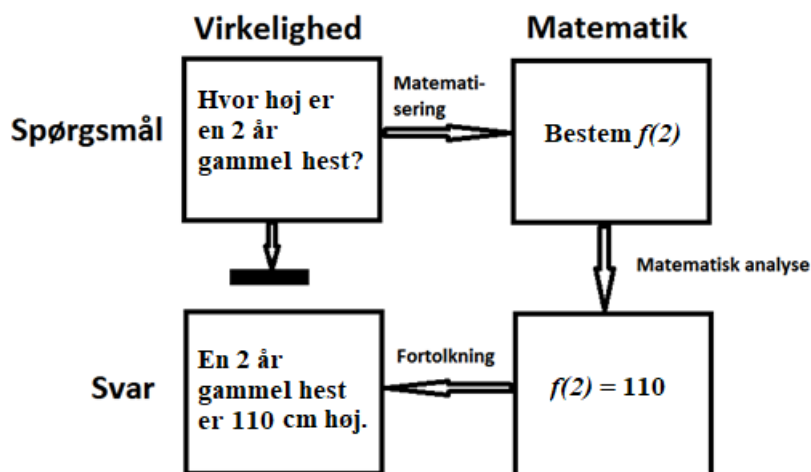
Af en god model vil vi generelt forvente: $-5\% \leq p \leq 5\%$

Fra teorien kan vi fortolke tallene a og b på følgende måder:

”Når x vokser med 1, så vokser $f(x)$ med a ” og ”når $x = 0$ så er $f(x) = b$ (dvs. $f(0) = b$)”.

Generelt kan en matematisk model siges at være en forbindelse mellem *virkeligheden* og *matematikken*, som gør det muligt at besvare spørgsmål fra virkeligheden via matematikken.

Hvis vi f.eks. ønsker at bestemme højden på en 2 år gammel hest, støder vi mod en mur hvis vi bare prøver at tænke os om.



Hvis vi derimod har en *matematisk model* kan vi *matematisere* spørgsmålet ind i modellen, så det virkelige spørgsmål oversættes til matematik-spørgsmålet ”bestem $f(2)$ ”.

Besvarelse af matematiske spørgsmål (kaldet *matematisk analyse*) leder frem til et matematisk svar.

Et matematisk svar kan ved en *fortolkning* oversættes tilbage til et svar på det virkelige spørgsmål.