

## 7 – Lineær regression i Nspire

Ofte ser vi, at fx befolkningsudvikling ikke beskrives ved en forskrift, men ved en række *data*, hvorefter vi selv skal finde den bedste lineære funktion til at beskrive data. Dette kaldes *lineær regression*. Et eksempel kunne være

Opgave:

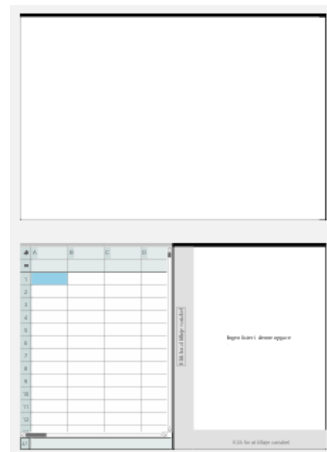
År	2005	2006	2007	2008	2009	2010
Øl (mio. liter)	574	554	527	505	476	453

Tabellen viser danskernes forbrug af øl i perioden 2005-2010.

Denne udvikling kan med god tilnærmelse beskrives ved en lineær model.

- Benyt alle tabellens data til at opstille en lineær model, der beskriver udviklingen i danskernes forbrug af øl som funktion af antal år efter 2005.
- Tegn et residualplot og bestem den største afvigelse mellem modellen for danskernes ølforbrug og de observerede værdier
- Hvor meget falder danskernes forbrug af øl pr. år ifølge modellen?  
Bestem danskernes forbrug af øl i 2015 ifølge modellen

### Sideopsætning og indskrivning af data:



Vi arbejder os gennem opgaven sammen:

Start med at åbne Nspire og lav fx en side til **Noter**. Vælg derefter **INDSÆT > SIDE** (vigtigt at du ikke tilføjer OPGAVE) og vælg her en side delt mellem **Lister og regneark** og **Statistik og diagrammer**

Start med at skrive data ind i to søjler i regnearket.

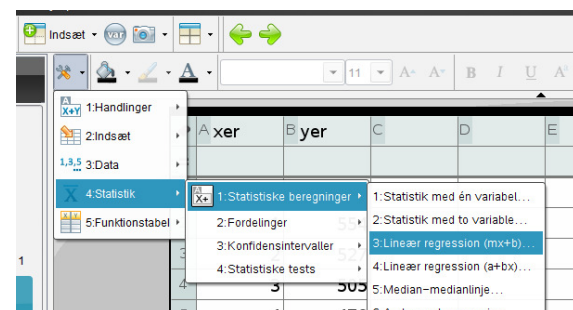
Den første søjle angiver *antal år efter 2005* som vi i modellen er vores x-værdi. Vi skal give søjlen et navn og vi kan kalde den fx 'xer' eller 'xx' eller 'aar'. Du må ikke kalde den 'x', fordi det vil give dig problemer senere

	A xer	B yer	C
=			
1		0	574
2		1	554
3		2	527
4		3	505
5		4	476
6		5	453
7			
8			

Tilsvarende med den anden søjle, som angiver forbrug af øl i millioner liter i Danmark. Vi skriver data ind og angiver et navn til listen fx 'yy' eller 'yer' eller 'forbrug'

Dit regneark skal se således ud når du har indskrevet det hele.

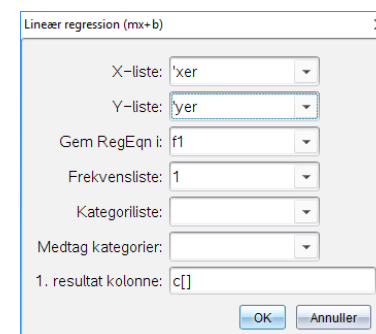
### Lav lineær regression i LISTER OG REGNEARK og hent den ind i NOTER



Klik i det felt lige under 453 (det der er blå på billedet og vælg:

**Dokumentværktøjer > Statistik > Statistiske beregninger > Lineær regression**

Det er ligegyldigt om du vælger nr 3 eller 4 på listen



Nu åbner følgende boks, hvor du bare skal finde dine to lister i de to øverste felter.

