

Uddrag fra:

MAT stx Grundforløb

© 2017 Esben Wendt Lorenzen, Michael Jørgensen, Jens Carstensen og Jesper Frandsen og Systime a/s

Kopiering og anden gengivelse af dette værk eller dele deraf er kun tilladt efter reglerne i gældende lov om ophavsret eller inden for rammerne af en aftale med Copydan. Al anden udnyttelse forudsætter en skriftlig aftale med forlaget.

Grafisk tilrettelæggelse og produktion: Systime a/s

Omslag/omslagsillustration: Systime

1. udgave 1. oplag

ISBN 978-87-616-8830-9

systime 

Sonnesgade 11

DK-8000 Aarhus C

Tlf.: 70 12 11 00

systime.dk

LINEÆR REGRESSION

Ofte har man givet en række datapunkter, som ikke ligger på en ret linje. Alligevel vil man gerne benytte en lineær model for at kunne lave prognoser (forudsigelser). Vi ser på et eksempel:

EKSEMPEL 6

Man har undersøgt højden af et stort antal piger og beregnet middelhøjden for hver årgang. Pigerne's højde ved forskellige aldre fremgår af skemaet.

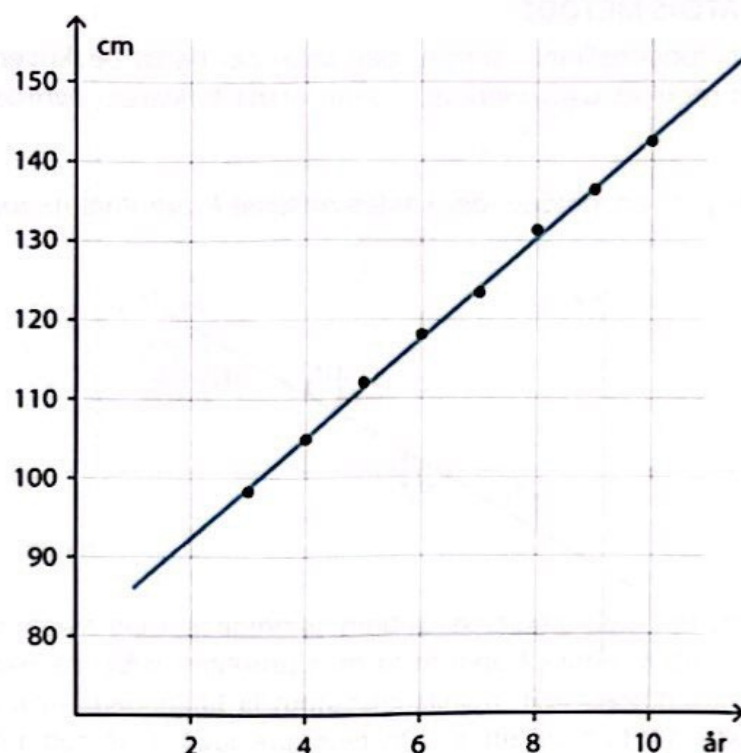
Alder (år)	x	3	4	5	6	7	8	9	10
Højde (cm)	y	98,3	104,9	112,0	118,1	123,4	131,3	136,4	142,5

På figur 9 ovenfor er punkterne indtegnet. Endvidere er *tendenslinjen* tegnet. Tendenslinjen er den linje, der bedst beskriver datapunkterne. I det næste afsnit skal vi se, hvordan tendenslinjen bestemmes – i dette eksempel vil vi blot benytte ligningen:

$$y = 6,31x + 79,8,$$

hvor x er alderen i år, og y er pigerne's gennemsnitlige højde i cm.

Vi ser på figur 8, at linjen følger punkterne med stor nøjagtighed. Modellen givet ved tendenslinjen er derfor gyldig i intervallet 3 år til 10 år. Det kan formodes, at modellen også er gyldig et par år over ud over dette.



Figur 9

Ligningen viser, at væksthastigheden for pigerne er på 6,31 cm om året.

Det er muligt at bruge ligningen for den rette linje til at lave prognoser. Eksempelvis kan man skønne 12-årige pigers gennemsnitlige højde ved at indsætte $x = 12$ i ligningen. Det giver

$$y = 6,31 \cdot 12 + 79,8 = 155,5 \text{ cm.}$$

Altså forudsiger denne model, at piger på 12 år har en gennemsnitlig højde på 155,5 cm. Det skal understreges, at der er tale om en prognose, som bygger på en række forudsætninger. I dette tilfælde forudsætter modellen, fordi det er en *lineær* model, at piger vokser med det samme antal cm hvert år, i dette tilfælde 6,31 cm.

