2. Matematiske modeller

En matematisk beskrivelse af et virkeligt fænomen kaldes en *matematisk model*. Beskrivelsen vil ofte bestå af en eller flere ligninger sammen med en forklaring på, hvad størrelserne i ligningen/ligningerne betegner. De lineære modeller, som vi arbejder med i kapitel 1 og 2, består altid af ligningen $y = ax + b$ samt en beskrivelse af, hvad x og y betegner. Senere skal vi se eksempler på mange andre typer matematiske modeller, herunder også matematiske modeller givet ved grafer og diagrammer.

Matematiske modeller bruges både videnskabeligt, kommercielt, politisk og i administrationen til at forstå, kontrollere og forudse virkelige fænomener. De er i dag blevet en uundværlig del af samfundet. Når teknologi udvikles, når bankerne investerer penge, når myndighederne træffer beslutninger om alt fra skatteprocent, SU, udledning af kvælstof, til nye veje og vaccinationsprogrammer, når DMI laver vejrudsigten, eller når firmaer annoncerer på internettet, og Google udvælger og ranglister søgeresultater, så er der i alle tilfælde matematiske modeller i spil.

iStockphoto.com/ALLVISIONN

DETTE KAPITEL

Vi ser eksempler på matematiske modeller og belyser nogle at de faldgruber, der er forbundet med at opstille en matematisk model på baggrund af lineær regression.

Vi undersøger, hvordan forklaringsgraden R^2 kan fortolkes, og vi illustrerer, hvordan forklaringsgraden ikke siger noget om kausalitet, dvs årssagssammenhæng.

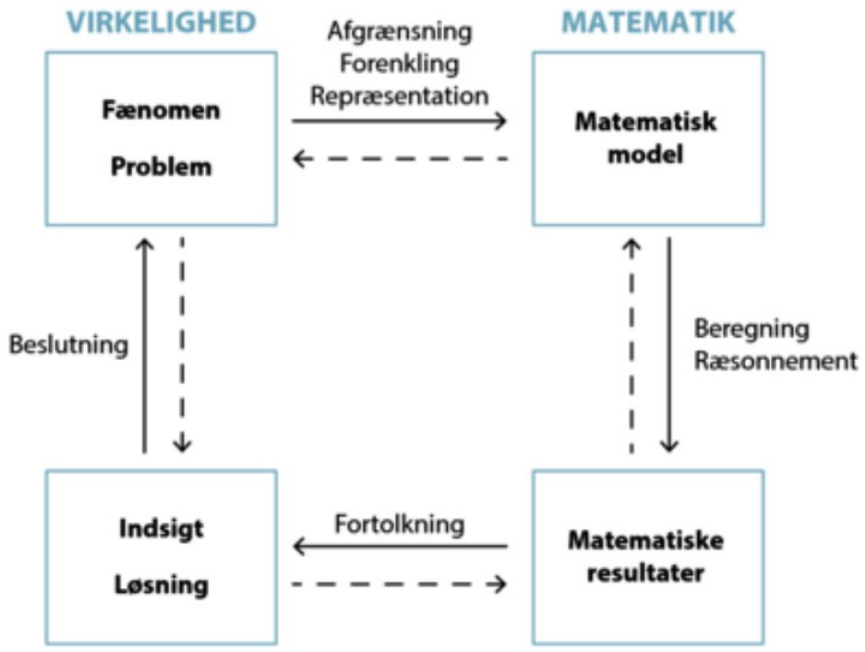
Med forklaringsgraden har man kun delvist muligheden for at vurdere kvaliteten af en lineær regression. I kapitlet viser vi, hvorfor det er nødvendigt også at inddrage en grafisk vurdering. *Punktplottet* og *residualplottet* er to muligheder, og vi viser i kapitlet, hvordan de fremstilles og benyttes.

Matematisk modellering

En brugbar matematisk model er resultatet af en proces, der kaldes *matematisk modellering,* og som involverer en række faser:

- Et virkeligt fænomen, der ønskes forklaret, afgrænses, forenkles og oversættes til matematik.
- En matematisk model opstilles og benyttes til beregninger og matematiske ræsonnementer.
- Matematiske resultater fortolkes til virkelige indsigter og løsninger.
- Nye indsigter og løsninger omsættes til beslutninger og handlinger.