Bemærk at der står **Gem RegEqn i f1**

Det betyder at Nspire vil kalde den funktion vi leder efter for f1(x). Du kan vælge at skrive fx f i stedet for f1. Så definerer Nspire funktionen med det navn. Jeg lader det bare stå her.

Tryk OK

Derefter udregner Nspire den bedste model:

Du behøver ikke se nærmere på dette lige nu.

Nu går vi tilbage til forsiden (Du kan bladere med ctrl+pil højre/venstre) og åbner en mat-box i **Noter** og skriver **f1(x)**. Nu skal der gerne komme til at stå en forskrift

$$f1(x) \rightarrow 576.333 - 24.6 \cdot x$$

Nu har du udregnet det der bliver spurgt om i spm a) men skal lige have tekst med der præsenterer problem og svaret:

**Opgave:** Data på næste side viser danskernes forbrug af øl (målt i mio l) i perioden 2005–2010. Vi antager at det er en lineær model.

**a) Bestem ud fra data en model for ølforbruget som funktion af antal år efter 2005.**

Vha lineær regression i regnearket på næste side har vi bestemt forskriften til

$$f1(x) \rightarrow 576.333 - 24.6 \cdot x, \text{ hvor } x \text{ er år efter 2005 og } f1(x) \text{ er forbrug målt i mio liter.}$$

## Lav residualplot og udskriv listen med residualer

For at svare på spørgsmål b) skal vi tegne et residualplot, der er et plot over punkternes afvigelser fra linjen. Det gør vi i STATISTIK-vinduet (Diagrammer og statistik), samtidig med at vi kontrollerer svaret fra spørgsmål a).

Gå ind i **Statistikvinduet** og klik på firkanten under x-aksen, og vælg det du kaldte din x-liste (jeg har kaldt det **xer**), og gør det tilsvarende på y-aksen

Gå derefter - mens du står i STATISTIKvinduet - ind i **dokumentværktøjer > Undersøg data > Regression > Vis lineær (m+b)**

Nu får du tegnet linjen ind med forskrift.

	A xer	B yer	C	D	E
=				=LinRegM	
1	0	574	Titel	Lineær r...	
2	1	554	RegEqn	m*x+b	
3	2	527	m	-24.6	
4	3	505	b	576.333	
5	4	476	r <sup>2</sup>	0.998065	
6	5	453	r	-0.9990...	
7			Resid	{-2.3333...	

Vælg nu - mens du stadig står i regressionsvinduet **dokumentværktøjer > Undersøg data > Residualer > Vis residualplot** hvorefter afvigelserne fra linjen tegnes under grafen (se resultatet næste side).

Vi ser af residualplottet at afvigelserne ligger mellem -2 og 2. For at svare præcist på hvad den største afvigelse er, kan det være rart at få udskrevet residualerne som tal.

Åben et matematik-felt og skriv **stat. ...** og vent

Nu åbner en liste med de muligheder vi har for at vælge.

Klik på **7:resid** og du får svaret

$$\text{stat.Resid} \rightarrow \{-2.33333, 2.26667, -0.133333, 2.46667, -1.93333, -0.333333\}$$

Vi ser at den største afvigelse er på 2.46667 og at det er det svarer til den 4 observation hvor x=3 dvs i år 2008.

Vi kan også få Nspire til selv at finde største og mindste værdi med en **max-** og en **min-**kommando. Dette kan være en fordel ved større datasæt

$$\max(\text{stat.Resid}) \rightarrow 2.46667$$

$$\min(\text{stat.Resid}) \rightarrow -2.33333$$

Vi kan nu svare på spm b)

**b) Tegn et residualplot og bestem den største afvigelse mellem modellen og de observerede værdier**

$$\text{Residualerne er } \text{stat.Resid} \rightarrow \{-2.33333, 2.26667, -0.133333, 2.46667, -1.93333, -0.333333\}$$

Den største afvigelse er på 2.47 (mio liter) i 2008. Plottet er på næste side

I spørgsmål c) spørges der om hvor meget danskernes ølforbrug falder med om året og hvad forbruget - iflg modellen - er i 2015. Hældningskoefficienten svarer netop til det ølforbruget falder med pr år, og værdien i 2015 svarer til f1(10), idet 2015 er 10 år efter 2005. I kan se den endelige og færdige besvarelse på næste side

**Opgave:** Data på næste side viser danskernes forbrug af øl (målt i mio l) i perioden 2005–2010. Vi antager at det er en lineær model.

