Lineær regression

Regression:

regression [rægræʒsjoɔn] sb. -en, -er, -erne
□ = TILBAGEGANG c PROGRESSION
ETYMOLOGI: fra latin *regressio* 'tilbagegang' afl. af *regredi* 'gå tilbage' (jf. *regredere*)

Kilde: Politikens Nudansk Ordbog med etymologi, 3. udg.

En regression kan på matematik'sk siges at betyde at man tilpasser en graf til en række målepunkter, eller for at tage ordbogens betydning, at man går tilbage til en kurve med udgangspunkt i en række (måle)punkter.

Her diskuteres udelukkende "Lineær regression", hvilket begrænser regressionsprincippet til, at de udvalgte punkter skal ligge på **en ret linje**.

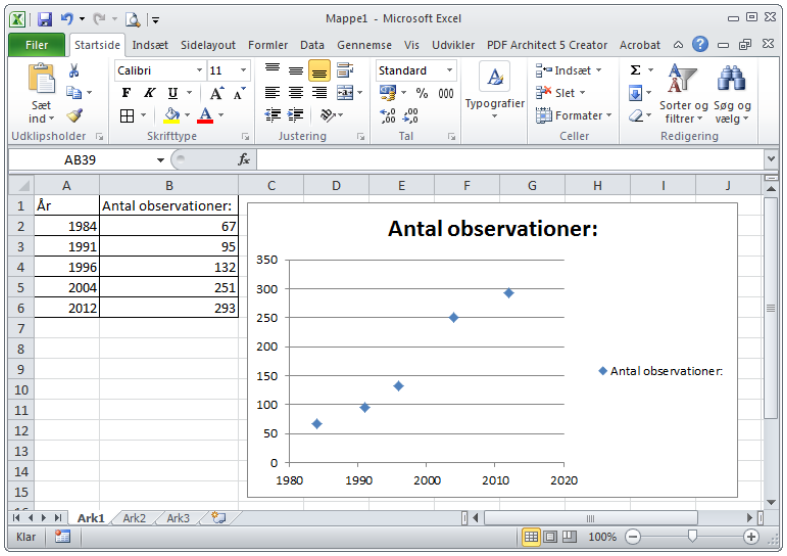
I den virkelige verden, kan man vælge (næsten) at benytte en hvilken som helst funktionstype til sin regression. Dette er praktisk, hvis man f.eks. antager en eksponentiel befolkningstilvækst.

Følgende regneeksempel er taget fra: Mat B1 for HTX, Systime.

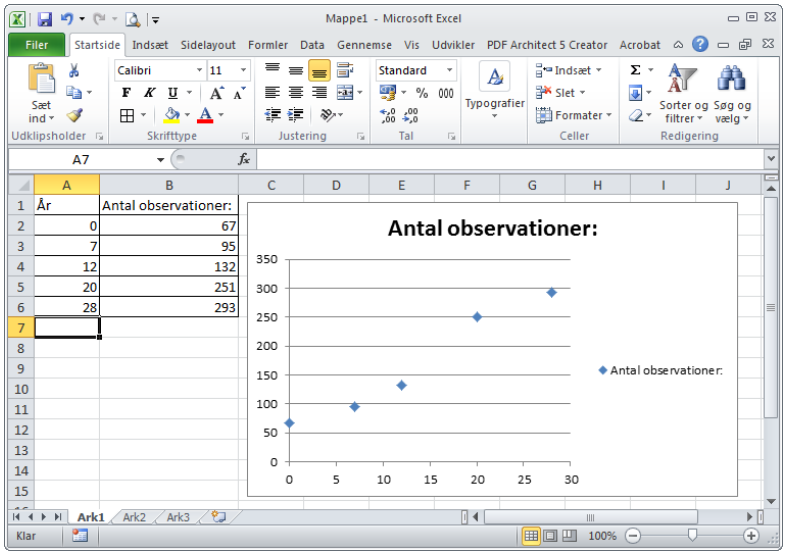
Eksempelmaterialet består af en række talpar (x;y) hvor abcsisseaksen repræsenterer årstallet og ordinataksen repræsenterer antallet af observerede oddere i Danmark i perioden fra år 1984 til år 2012.