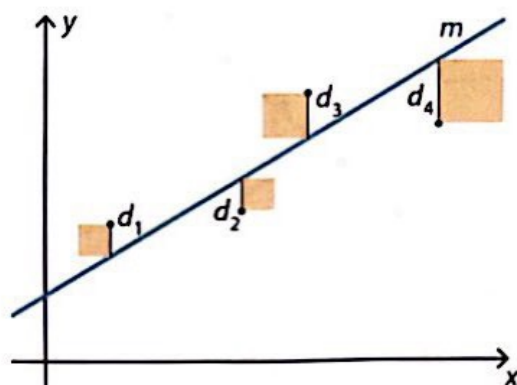
Denne forudsætning holder ikke, når pigerne passerer puberteten. Endvidere er der tale om et gennemsnit. Alligevel bliver sådanne modeller brugt i stor udstrækning, og det er derfor vigtigt at kende modellernes forudsætninger.

I en lineær model forudsættes det, at væksthastigheden er konstant.

MINDSTE KVADRATERS METODE

For at finde frem til tendenslinjen, som er den linje, der bedst beskriver datapunkterne, benytter man som regel et CAS-værktøj. I dette afsnit forklares, hvordan tendenslinjen findes.

CAS-værktøjer benytter en metode, der kaldes *mindste kvadraters metode*.



Figur 10

På figur 10 er en række punkter afsat, og en vilkårlig linje m er tegnet. Desuden er de lodrette afstande $d_1, d_2, d_3, \dots, d_n$ fra målepunkterne til linjen afsat, samt de kvadrater, som de lodrette afstande danner. Man kan så udregne summen D af kvadraternes arealer

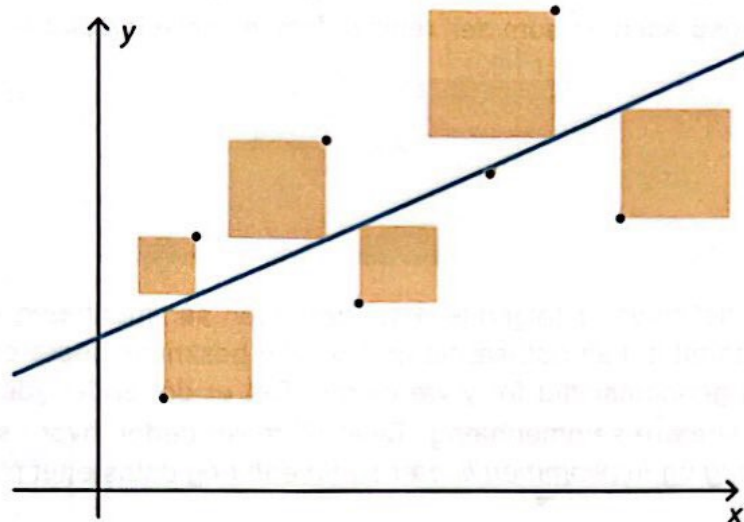
$$D = d_1^2 + d_2^2 + d_3^2 + \dots + d_n^2.$$

Hvis man vælger en anden linje, får man selvfølgelig i reglen en anden kvadratsum.

Man har vedtaget, at den linje, der gør kvadratsummen D mindst mulig, er den "bedste" linje. Metoden kaldes *de mindste kvadraters metode*, og den linje, der fremkommer, kaldes *regressionslinjen* eller *tendenslinjen* svarende til punkterne. Vi går ikke ind på, hvordan linjen matematisk bestemmes.

FORKLARINGSGRAD

Når vi laver lineær regression med et værktøjsprogram, er regressionsligningen i de fleste værktøjsprogrammer ledsaget af *forklaringsgraden* R^2 . Forklaringsgraden er et tal mellem 0 og 1 og fungerer som et mål for, hvor godt regressionsligningen beskriver data: Jo tættere tallet er på 1, jo bedre beskriver regressionsligningen datasættet. En mere præcis definition af R^2 kræver lidt teori.