**a) Bestem ud fra data en model for ølforbruget som funktion af antal år efter 2005.**

Vha lineær regression i regnearket på næste side har vi bestemt forskriften til

$f_1(x) = 576.333 - 24.6 \cdot x$ , hvor  $x$  er år efter 2005 og  $f_1(x)$  er forbrug målt i mio liter.

**b) Tegn et residualplot og bestem den største afvigelse mellem modellen og de observerede værdier**

Residualerne er **stat.Resid** =  $\{-2.33333, 2.26667, -0.133333, 2.46667, -1.93333, -0.333333\}$

Den største afvigelse er på 2.47 (mio liter) i 2008. Plottet er på næste side

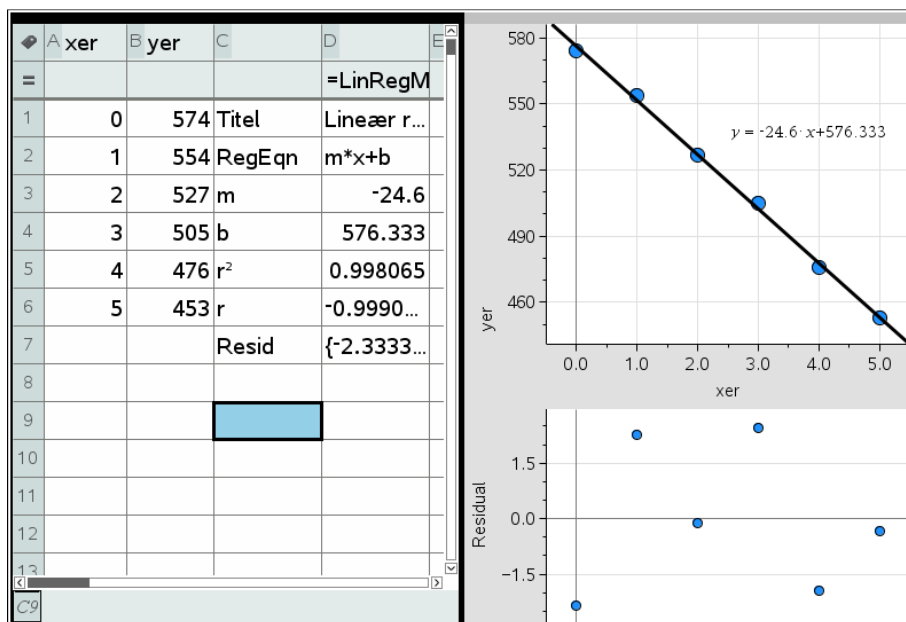
**c) Hvor meget falder danskernes forbrug af øl pr. år ifølge modellen?**

Bestem danskernes forbrug af øl i 2015 ifølge modellen

Af hældningskoefficienten  $f_1(x)$  aflæses, at forbruget af øl falder 24.6 mio l pr år.

I 2015 svarer det til  $x=10$ :  $f_1(10) = 330.333$

Forbruget i 2015 er på 330 mio l iflg modellen



## Opg 1:



Befolkningsudviklingen i Sverige kan beskrives ved tabellen

Årstal	1960	1970	1980	1990	2000	2010
Anal (i mio.)	7,49	8,04	8,31	8,56	8,87	9,38

[Bemærk at talmateriale i eksamens-opgaver altid angives med ”komma”, som er dansk standard, mens vores programmer altid regner med ”punktum”, som er amerikansk/engelsk standard. Du skal altså altid skrive data ind med punktum i stedet for komma]

Vi antager at befolkningsantallet i Sverige kan skrives på formen

$$f(x) = a \cdot x + b$$

hvor  $x$  er antal år efter 1960 og  $f(x)$  angiver befolkningstallet i mio.

a) Bestem ud fra datasættet konstanterne  $a$  og  $b$ . Hvad fortæller  $a$  om befolkningsudviklingen.

