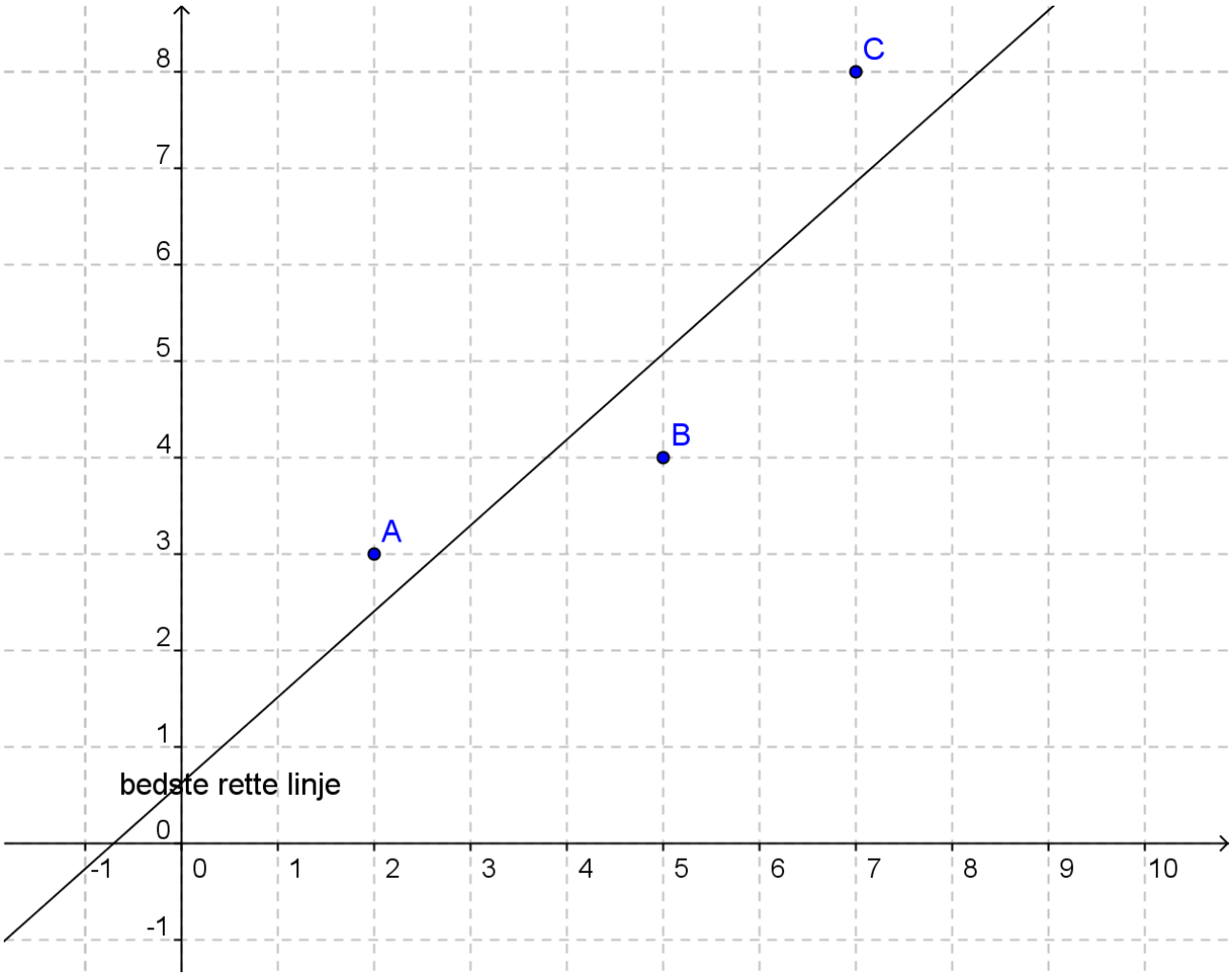
Bedste rette linje ved mindste kvadraters metode

- fra www.borgeleo.dk



Figur 1: Tre datapunkter og den bedste rette linje bestemt af A, B og C

Målepunkter og bedste rette linje

I ovenstående koordinatsystem er indtegnet tre målepunkter sammen med en ret linje, som skal placeres bedst muligt - sådan at målepunkterne ligger så tæt på linjen som muligt.

Punkterne har koordinaterne

$A(x_1, y_1)$ $B(x_2, y_2)$ $C(x_3, y_3)$

Hvordan beregner vi så ligningen for den bedste rette linje (modelfunktionen)?

Formlen er

(1) $y = ax + b$

hvor konstanterne *a* og *b* er givet ved formlerne

(2) $a = \frac{n*S_{xy} - S_x*S_y}{n*S_{xx} - S_x^2}$

og

(3) $b = \frac{S_y*S_{xx} - S_x*S_{xy}}{n*S_{xx} - S_x^2}$

Her har vi indført betegnelserne

$n = antal\ datapunkter$ (som er 3 på figur 1 – men der kunne være flere)

$S_x = \sum x_i$ = summen af *x*-værdierne for alle datapunkterne
$S_y = \sum y_i$ = summen af *y*-værdierne for alle datapunkterne
$S_{xy} = \sum x_i \cdot y_i$ = summen af produkterne af *x* og *y* for hvert datapunkt
$S_{xx} = \sum x_i^2$ = summen af kvadraterne af *x*-værdierne
$S_{yy} = \sum y_i^2$ = summen af kvadraterne af *y*-værdierne

Vær opmærksom på, at disse summer nemt kan beregnes, når datapunkterne er kendt!
Når *a* og *b* er beregnet, kan den rette linje med ligningen (1) tegnes sammen med datapunkterne.

Eksempel:

Lad os se på hvordan vi beregner *a* og *b* i et konkret tilfælde:

$A(x_1, y_1) = (2,3)$ $B(x_2, y_2) = (5,4)$ $C(x_3, y_3) = (7,8)$

Vi beregner de 4 forskellige summer:

$$S_{xx} = \sum x_i^2 = 2^2 + 5^2 + 7^2 = 78$$

$$S_x = \sum x_i = 2 + 5 + 7 = 14$$

$$S_y = \sum y_i = 3 + 4 + 8 = 15$$

$$S_{xy} = \sum x_i \cdot y_i = 2 \cdot 3 + 5 \cdot 4 + 7 \cdot 8 = 82$$

Og indsætter tallene i formlerne for *a* og *b*:

$$a = \frac{3 \cdot 82 - 14 \cdot 15}{3 \cdot 78 - 14^2} = \frac{36}{38} = 0{,}947368$$

$$b = \frac{78 \cdot 15 - 14 \cdot 82}{3 \cdot 78 - 14^2} = \frac{22}{38} = 0{,}578947$$

Altså er ligningen for den bedste rette linje:

$$\boldsymbol{y = ax + b = 0{,}947368 \cdot x + 0{,}578947}$$

Dette er ligningen for den bedste rette linje bestemt af datapunkterne A, B og C

Øvelse 1:

Prøv med et værktøj som Excel eller Tinspire at lave lineær regression på de tre datapunkter A, B og C fra eksemplet ovenfor for at se, om du får samme løsning for den bedste rette linjes ligning som ovenfor.

Bevis for formlerne for *a* og *b* for den bedste rette linje

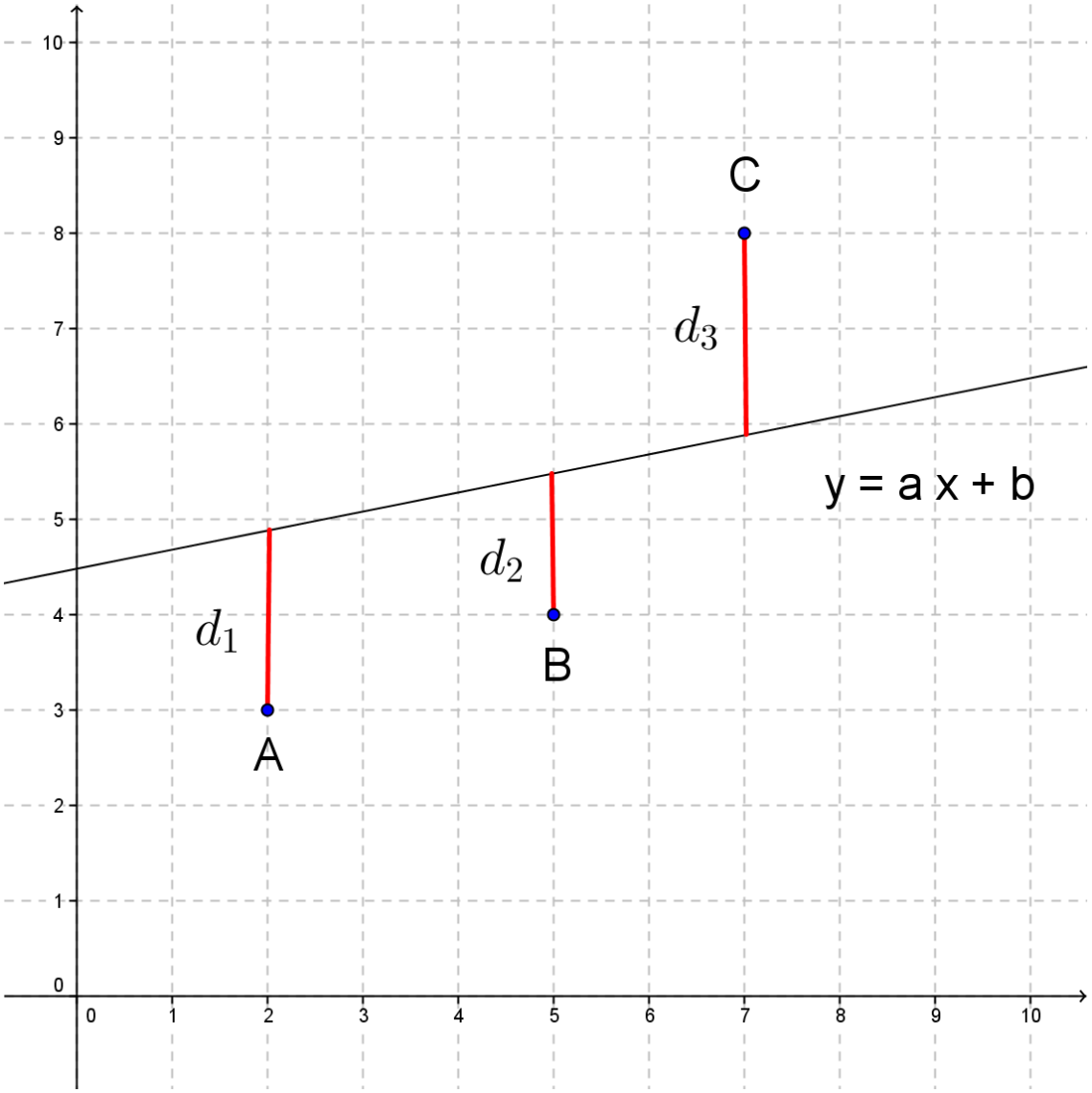
Men hvordan begrunder vi nu formlerne (2) og (3)?

Vi vil her nøjes med at se på et eksempel med 3 datapunkter for at beregningerne kan blive mere overskuelige. Det er nemt senere at vende tilbage til de generelle formler.

Først en figur, hvor vi kan se afvigelserne d_1, d_2 og d_3 mellem en ret linje med ligningen $y = a \cdot x + b$ og de 3 datapunkter A, B og C. Se figuren nedenfor (figur 2).

De tre datapunkter er, som nævnt tidligere:

$A(x_1, y_1) = (2,3)$ $\quad$ $B(x_2, y_2) = (5,4)$ $\quad$ $C(x_3, y_3) = (7,8)$



Figur 2: de tre datapunkter A, B og C og den rette linje med ligningen y = a x + b

Her skal vi opstille et udtryk for summen af de *kvadratiske* afvigelser mellem linjens *y*-værdier i de *x*-værdier, der stammer fra målepunkterne - og så datapunkternes *y*-værdier:

(4) $S(a,b) = (d_1)^2 + (d_2)^2 + (d_3)^2$

Afvigelserne d_1, d_2 og d_3 afhænger af linjens placering og dermed tallene *a* og *b*. For hver *x*-værdi for datapunkterne kan vi udregne den tilhørende *y*-værdi for den rette linje, fx for det første datapunkt: $y_{linje} = a \cdot x_1 + b$.

Datapunktets *y*-værdi er y_1, og dermed bliver afvigelsen $d_1 = a \cdot x_1 + b - y_1$

Øvelse 2:

Indtegn *y*-værdien $y_{linje} = a \cdot x_1 + b$ og datapunktets *y*-værdi y_1 på *y*-aksen på figur 2 ovenfor.

For de to øvrige datapunkter kan vi på samme måde lave formler for afvigelserne, og herved bliver summen af de kvadratiske afvigelser:

(5) $$S(a,b) = (a \cdot x_1 + b - y_1)^2 + (a \cdot x_2 + b - y_2)^2 + (a \cdot x_3 + b - y_3)^2$$

summen af de kvadratiske afvigelser mellem modelfunktion y = ax+b og datapunkters y-værdier.

I dette udtryk indgår alle datapunkterne, men også de to variable *a* og *b* - hvor *a* er hældningskoefficienten for den rette linje, og *b* er denne linjes skæring med *y*-aksen.

Opgaven:

Vi skal i udtrykket S(a,b) vælge de værdier af a og b, der gør summen mindst mulig, sådan at den rette linje ligger så tæt på datapunkterne som mulig (mindst mulig afvigelse). Og vise, at det netop er de to værdier fra formlerne (2) og (3).

I udtrykket (5) skal vi kunne udregne en parentes med 3 led opløftet til anden potens! Og så oven i købet 3 af dem.

Det gør vi i to step:

Øvelse 3

Vis, at følgende ligning er rigtig – fx ved at gange to ens parenteser sammen, hvor hver parentes indeholder $a \cdot x + b - y$:

$$(a \cdot x + b - y)^2 = a^2 \cdot x^2 + b^2 + y^2 + 2a \cdot x \cdot b - 2a \cdot x \cdot y - 2b \cdot y$$

Denne udregning skal vi bruge på hvert led i ligning (5).

Resultatet er:

$$\begin{aligned} S(a,b) = \quad & a^2 \cdot {x_1}^2 + b^2 + {y_1}^2 + 2a \cdot x_1 \cdot b - 2a \cdot x_1 \cdot y_1 - 2b \cdot y_1 \\ + \ & a^2 \cdot {x_2}^2 + b^2 + {y_2}^2 + 2a \cdot x_2 \cdot b - 2a \cdot x_2 \cdot y_2 - 2b \cdot y_2 \\ + \ & a^2 \cdot {x_3}^2 + b^2 + {y_3}^2 + 2a \cdot x_3 \cdot b - 2a \cdot x_3 \cdot y_3 - 2b \cdot y_3 \end{aligned}$$

Dette kan reduceres til

(5a) $(a,b) = a^2 \cdot ({x_1}^2 + {x_2}^2 + {x_3}^2) + 3b^2 + ({y_1}^2 + {y_2}^2 + {y_3}^2) + 2a \cdot (x_1 + x_2 + x_3) \cdot b$

$$-2a \cdot (x_1 \cdot y_1 + x_2 \cdot y_2 + x_3 \cdot y_3) - 2b \cdot (y_1 + y_2 + y_3)$$

Øvelse 4

Prøv om du selv kan lave denne omformning!

I udtrykket (5a) optræder der i parenteserne nogle summer, der kan beregnes ud fra datapunkterne.

Vi indfører nogle kortere betegnelser for disse summer (*S* for sum!) og udregner dem, idet vi husker, at $A(x_1, y_1) = (2,3)$, $B(x_2, y_2) = (5,4)$ og $C(x_3, y_3) = (7,8)$:

$S_{xx} = {x_1}^2 + {x_2}^2 + {x_3}^2 = 2^2 + 5^2 + 7^2 = 78$

$S_{yy} = {y_1}^2 + {y_2}^2 + {y_3}^2 = 3^2 + 4^2 + 8^2 = 89$

$S_x = x_1 + x_2 + x_3 = 2 + 5 + 7 = 14$

$S_{xy} = x_1 \cdot y_1 + x_2 \cdot y_2 + x_3 \cdot y_3 = 2 \cdot 3 + 5 \cdot 4 + 7 \cdot 8 = 82$

$S_y = y_1 + y_2 + y_3 = 3 + 4 + 8 = 15$

$n = antal\ datapunkter = 3$

(vi lavede disse udregninger allerede på side 2!)

Indfører vi disse tal i det forrige udtryk (5a) for $S(a,b)$, får vi:

(6) $S(a,b) = 78 \cdot a^2 + 3 \cdot b^2 + 89 + 2 \cdot 14 \cdot a \cdot b - 2 \cdot 82 \cdot a - 2 \cdot 15 \cdot b$

Øvelse 5

Tjek lige, at det er rigtigt!

Vi ser så først på hvordan denne funktion $S(a,b)$ afhænger af *a* (*b* betragtes her som et fast tal!)
En lille omskrivning giver:

(6a) $S(a) = 78 \cdot a^2 + 2 \cdot (14 \cdot b - 82) \cdot a + (3 \cdot b^2 + 89 - 2 \cdot 15 \cdot b)$

Øvelse 6

Kontroller, at omformningen er rigtig!

Men (6a) er jo et 2. gradsudtryk i den variable *a*! (når *b* er et fast tal)

Derfor kan vi beregne toppunktets *a*-værdi:

$$a = -\frac{2 \cdot (14 \cdot b - 82)}{2 \cdot 78} = \frac{82 - 14 \cdot b}{78}$$

Øvelse 7

Vis, at dette er rigtigt – du skal bruge formlen for toppunktets 1. koordinat for en parabel! Se evt. hjælpen i slutningen af dette dokument.

Dette kan omskrives til ligningen

(7) $78 \cdot a + 14 \cdot b = 82$

Øvelse 8

Vis, at omformningen er rigtig!

Men *b* skal jo også vælges, så udtrykket $S(a, b)$ bliver mindst mulig! Derfor omskriver vi igen udtrykket til det følgende, hvor vi nu ser på *a* som en konstant:

(6b) $S(b) = 3 \cdot b^2 + 2 \cdot (14 \cdot a - 15) \cdot b + (78 \cdot a^2 - a \cdot 2 \cdot 82 + 89)$

Øvelse 9

Vis igen, at dette er en rigtig omformning af formel (6)!

Udtrykket er jo nu et 2. grads udtryk i den variable *b*! (når *a* er et fast tal)

Vi beregner så *b*-værdien for toppunktet:

$$b = -\frac{2 \cdot (14 \cdot a - 15)}{2 \cdot 3} = \frac{15 - 14 \cdot a}{3}$$

Denne ligning kan omskrives til:

(8) $14 \cdot a + 3 \cdot b = 15$

Vi samler nu de to ligninger, der sikrer os den mindste værdi for summen af de kvadratiske afvigelser:

(9) $\begin{matrix} 78 \cdot a + 14 \cdot b = 82 \\ 14 \cdot a + 3 \cdot b = 15 \end{matrix}$

Løsning med lige store koefficienters metode

For at løse de to ligninger (9) med to ubekendte bruger vi *lige store koefficienters metode*.

Først skal vi finde talværdien for den variable *a*:

Vi ganger nu den første ligning i (9) igennem med 3 og den anden med 14:

$$3 \cdot 78 \cdot a + 14 \cdot 3 \cdot b = 3 \cdot 82$$

$$14^2 \cdot a + 14 \cdot 3 \cdot b = 14 \cdot 15$$

Og vi trækker nu den nederste ligning fra den øverste (*b*-leddene forsvinder fordi der er lige mange *b*-er i de to ligninger):

$$3 \cdot 78 \cdot a - 14^2 \cdot a = 3 \cdot 82 - 14 \cdot 15$$

Så sætter vi *a* uden for en parentes:

$$a \cdot (3 \cdot 78 - 14^2) = 3 \cdot 82 - 14 \cdot 15$$

Og endelig deles med parentesen på begge sider:

(10) $$a = \frac{3 \cdot 82 - 14 \cdot 15}{3 \cdot 78 - 14^2} = \frac{36}{38} = 0{,}947368$$

For at finde *b* gør vi noget tilsvarende: vi ganger den første ligning i (9) igennem med 14 og den anden med 78 – men det må du selv gøre for at kunne lave den næste øvelse!

Øvelse 10

Løs de to ligninger (9) for at finde den ubekendte *b* ! Vis, at

(11) $$b = \frac{78 \cdot 15 - 14 \cdot 82}{78 \cdot 3 - 14^2} = \frac{22}{38} = 0{,}578947$$

Øvelse 11

Beregn summen af de kvadratiske afvigelser fra formel (5) når du har *a* og *b* som i formlerne (10) og (11). Denne værdi af summen er den mindst mulige! Du kan prøve vælge andre værdier for *a* og *b* – så vil du opdage, at summen af de kvadratiske afvigelser altid er større med de samme datapunkter!

Øvelse 12

For at få de generelle formler for *a* og *b*, skal du genindføre summerne S_x, S_y, S_{xy} og S_{xx} fra side 6 i dette dokument i formlerne (10) og (11). Vis, at du herved vender tilbage til de generelle formler (2) og (3) fra side 2.

Øvelse 13 (sværere!)

Prøv at gennemføre beviset for formlerne for *a* og *b* ved at erstatte summerne i parenteserne i formel (5a) med symbolerne S_x, S_y, S_{xy} og S_{xx} fra side 6 – altså uden at sætte tal ind. Prøv så derfra at gennemføre beviset for formlerne for *a* og *b*, sådan at du viser de generelle formler (2) og (3) direkte!

Øvelse 14

Bestem ligningen for den bedste rette linje, når datapunkterne er

$A(x_1, y_1) = (-2{,}11)$, $B(x_2, y_2) = (3{,}9)$, $C(x_3, y_3) = (9{,}4)$ og $D(x_4, y_4) = (17{,}1)$

Indtegn datapunkter og den bedste rette linje i samme koordinatsystem.

Sådan finder man 1. koordinat for parabels toppunkt

Parablen med ligningen

(12) $$y = \alpha \cdot x^2 + \beta \cdot x + \gamma$$

har et toppunkt med x-koordinaten

(13) $$x = -\frac{\beta}{2\alpha}$$