Kurvetilpasning og modellering

Stikkord:

Kurvetilpasning, korrelasjonskoeffisient, sum av kvadratavvik.
Kurvetilpasning kalles ofte regresjon, men regresjon er et statistisk begrep som omhandler den statistiske og kausale sammenhengen mellom to variabler, altså om det er en tilfeldig sammenheng eller en årsakssammenheng mellom de to variablene.

De forskjellige kurvetilpasningene vi kan gjøre med ***lommeregnere*** er vist på side 188 i læreboken.

På ***GeoGebra*** har vi tilsvarende:

("listenavn" betyr navnet på listen vi har lagt datapunktene i.)

Funksjonstype:	**Kommando**:	**Krav til x og y**:
Lineær: $ax + b$	RegPoly[listenavn,1]	Ingen
Andregrad: $ax^2 + bx + c$	RegPoly[listenavn,2]	Ingen
$ax^n + bx^{n-1} + \ldots$	RegPoly[listenavn,n]	Ingen
Logaritmisk: $a + b\ln x$	RegLog[listenavn]	$x > 0$
Eksponentiell: ae^{bx}	RegEksp[listenavn]	$y > 0$
Potens: $ax\hat{}b$	RegPot[listenavn]	$x > 0, y > 0$
Sinus: $a\sin(bx + c) + d$	RegSin[listenavn]	Ingen
Logistisk: $\frac{a}{1+be^{cx}}$	RegLogist[listenavn]	Ingen

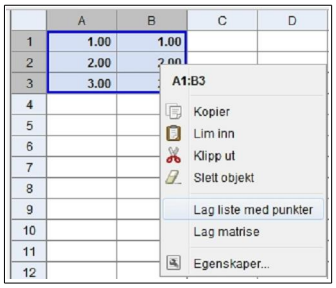
Viktig:
Må ha minst like mange punkter som parametere $(a, b, c, \ldots)$ for at kurvetilpasningen skal kunne regnes ut.

Fremgangsmåte:

Vi skal lage en rett linje og en andregradskurve gjennom punktene gitt av tabellen:

x :	1	2	3
$f(x)$:	1	2	2.6

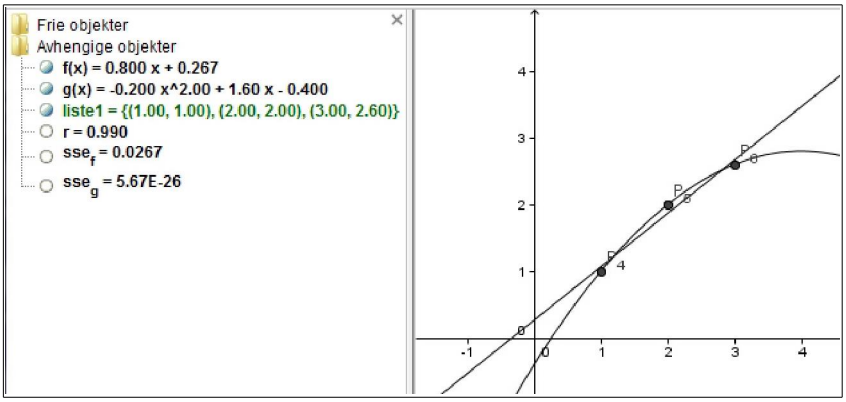
1. Vi legger inn punktene i regnearket. (**Menyvalg**: **Vis**, **Regneark**)
2. Markerer, høyreklikker og velger **Lag liste med punkter**:



GeoGebra lager da en liste med navnet liste1= {(1,1),(2,2),(3,2.6)}
og legger punktene inn i koordinatsystemet.

3. Gjør kurvetilpasningen med:
   f(x)=RegPoly[liste1,1] (lineær)
   g(x)=RegPoly[liste1,2] (andregrad)

Vi får da:



Her er også lagt inn:

Korrelasjonskoeffisient: **r=Korrelasjonskoeffisient[liste1]**

Sum av kvadrater av feilavvik: **sse_f=Sum[(f(x(liste1))-y(liste1)) ^2]**
sse_g=Sum[(g(x(liste1))-y(liste1)) ^2]

(Matematisk definisjon:
Sum Squared Errors: $sse = \sum_i (f(x_i) - y_i)^2$, der (x_i, y_i) er punkt nummer i.)