Årstal:	1984	1991	1996	2004	2012
Antal observationer:	67	95	132	251	293

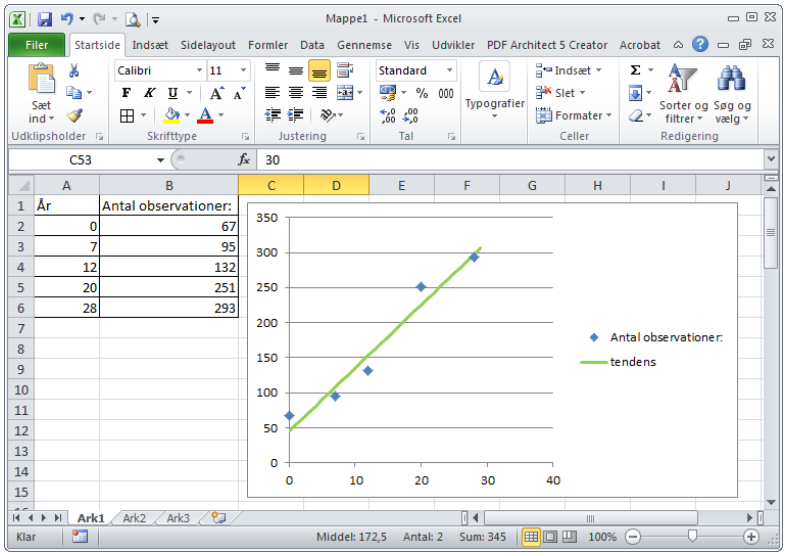
Bortset fra den åbenlyse konklusion at Danmark er ved at blive oversvømmet af oddere, så kan punkterne indtegnes i et koordinatsystem:



Det kan være en fordel at regne årstallene som afstanden i år fra det første årstal (1984), som derfor ”omdøbes” til år 0.



Der indlægges en grøn linje med ligningen $y = 9t + 45$. Denne linje kaldes for en tendenslinje, idet den kan bruges til at forudse fremtidige punkter (tendensen) i det omfang at udviklingen fortsætter i samme tempo.



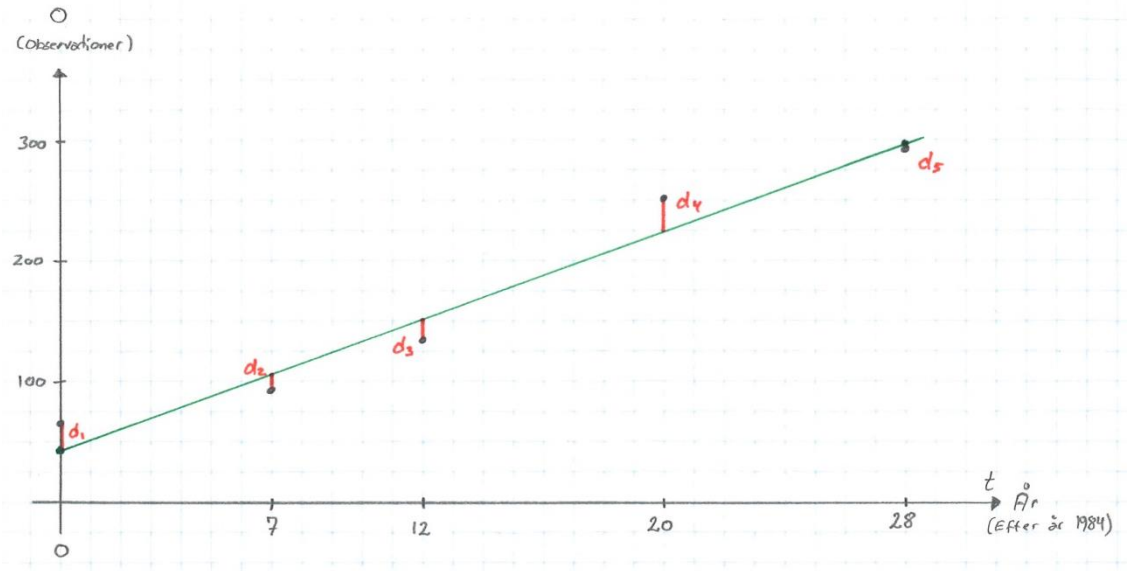
Tendensen vil f.eks. pege på, at der i år 2024 vil være 405 oddere i Danmark, men dette er meget usikkert, da det er ret langt ude i fremtiden. Der er desuden mange ukendte parametre. Modellen kan f.eks. ikke forudsige, hvad der sker med oddernes levevilkår i Danmark.

Dette eksempel er en simpel lineær matematisk model, hvor væksthastigheden er fastlagt ($a = 9$). Spørgsmålet er nu, om der findes en matematisk metode til at optimere denne proces, så punkterne bliver tilnærmet så godt som muligt.

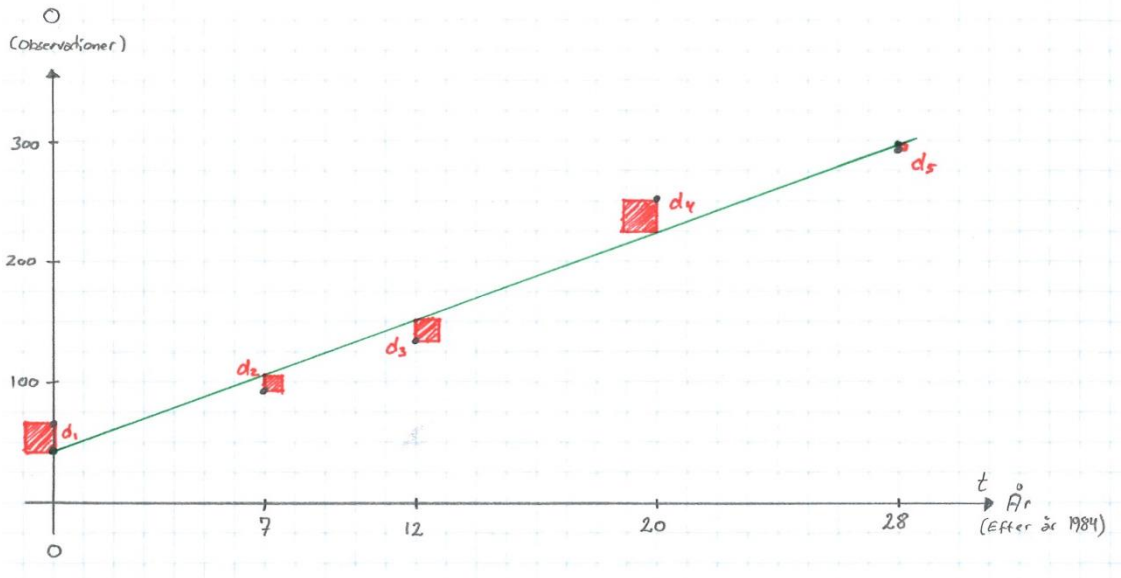
Der er flere metoder, som kan anvendes til netop dette formål, og en af de mest brugte kaldes for ”Mindste kvadraters metode”.

Mindstekvadraters metode er en matematisk måde at bestemme den bedste rette linje på. Metoden undersøger den lodrette afstand fra de kendte punkter i koordinatsystemet til en given linje.

På næste billede (der skiftes til håndtegninger), ses de samme talpar, som er benyttet i eksemplet. Dog er der her indtegnet de lodrette afstande fra de kendte punkter i koordinatsystemet til den tegnede rette linje.



De lodrette afstande kaldes også *residualer*. De røde kvadrater er en illustration af den samme lodrette afstand, angivet som siderne på et kvadrat. Man kalder også arealet af kvadratet for den *kvadratiske afvigelse* af talsættet.



Når summen *D* af alle kvadraternes arealer for en given linje er **mindst mulig**, siges den rette linje at være den bedste rette linje til at repræsentere tallene.

$D = \sum_{i=1}^{n} d_i^{\,2} = d_1^{\,2} + d_2^{\,2} + d_3^{\,2} + \cdots + d_n^{\,2}$, hvor *n* er antallet af målepunkter.

i dette tilfælde:

$$D = \sum_{i=1}^{n} d_i^{\,2} = d_1^{\,2} + d_2^{\,2} + d_3^{\,2} + d_4 + d_5^{\,2}$$

Ændrer man den rette linje (tendenslinjen), så vil man samtidig også ændre punkternes residualer og dermed også den kvadratiske afvigelse af talsættet.

Korrelationskoefficienten

Vil man have en ide om, h vor godt modellen passer til de givne måledata, kan man beregne **korrelationskoefficienten, *r***. (Det sker, at denne værdi også betegnes med et stort *R*).

Det er en god ide at beregne *r*, da selve regressionsligningen (tendenslinjen) ikke i sig selv udtrykker hvor god lineær sammenhæng der er mellem variablene *x* og *y* i datasættet. Man kan jo altid lave en tendenslinje uanset hvordan og hvor meget målepunkterne ligger spredt.

Uden at gå for meget i detaljer om, hvordan *r* udregnes (det er en lang proces), så viser det sig, at *r* kan blive enten positiv eller negativ. Derfor kvadrerer man ofte *r*, hvilket bliver til r^2 (eller R^2) når man sammenligner korrelationer.

r^2 kaldes for **forklaringsgraden**.

Kvadratet på *r*, (r^2), vil altid være i intervallet fra 0 til 1. $\left(r \in [0;1]\right)$.

Hvis r^2 er lig med 0 eller tæt på 0, så er der slet ingen sammenhæng mellem måledata og modellen (tendenslinjen). Er værdien derimod lig med 1, så passer alle punkter nøjagtig med modellen. Så hvis r^2 er lig med 1, så er tendenslinjen den optimale rette linje.
Jo tættere på 1 r^2 er, desto bedre beskriver tendenslinjen målepunkterne. Hvis værdien er tæt på 1, siger man at der er **høj lineær korrelation mellem de to variable *x* og *y***.

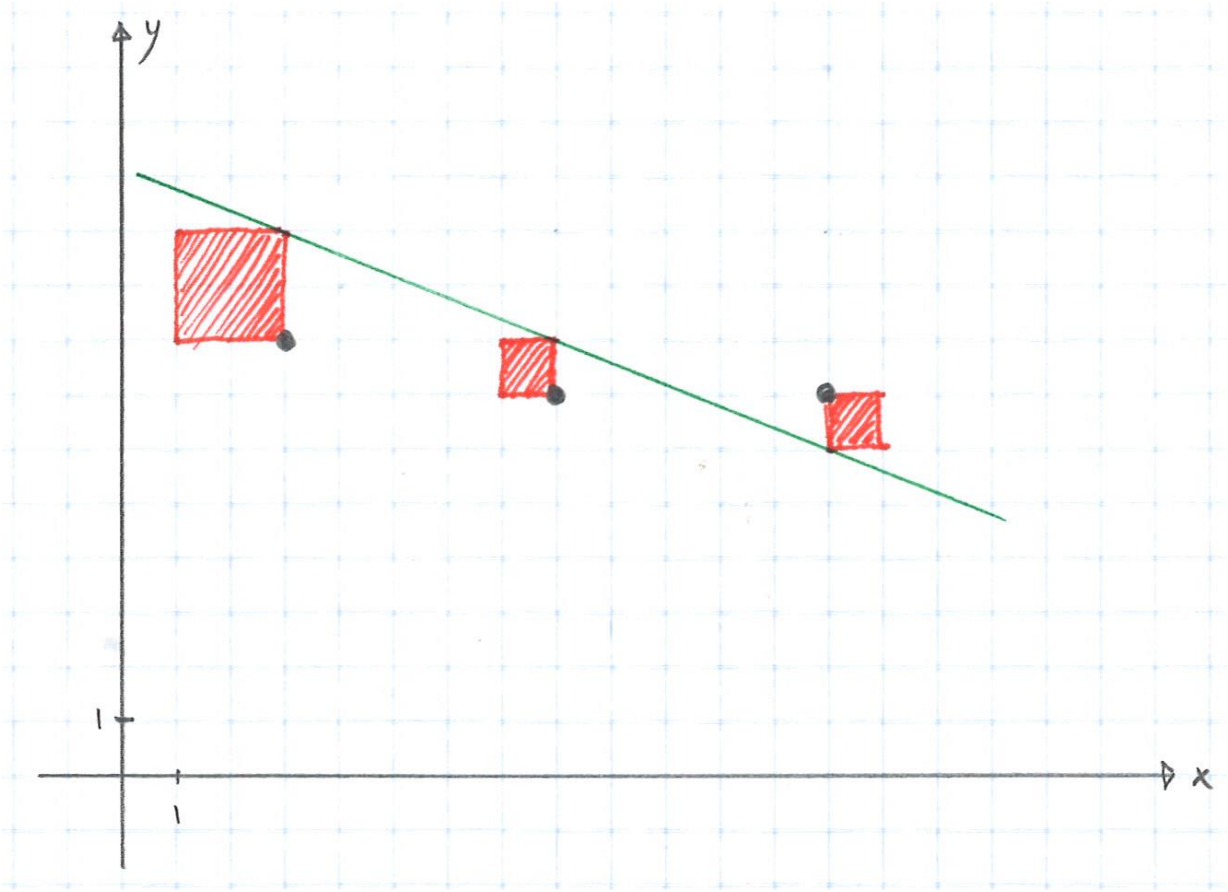
Så for at samle lidt op:
Lineær regression er et vigtigt redskab til at bygge matematiske modeller med.
Når man med lineær regression bestemmer en ret linje med høj korrelationskoefficient, r^2, så siger det noget om, at linjen matematisk set passer godt med talmaterialet.

Dog siger resultatet ikke noget om, hvorvidt det man undersøger rent faktisk er linært. En høj korrelations betyder ikke, at den ene variabel *x* nødvendigvis er årsag til den anden variabel *y*.

Det lyder vel – om ikke selvmodsigende – så vel en smule kryptisk, men det giver mening. Ud fra det tidligere eksempel kan man f.eks. ikke sige, at antallet af oddere i Danmark vokser lineært afhængigt af tiden, men det er helt fint at påstå at udviklingen af oddere KAN beskrives SOM OM der var en lineær sammenhæng (i den observerede periode).

Øvelser:

Øvelse 01



a) Bestem tallet *D*.

b) Find en bedre ret linje end den grønne, og angiv forskellen på de to linjer.
Prøv evt. til sidst at taste oplysningerne ind i f.eks. Excel, og prøv at give et bud på, hvilken af tendenslinjerne som er bedst – og hvorfor …

Øvelse 02

En ligefrem proportionalitet, som går igennem $(x;y)=(17;13)$ antages at være en tendenslinje for punktsættet:

$P_1=(2;2)$, $P_2=(7;6)$, $P_3=(10;5)$ & $P_4=(14;10)$

Bestem tallet *D*.

Husk, at residualet udregnes som den numeriske værdi af punktets *y*-værdi subtraheret med linjens funktionsværdi i punktets *x*-værdi.

Find en bedre tendenslinje.

Brug evt. (til sidst) Excel (eller andet program) til at bestemme en tendenslinje.

Øvelse 03

Sammenhængen mellem ugentlig salg i tusinde kr. og testscores for en stikprøve bestående af 8 salgskonsulenter fremgår af tabellen:

Ugentligt salg	10	12	28	24	18	16	15	12
Test scores	55	60	85	75	80	85	65	60

Bestem den ”bedste” rette linje samt r og r^2 vha. lineær regression.

Øvelse 04

Prisen på DVD-afspillere sættes forskelligt i 8 forskellige regioner af landet, se nedenfor.
Prisen er opgivet i hundrede dollar, se tabellen:

Antal solgt	420	380	350	400	440	380	450	420
Pris	5,5	6,0	6,5	6,0	5,0	6,5	4,5	5,0

Bestem den ”bedste” rette linje samt r og r^2 vha. lineær regression.