NB! Husk at  $x$ -listen har 0 i første felt og ikke 1960

b) Lav et residualplot og angiv den største afvigelse mellem model og data

c) Bestem befolkningstallet i 2020 og angiv hvornår befolkningstallet runder 10 millioner iflg modellen

### Er den lineære model god?

Vi kan vurdere en lineær model ud fra data-plot og residual-plot. Hvis punkterne ligger tæt på linjen og fordelt tilfældigt omkring linjen i punktplottet er modellen god.

Det sidste viser sig også ved at punkterne i residualplottet ligger tilfældigt rundt om  $x$ -aksen.

**Opg 2:** Under et langvarigt regnvejr aflæses en regnmåler hver halve time. Nedenstående tabel viser aflæsningerne.

Tid (timer)	0	0,5	1	1,5	2	2,5	3
Regnmængde (mm)	3	9	19	31	39	50	57

I en model beskrives regnmængden ved en lineær funktion

$$f(x) = a \cdot x + b$$

hvor  $f(x)$  er regnmængden (målt i mm) og  $x$  er tiden (målt i timer), efter at aflæsningerne begyndte.

- Tegn et punktplot af tabellens data
- Bestem tallene  $a$  og  $b$  ved lineær regression
- Tegn et residualplot, og kommenter på baggrund heraf om den lineære model er god til at beskrive regnmængden som funktion af tiden



**Opg 3:** Tabellen viser en opgørelse over antallet af praktiserende læger i Nordjylland i perioden 2010-2017

År efter 2010	0	1	2	3	4	5	6	7
Antal praktiserende læger	352	343	336	340	322	311	307	303

I en model kan antallet af praktiserende læger i Nordjylland beskrives ved

$$f(x) = a \cdot x + b$$

hvor  $f(x)$  er antallet af praktiserende læger  $x$  år efter 2010.

- Bestem tallene  $a$  og  $b$  ved lineær regression.
- Tegn residualplot.
- Bestem det årstal, hvor der er størst forskel mellem det observerede antal praktiserende læger og den tilsvarende modelværdi, og bestem denne forskel.
- Forklar hvad tallet  $a$  siger om antallet af praktiserende læger i Nordjylland. (NB! Brug altid det konkrete tal i dit svar)

Hvis vi kun har to punkter, så går *den bedste rette linje* lige gennem disse to punkter. Dermed kan vi bruge regression til at finde forskriften gennem to punkter:

**Opg 4:** For en lineær funktion gælder at den går gennem punkterne (2,7) og (9,33)

$x$	2	9
$y$	7	33

- Bestem forskriften for funktionen ved regression
- Bestem forskriften vha to-punkts-formlerne
- Bestem forskriften for funktionen med graf gennem (29, 54) og (43, 80)

### Facitliste:

Opg 1: a)  $a=0.035$   $b=7.571$  Befolkningen vokser med 0.035 mio pr år  
b) Største-afvigelse i er på 0.121 mio personer (det var i 1970) c) 9.66mio ca 2030

Opg 2: b)  $f(x) = 18.9 \cdot x + 1.4$  c) God model. Punkterne ligger tæt på linjen i dataplot og tilfældigt omkring linjen

Opg 3: a)  $a = -7.3$   $b = 352.4$  c) ved  $x=3$  dvs i 2013 er afvigelsen 9.6  
d) Antallet af praktiserende læger falder i gennemsnit med 7.3 pr år

Opg 4: a+b)  $f(x) = 3.71 \cdot x - 0.43$  c)  $f(x) = 1.86x + 0.14$