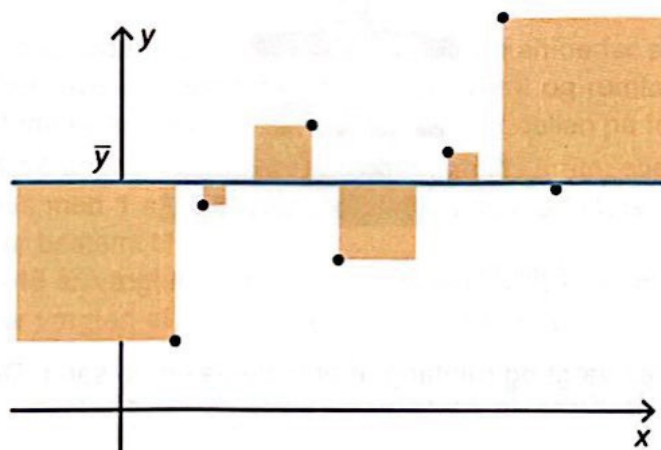


Figur 11

Figur 11 viser en række punkter sammen med *regressionslinjen*. For ethvert af punkterne er den *kvadratiske afvigelse* markeret som et orange kvadrat. Siden i kvadratet er den lodrette afstand fra målepunktet til regressionslinjen. Regressionsligningen er bestemt som ligningen for den linje, hvor summen af alle de kvadratiske afvigelser er mindst mulig. Det er altså ligningen for den linje, der kommer tættest på alle punkterne, når vi som målestok bruger det samlede areal af de orange kvadrater. Vi lader A_r betegne dette mindst mulige samlede areal.

Jo mindre tallet A_r er, jo tydeligere er der en lineær sammenhæng i datasættet. I det ekstreme tilfælde, hvor data faktisk ligger på en ret linje, er der ingen afvigelse. Det giver $A_r = 0$, som derfor er et udtryk for en fuldkommen lineær sammenhæng.

Figur 11 viser de samme punkter som ovenfor, men nu sammen med en vandret linje gennem gennemsnittet $\bar{y}$ af datasættets y -værdier. Punkternes kvadratiske afvigelser til gennemsnittet er også markeret med orange kvadrater. Vi sætter A_g til at være summen af de kvadratiske afvigelser fra gennemsnittet $\bar{y}$. Tallet A_g er altså det samlede areal af de orange kvadrater på figur 12.



Figur 12

Forklaringsgraden fremkommer ved at sammenligne tallet A_r med tallet A_g . Mere præcist defineres forklaringsgraden R^2 som den relative forskel mellem A_g og A_r :

$$R^2 = \frac{A_g - A_r}{A_g}.$$

Ideen bag denne definition er følgende: Hvis der ingen sammenhæng er mellem x - og y -værdierne i datasættet, kan datasættet grafisk ikke beskrives bedre end med en vandret linje gennem gennemsnittet for y -værdierne. Det er det andet yderlighed i forhold til den fuldkomne lineære sammenhæng. Tallet R^2 måler derfor, hvor i spændet mellem *ingen sammenhæng* og *fuldkommen lineær sammenhæng* datasættet placerer sig. Specielt ser vi at:

- Hvis der ingen sammenhæng er, bliver A_r ligeså stor som A_g , dvs.

$$R^2 = \frac{A_g - A_r}{A_g} = \frac{A_g - A_g}{A_g} = \frac{0}{A_g} = 0.$$

- Hvis datasættet udviser nogen lineær sammenhæng, er A_r mindre end A_g , dvs. R^2 får en værdi mellem 0 og 1.
- Hvis datasættet udviser en fuldkommen lineær sammenhæng, er $A_r = 0$, dvs.

$$R^2 = \frac{A_g - A_r}{A_g} = \frac{A_g - 0}{A_g} = 1.$$

EKSEMPEL 7



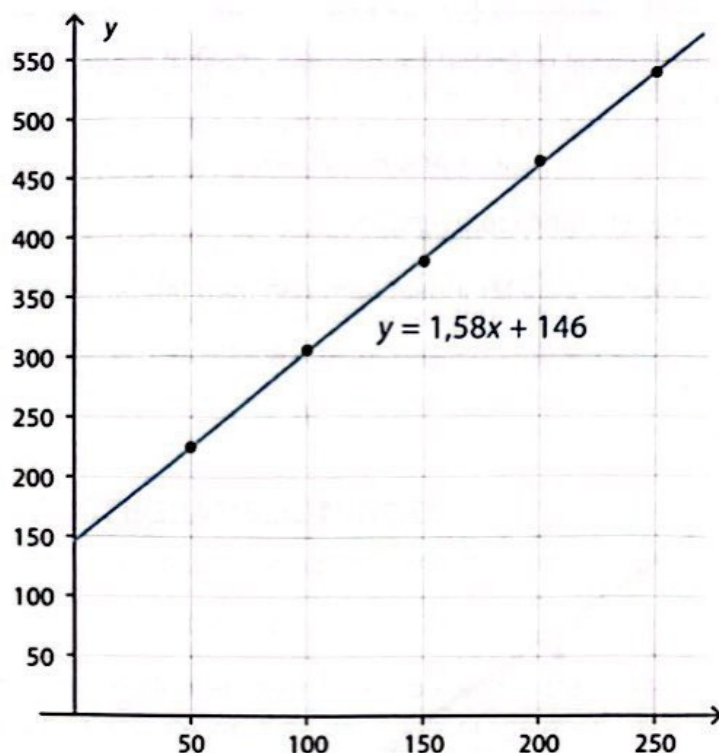
Figur 13

I et fysikforsøg måles vægt og rumfang af et måleglas med sand. Det giver følgende værdier:

Rumfang af sandet (ml)	x	50	100	150	200	250
Vægt (g)	y	225	305	380	465	540

hvor x er rumfanget af sandet i ml, og y er vægten i gram.

Målingerne er vist i figur 14:



Figur 14

På figuren er også indtegnet tendenslinjen givet ved ligningen

$$y = 1,58x + 146.$$

Punkterne passer overordentligt godt med linjen, og vi kan derfor sige, at linjen er en god model til at beskrive sammenhængen mellem vægt og rumfang i dette forsøg. Endvidere kan vi fortolke konstanterne 1,58 og 146 i modellen på følgende måde:

- Konstanten 1,58 angiver, hvor meget vægten (målt i gram) stiger, når rumfanget af sandet øges med 1 ml. Dette er det samme som sandets massefylde, som i dette forsøg er bestemt til 1,58 g/ml.
- Konstanten 146 er vægten (i gram), når måleglasset er tomt. Det vil sige, at i dette forsøg er vægten af det tomme måleglas 146 gram.

Forklaringsgraden er af CAS-værktøjet angivet til at være $R^2 = 0,9997$, hvilket bekræfter, at modellen passer særdeles godt med målingerne. Derfor er det rimeligt at konkludere, at der er en lineær sammenhæng mellem rumfang og vægt af sand, og at sandets massefylde er konstant.

EKSEMPEL 8

I statistikbanken.dk kan man finde data over "tilskadekomne og dræbte i spiritusuheld" i Danmark, som angivet i følgende tabel:

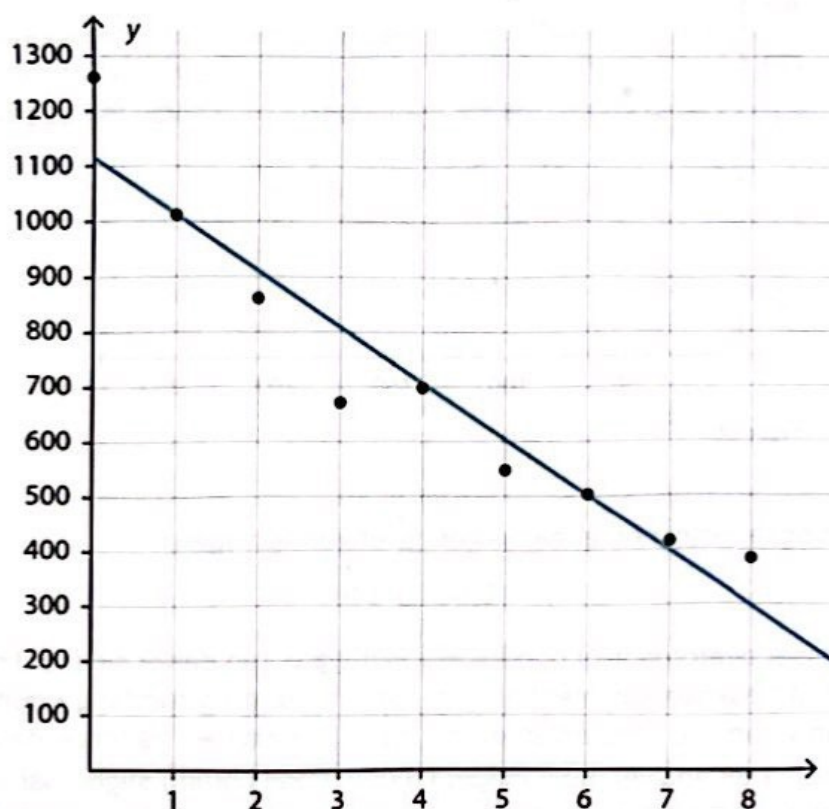
Antal år efter 2007	0	1	2	3	4	5	6	7	8
Årstal	2007	2008	2009	2010	2011	2012	2013	2014	2015
Antal	1261	1012	861	671	698	548	504	417	384

Når den uafhængige variabel er årstal, vælger man oftest at sætte $x = 0$ ved det første årstal. Det vil sige, at

$$x = \text{årstal} - 2007.$$

Dette er ligeledes angivet i tabellen ovenover.

På figur 15 er ovenstående punkter indsat sammen med tendenslinjen.



Figur 15

Tendenslinjen er udregnet på CAS værktøj til

$$y = -102,2x + 1115.$$

Punkterne passer nogenlunde med linjen, og vi kan derfor med et vist forbehold sige, at der er en tilnærmelsesvis lineær sammenhæng mellem antal år efter 2007 og det årlige antal tilskadekomne og dræbte i spiritusuheld i Danmark, i hvert fald inden for

intervallet nul til otte år efter 2007. Dog ligger begge de yderste punkter (ved $x = 0$ og $x = 8$) markant over linjen, og modellen kan derfor næppe bruges udenfor dette interval.

Forklaringsgraden er 0,92 dvs. tæt på 1. Dette antyder en god overensstemmelse mellem data og model.

Konstanterne $-102,2$ og 1115 i modellen kan fortolkes på følgende måde:

- Konstanten $-102,2$ angiver, at hvert år reduceres antallet af tilskadekomne og dræbte i spiritusuheld i Danmark med ca. 102 personer.
- Konstanten 1115 fortæller, at der ved modellens begyndelse, dvs i år 2007, var 1115 tilskadekomne og dræbte i spiritusuheld i Danmark. Dette passer åbenlyst ikke med det rigtige antal på 1261 personer i tabellen herover. Forskellen er et udtryk for, at modellen afviger fra de faktiske antal.

Denne lineære model har dog en indbygget svaghed, nemlig at antallet af personer fortsætter med at falde. På et tidspunkt (efter yderligere ca. fire år) vil modellen angive et negativt antal personer, hvilket er åbenlyst forkert. Så ligesom i det tidligere eksempel med pigers højde, er modellens anvendelighed begrænset.

LØSNING AF FØRSTEGRADSLIGNINGER

En ligning er et udtryk, der indeholder et lighedstegn, som f.eks.

$$5 + 7 = 12 \quad , \quad a + b = b + a \quad , \quad 2x - 3y = 7 \quad , \quad 3x - 4 = 11.$$

En ligning indeholder i reglen en eller flere ubekendte størrelser. At løse en ligning vil sige at bestemme den eller de værdier, der passer i ligningen, dvs. gør den sand, når de indsættes. Ligningen

$$3x - 4 = 11$$

har $x = 5$ som løsning, fordi 5 passer i ligningen:

$$3 \cdot 5 - 4 = 11.$$

På samme måde er $(x, y) = (5, 1)$ og $(x, y) = (8, 3)$ løsninger til ligningen

$$2x - 3y = 7,$$

fordi $x = 5$, $y = 1$ passer, og $x = 8$, $y = 3$ passer:

$$2 \cdot 5 - 3 \cdot 1 = 7 \text{ og } 2 \cdot 8 - 3 \cdot 3 = 7.$$

Vi har hermed ikke udtalt os om, hvorvidt der findes flere løsninger til ligningen end disse to.

De tal, der står foran x og/eller y i ligningen, kaldes ligningens *koefficienter*. Således er 2 og -3 koefficienter til x og y i ligningen $2x - 3y = 7$.