Der er tale om en vekselvirkning mellem faserne, som også kan gå baglæns. F.eks. kan resultatet af en beregning give anledning til, at man går tilbage og afgrænser problemstillingen anderledes, eller justerer i den matematiske model.



Figur 1

EKSEMPEL 1

iStockphoto.com/sestovic

I kapitel 1 eksempel 8 opstillede vi vha. lineær regression den matematiske model

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

Den uafhængige variabel x er antal år efter 2007, og y er antal tilskadekomne eller dræbte i trafikken ved spiritusuheld. Vi kan konstatere, at modellen er en god matematisk beskrivelse af den virkelige problemstilling:

Færdselsuheld, hvor alkohol er involveret,

når vi begrænser os til perioden 2007-2015 og forenkler problemstillingen ved kun at se på spiritusuheld, hvor der hos politiet er registreret personskade eller dødsfald.

Af den lineære model aflæser vi hældningskoefficienten a til –102,2 og ved beregning finder vi for x = 9

$$y = -102{,}2 \cdot 9 + 1115 = 195{,}2.$$

De to resultater kan fortolkes således, at antallet af spiritusuheld i perioden 2007-2015 aftager lineært med ca. 102 uheld om året, samt at modellen forudser ca. 195 tilskadekomne eller dræbte ved spiritusuheld i 2016. På den baggrund kan myndighederne f.eks. vurdere, at faldet er tilfredsstillende, og at man derfor skal videreføre de initiativer mod spirituskørsel, der blev benyttet i perioden 2007-2015.

erfaringer om verden. Det er også grunden til, at der i diagrammet over matematisk modellering (figur 1) står *fortolkning* over den ene pil.

Vi skal i de følgende afsnit se på nogle af de problemer, der er forbundet med at benytte matematiske modeller, som er opstillet ved hjælp af lineær regression.

Omregning af data

Mange data fremkommer ved optælling eller måling og optræder derfor som absolutte tal:

antal cykeltyverier, antal arbejdsløse, overskud i kr., antal ton udledt CO_2*, antal gram udledt pesticid …*

Absolutte tal kan imidlertid, når de benyttes i en matematisk model, føre til resultater, der nemt kan misforstås og på den måde give svage (eller helt fejlagtige) konklusioner. I de situationer kan det være bedre at omregne data til *relative tal*, hvor tallene er sat i forhold til noget:

antal cykeltyverier pr. 1000 indbyggere, antal arbejdsløse i procent af den samlede arbejdsstyrke, overskud i kr. pr. medarbejder, antal ton udledt CO_2 *pr. dansker, antal gram pesticid pr.* m^3 *grundvand …*

EKSEMPEL 2

Vi betragter igen eksempel 8 i kapitel 1. En nærliggende konklusion på den lineære regression kunne være:

Bilisterne i Danmark bliver bedre og bedre til at lade bilen stå, når de har fået for meget at drikke.

Men det er en usikker konklusion, som ikke kan begrundes ud fra regressionen alene. Hvis f.eks. antallet af bilister i Danmark i samme periode er faldende, så kunne resultatet blive, at brøkdelen af bilisterne, der kører spirituskørsel er stigende. Det problem kan vi prøve at håndtere ved at sætte antallet af spiritusuheld i forhold til noget.

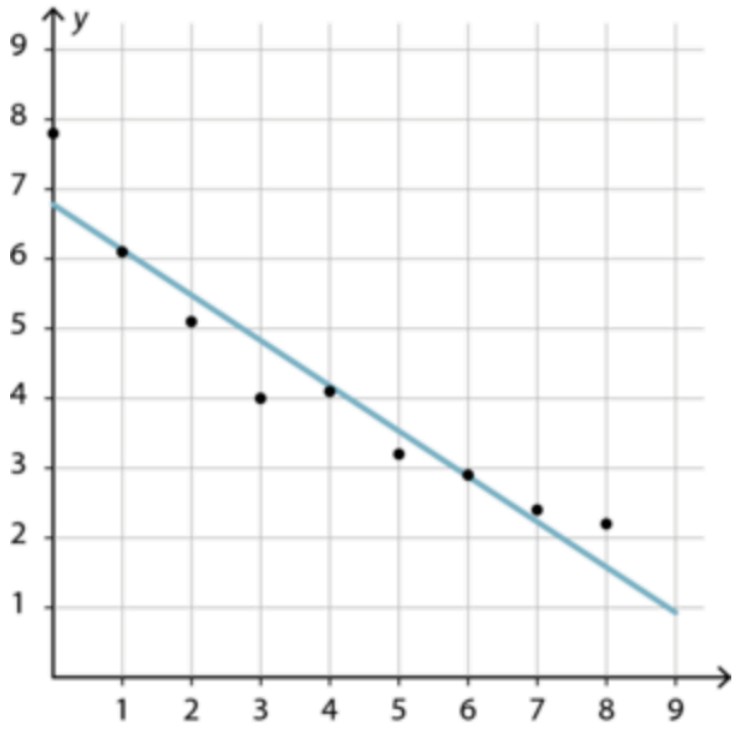
Tal fra Danmarks Statistik viser for perioden 2007-2015 udviklingen i antallet af familier med bil. Vi kan derfor lave en ny tabel, der, for hvert år (x) i samme periode, viser udviklingen i antal spiritusuheld pr. 10.000 familier med bil (y):

x	0	1	2	3	4	5	6	7	8
y	7,8	6,1	5,1	4,0	4,1	3,2	2,9	2,4	2,2

Med CAS-værktøjet fås

$$y = -0{,}65x + 6{,}78, \qquad R^2 = 0{,}91.$$

Den høje forklaringsgrad sammen med figur 2 fortæller, at også i det nye datasæt er der en tydelig lineært aftagende tendens.



Figur 2

Den nye regression underbygger bedre konklusionen ovenover, som vi derfor (måske) har fået større tiltro til.

Et egentligt bevis for konklusionen har vi ikke, og den mulighed, at konklusionen er forkert, er stadig til stede. Bliver der kørt mere og mere spirituskørsel, mens der samtidigt bliver registreret færre og færre uheld hos politiet? En yderligere afklaring kræver flere undersøgelser, f.eks. på skadestuerne, hvor skader i forbindelse med trafikuheld også bliver registreret.

Fortolkning af forklaringsgraden

Når lineær regression benyttes til modellering, er det væsentligt at forholde sig til kvaliteten af regressionen. Hvornår er kvaliteten af regressionen så god, at vi med rimelighed kan opstille en konklusion om lineær sammenhæng? At besvare det spørgsmål kræver også en fortolkning, f.eks. af forklaringsgraden R^2.

Forskellige videnskaber bruger og fortolker R^2 forskelligt. I fysik og kemi kræves almindeligvis en forklaringsgrad over 0,95 før en evt. konklusion om lineær sammenhæng. Inden for fag som biologi, samfundsfag og psykologi er praksis en anden. Her kan en forklaringsgrad på f.eks. 0,4 godt blive opfattet som et udtryk for en lineær tendens i datasættet.

Forskellen skyldes bl.a. de meget forskellige omstændigheder inden for de forskellige videnskaber som data bliver indsamlet under.

I fysik, kemi og til dels biologi fås data i forbindelse med kontrollerede forsøg, hvor der er godt styr på de faktorer, der har betydning for forsøget. Der er kun i meget lille udstrækning faktorer (fejlkilder), der "forstyrrer" målingerne. Derfor tillader man sig at kræve en meget tydelig lineær tendens i datasættet, før en evt. konklusion om lineær sammenhæng.

I f.eks. samfundsvidenskab, psykologi og tildels biologi er situationen ofte en anden. Her fremskaffes data mange gange under betingelser, hvor der ikke er god kontrol med alle de faktorer, der har betydning for det, der undersøges. Man er derfor mere tilbøjelig til at tale om lineær sammenhæng, selvom der er store variationer i data – variationerne kan jo skyldes de faktorer, der ikke er kontrol over.

Man støder af og til på følgende fortolkning af forklaringsgraden:

R^2 angiver den procentdel af datapunkternes variation, der kan *forklares* af modellen.

EKSEMPEL 3

I perioden 2009-2015 falder både de samlede lønudgifter x (mia. kr) og det samlede bogudlån y (mio. udlån) på de danske biblioteker, hvilket kan ses i følgende tabel med tal fra Danmarks Statistik:

Lønudgifter (mia.)	1,80	1,81	1,75	1,72	1,70	1,69	1,68
Bogudlån (mio.)	32,3	31	30,3	29,4	28,5	27,6	26,8

CAS-værktøjet giver følgende lineære model for sammenhængen mellem x og y:

$$y = 34{,}85x - 31{,}07, \qquad R^2 = 0{,}89.$$

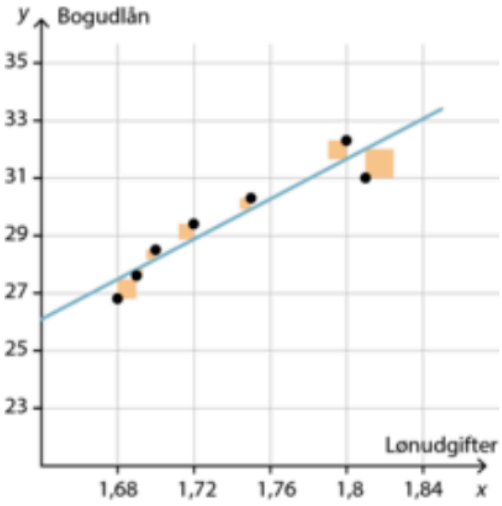
Som en fortolkning af forklaringsgraden siger vi, at den fundne lineære model forklarer 89 % af variationen i datasættet, mens den ikke forklarer de resterende 11 %.

For at forstå, hvad der egentlig menes med denne fortolkning, ser vi på formlen for forklaringsgraden, der blev introduceret i kapitel 1 i afsnittet *Lineær regression*:

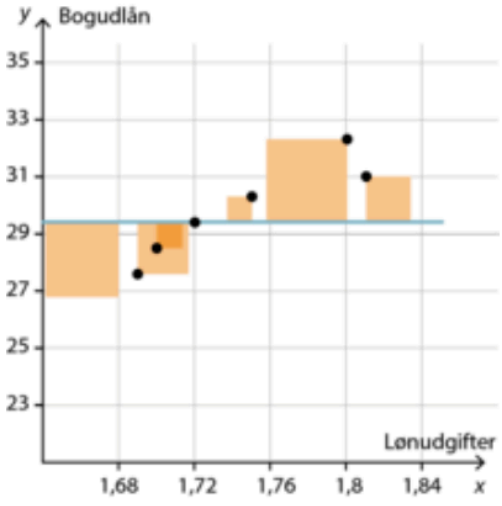
$$R^2 = \frac{A_g - A_r}{A_g}.$$

A_r er det samlede areal af de farvede kvadrater på figur 3. Det er et mål for den *restvariation* omkring regressionslinjen, der er i datasættet. Restvariationen bliver også (mere løst) omtalt som den variation i datasættet, den lineære model ikke kan *forklare.*

A_g er det samlede areal af de kvadrater på figur 4. Den er et mål for den totale variation i datasættet.



Figur 3



Figur 4

På den baggrund opfattes brøken $\frac{A_r}{A_g}$ som den del af den totale variation i datasættet, som modellen ikke kan forklare.

I dette eksempel finder man $A_r = 2{,}468$ og $A_g = 22{,}589$. Dvs.

$$\frac{A_r}{A_g} = \frac{2{,}468}{22{,}589} = 0{,}11.$$

Den lineære model for sammenhængen mellem bibliotekernes lønudgifter og samlede bogudlån kan altså ikke forklare 11 % af variationen i datasættet i tabellen ovenover.

Da

$$R^2 = \frac{A_g - A_r}{A_g} = 1 - \frac{A_r}{A_g} = 1 - 0{,}11 = 0{,}89\ ,$$

kan man i stedet sige, at forklaringsgraden angiver, at 89 % af variationen i datasættet forklares med modellen.

Korrelation og kausalitet

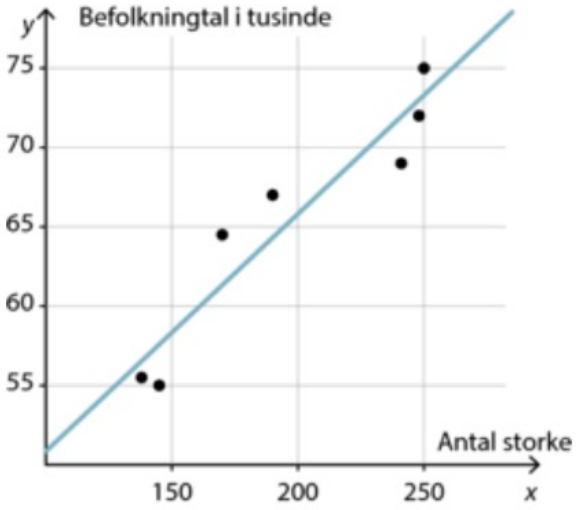
Når forklaringsgraden R^2 er stor, siger man også, at der er en høj *korrelation* mellem de to variable *x* og *y*. Det betyder, at der er en stor overenstemmelse mellem den måde *x* og *y* varierer på, der gør det muligt at forudsige den ene variabel, når man kender den anden.

En høj korrelation mellem *x* og *y* er ikke det samme, som at den ene variabel er årsag til den anden – dvs. det fænomen, den ene variabel beskriver, forårsager det fænomen, den anden variabel beskriver. Et klassisk eksempel illustrerer den pointe:

EKSEMPEL 4

Flere undersøgelser har påvist en stor overensstemmelse mellem antallet af storke og befolkningstal; når antallet af storke vokser, så vokser også befolkningstallet.

Blandt andet tal fra den tyske by Oldenburg opsamlet i perioden 1930-1936 viser en tydelig sammenhæng.



Figur 5

Mellem de to variable *antal storke* og *befolkningstal i tusinde* var der altså i Oldenburg en høj korrelation. Men der er oplagt ikke tale om en årsagssammenhæng.

En årsagssammenhæng kaldes også for en *kausal sammenhæng* (latin: causa: årsag), og man har som fast vending:

Korrelation medfører ikke nødvendigvis kausalitet.

Fra tid til anden finder den type misforståelser også vej til medierne, og ofte under opsigtsvækkende overskrifter:

- *Unge, der ryger, får dårligere karakterer*
- *Elever, der er gode til musik, klarer sig godt i skolen*
- *Personer, der ser meget fjernsyn, er usunde*
- *Børn, der kommer for sent i seng, bliver overvægtige*
- *Øget grænsekontrol reducerer antallet af asylansøgere*
- *Mere politi skaber mere vold.*

I mange af tilfældene er det tvivlsomt, om der tale om en kausal sammenhæng, selvom korrelationen kan vise sig at være høj.

Når der er en høj korrelation, men ingen kausal sammenhæng, kalder man det også en *spuriøs sammenhæng*. En spuriøs sammenhæng mellem x og y kan skyldes et af to forhold:

1. Overensstemmelsen mellem x og y er en ren tilfældighed
2. Overensstemmelsen mellem x og y skyldes en anden bagvedliggende variabel z, der er årsag til både x og y.

Den spuriøse sammenhæng mellem antallet af storke og befolkningstal har man f.eks. forsøgt at forklare med den bagvedliggende variabel *urbanisering*: I byer får familier i gennemsnit færre børn, og i byer lever der kun få storke.

EKSEMPEL 5

I eksempel 3 beskrev vi sammenhængen mellem lønudgifter x og antal bogudlån y på de danske biblioteker med en lineær model, hvor forklaringsgraden er 0,89. Selvom der er en høj korrelation mellem x og y, er det slet ikke sikkert, at x og y er kausalt forbundne.

Danske biblioteker har i en årrække oplevet store forandringer, hvor faktorer som digitalisering, nye læsevaner, konkurrence fra andre kulturinstitutioner og rationalisering har haft en betydning.

Sammenhængen mellem x og y er derfor måske en spuriøs sammenhæng med ovenstående faktorer blandt de (mange) bagvedliggende faktorer, der både er årsag til x og til y.

EKSEMPEL 6

Målinger viser, at de franskstuderende på Københavns Universitet har lavere skonummer end de ingeniørstuderende på DTU (Danmarks Tekniske Universitet). Får man derfor store fødder ved at læse til ingeniør? Eller små fødder af at studere fransk?

I det tilfælde, hvor undersøgelser viser, at der er en kausal sammenhæng mellem x og y, skal man også overveje, om det er x, der er årsag til y, eller y, der er årsag til x. Hvis x tidsligt går forud for y, er det x, der er årsag til y. Men i de situationer, hvor der ikke er nogen naturlig tidslig rækkefølge for x og y, kan det være svært at afgøre.

EKSEMPEL 7

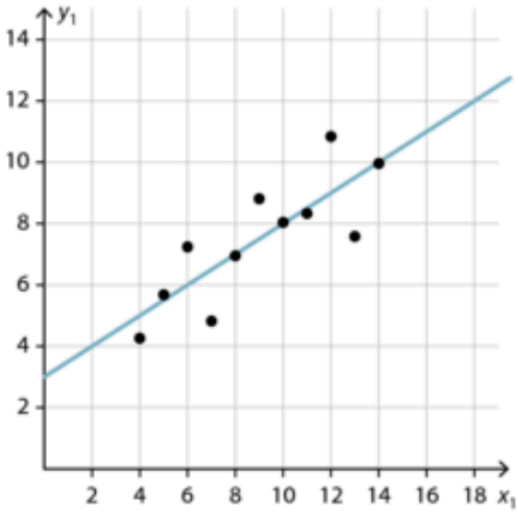
Undersøgelser peger på, at *politisk ståsted* og *valg af avis* er kausalt forbundne, men det er uklart, hvilken vej kausaliteten vender: Er det mit politiske ståsted, der bestemmer, hvilken avis jeg foretrækker, eller er det den avis jeg læser, der former mit politiske ståsted?

Punktplot og residualplot

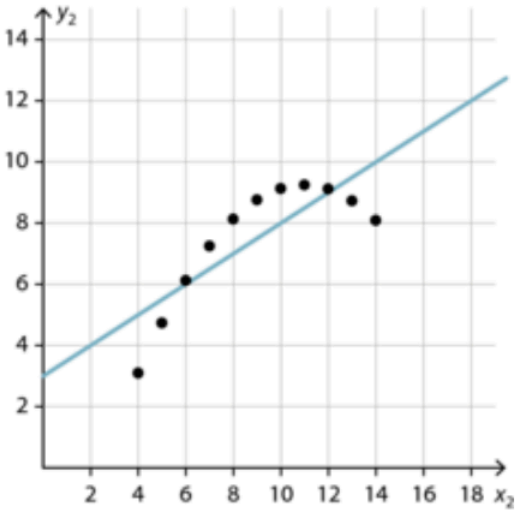
Figurerne nedenunder viser regressioner på fire forskellige datasæt. Det interessante er her, at man i alle fire tilfælde (med tre decimalers nøjagtighed) får den samme regressionsligning

$$y = 0{,}500x + 3{,}000$$

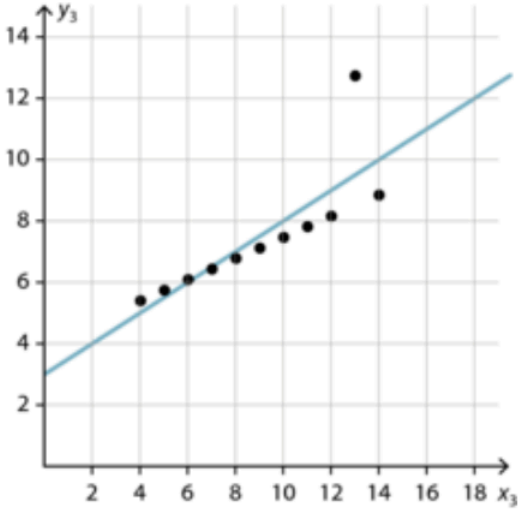
og (med to decimalers nøjagtighed) den samme forklaringsgrad $R^2 = 0{,}67$. Der er med andre ord tale om fire meget forskellige datasæt, der beskrives lige godt med den samme rette linje.



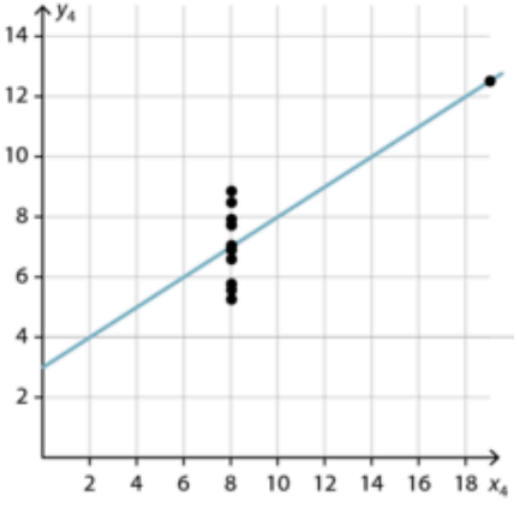
Figur 5



Figur 6



Figur 7



Figur 8

Eksemplet (første gang præsenteret af statistikeren Francis Anscombe i 1973), viser meget tydeligt, at man skal passe på med at slutte noget alene ud fra forklaringsgraden.

Mens det kan være rimeligt at beskrive det første datasæt ved den lineære model $y = 0{,}500x + 3{,}000$, så virker det som en meget dårlig ide i de tre andre tilfælde. Vi bemærker også, at data i det andet tilfælde meget bedre kan beskrives med en ikke-lineær model.

Kvaliteten af en lineær regression kan altså ikke alene vurderes ud fra R^2. Som supplement laver man også grafiske vurderinger.

Punktplot

Datasættet afsat som punkter i et koordinatsystem kaldes et *punktplot*.

Når vi også indtegner regressionslinjen, kan vi benytte, at:

> En god lineær regression er kendetegnet ved et punktplot, hvor datapunkterne har *små* og *tilfældige* variationer omkring regressionslinjen.

Blandt de fire punktplot på figur 5-8 ser vi kun på den første figur en lille og tilfældig variation i datapunkternes placering omkring regressionslinjen. Det viser, at kun i det første tilfælde er det rimeligt at tale om en lineær tendens.

Residualplot

I et residualplot afbilder man datasættets x-værdier sammen med datapunkternes afvigelse fra regressionslinjen. Teknikken illustreres med et eksempel. Vi ser på datasættet vist på figur 5:

x	4	5	6	7	8	9	10	11	12	13	14
y_{data}	4,26	5,68	7,24	4,82	6,95	8,81	8,04	8,33	10,84	7,58	9,96
y_{model}	5,0	5,5	6,0	6,5	7,0	7,5	8,0	8,5	9,0	9,5	10,0
residual	–0,74	0,18	1,24	–1,68	–0,05	1,31	0,04	–0,17	1,84	–1,92	–0,04

Tabellens to første rækker indeholder selve datasættet. Med betegnelsen y_{data} henviser vi til datasættets y-værdier. Regressionsligningen

$$y = 0{,}500x + 3{,}000$$

bruger vi til at beregne modellens y-værdier (y_{model}). F.eks. får vi til $x = 12$

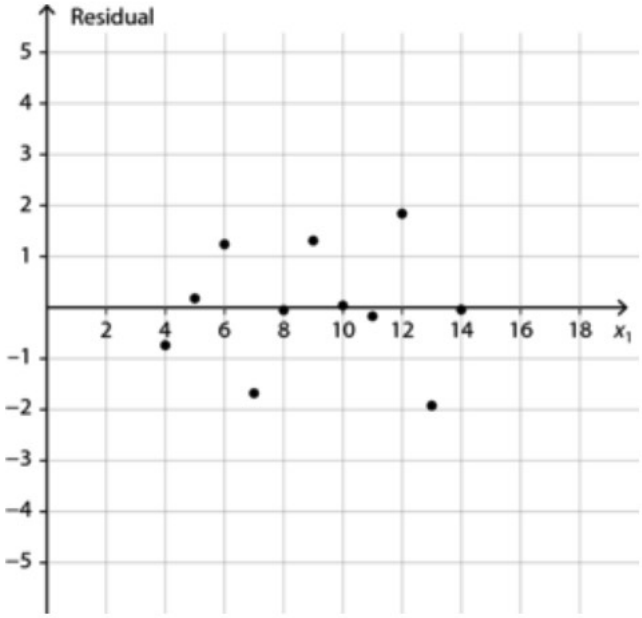
$$y_{\text{model}} = 0{,}500 \cdot 12 + 3{,}000 = 9{,}0.$$

Tabellens tredje række viser de tilsvarende resultater for de andre datapunkter. Forskellen

$$y_{\text{data}} - y_{\text{model}} = 10{,}84 - 9{,}0 = 1{,}84$$

kaldes *residualet* (resten) og er den lodrette afstand mellem datapunktet (12; 10,84) og regressionslinjen. Residualer regnes med fortegn og er negative, når datapunkterne er under regressionslinjen. I tabellens fjerde række er alle 11 residualer vist.

Residualplottet kan nu laves ved at afbilde datasættets x-værdier sammen med de tilhørende residualer (se figur 9).



Figur 9

Fordi residualerne angiver størrelserne på afvigelserne, har vi:

En god lineær regression er kendetegnet ved et residualplot, hvor punkterne har små og tilfældige variationer omkring x-aksen.

EKSEMPEL 8

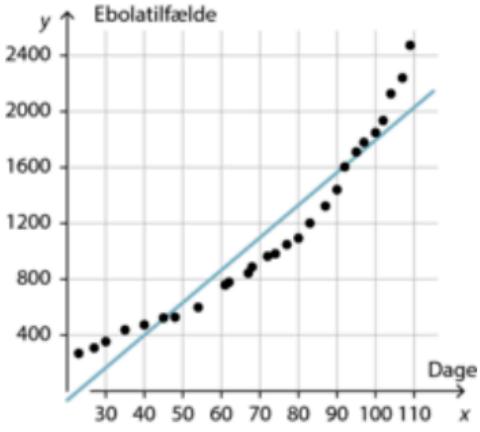
Vi undersøger data fra verdenssundhedsorganisationen WHO over udviklingen i antallet af registrerede ebolatilfælde i Vestafrika (y) i perioden 23 maj 2014 til 18 august 2014. Som uafhængig variabel x vælger vi antal dage efter 30 april 2014.

Pascal Guyot/AFP/Scanpix

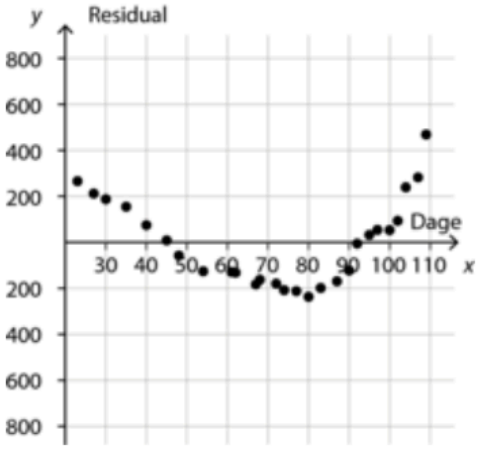
Lineær regression i et CAS-værktøj giver den lineære model

$$y = 23{,}3x - 530{,}9 \qquad R^2 = 0{,}91.$$

Selv om forklaringsgraden er høj, er der en klar ikke-lineær tendens i datasættet. Det er tydeligt, både når vi ser på punktplottet (figur 10) og når vi ser på residualplottet (figur 11). Residualplottet viser, at residualerne ikke varierer tilfældigt om x-aksen, men ændrer sig på systematisk vis over hele perioden.



Figur 10



Figur 11