Kommentarer:

Tabellen for lommeregnere side 188 i læreboken viser at lommeregnere regner ut korrelasjonskoeffisienten for flere typer kurvetilpasning, selvom den egentlig ikke har noe med kurven å gjøre, bare med en eventuell statistisk sammenheng mellom x – og y –verdier.
Legg merke til at GeoGebra bare trenger listenavnet for å regne ut korrelasjonskoeffisienten!

Korrelasjonskoeffisienten har i seg selv ingen mening, det er kvadratet r^2 som har mening!

r^2 betyr andelen av dataene som støtter en sammenheng:

$r = 0.9$ betyr at $r^2 = 0.81 = 81\%$ av datamaterialet støtter en lineær sammenheng,

mens 19% antageligvis skyldes rene tilfeldigheter...

I vårt eksempel passer andregradsfunksjonen $g(x)$ perfekt og har derfor $sse \approx 0$,
mens den lineære funksjonen $f(x)$ har $sse = 0.0267$ og er derfor dårligere enn $g(x)$,
selvom korrelasjonskoeffisienten er $r = 0.99$!

Vurdering av modeller laget med kurvetilpasning:

-Lærebøkene snakker bare om korrelasjonskoeffisienten r og sier modellen er "god"
hvis r ikke er altfor langt unna 1.
Dette er greit hvis man bare vurderer en ***lineær*** modell.

-Hvis man har flere ***alternative*** kurver å velge mellom hjelper ikke r, da
den bare avhenger av dataene, ikke funksjonen.
(Lommeregneren regner ut forskjellige r for forskjellige kurver, men dette
er egentlig direkte feil, av grunner det vil ta for lang tid å forklare her.)
Da er det bedre å bruke sse for sammenligning, og velge kurven med minst sse.

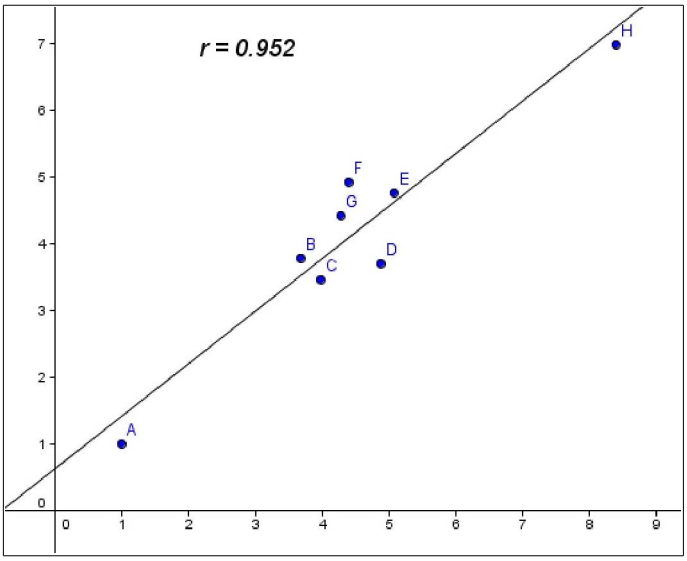
I tillegg bør man:

- ***Plotte*** punktene og ***se på*** plottet!
  - Ser modellen fornuftig ut?
  - Er punktene rimelig jevnt fordelt over det x –område man studerer? (Spredning? Konsentrasjon?)
  - Tyder punktene på at modellen er periodisk? (Svingning?)
  - Er noen punkter åpenbare feilmålinger?
  - Stemmer det bra med startverdier, for eksempel for $x = 0$?
  - Stemmer det bra for store x, for eksempel når det har gått lang tid? Flater ut? Øker? Minker?
    (Radioaktivitet bør gå mot null, populasjoner stabiliserer seg ofte asymptotisk,...)

Et siste spark til korrelasjonskoeffisienten: :-)

Hvis ikke punktene er jevnt fordelt, kan de mest ekstreme verdiene gi altfor god r.
En test, som selv profesjonelle ofte glemmer, er å fjerne de mest ekstreme verdiene og
se hva som skjer med r.

Eksempel:



Uten *A* og *H*: