

LINEÆR REGRESSION

Hvordan bestemmer man ligningen for tendenslinjen?

Version 6

InterMat april 2019.

Tendenslinjen for et datasæt er løst sagt defineret som den linje der "bedst" approksimerer alle punkterne i datasættet. Hermed menes at punkterne samlet set skal ligge så tæt som muligt på linjen.

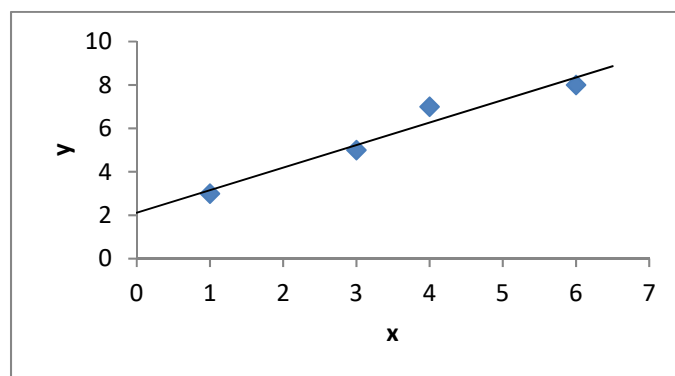
For at sætte tal på afstanden fra et bestemt datapunkt til linjen udregnes det såkaldte *fejlkvadrat*, nemlig kvadratet på den lodrette afstand fra datapunkt til linjen. Punkternes samlede afstand til linjen udtrykkes så ved summen af alle fejlkvadraterne.

Vi laver et **eksempel**:

Datasæt:

x	y
1	3
3	5
4	7
6	8

Tendenslinje iflg. Excel:



Tendenslinjens ligning er $y = ax + b$. Vores opgave: at bestemme a og b , sådan at summen af fejlkvadraterne er minimal. Udfyld tabellen:

(x, y)	Fejlkvadrat $(y - (ax + b))^2$	Udregnet
(1,3)	$(3 - (a \cdot 1 + b))^2$	$a^2 + b^2 - 6a - 6b + 2ab + 9$
(3,5)	$(5 - (a \cdot 3 + b))^2$	
(4,7)		
(6,8)		

Summen af fejlkvadraterne – som afhænger af hvordan vi vælger a og b – kalder vi nu $f(a, b)$. Denne funktion af to variable er et mål for hvor langt punkterne ligger fra tendenslinjen. Jeg har lagt udtrykkene i tabellen ovenfor sammen – check lige at jeg har regnet rigtigt (i hånden eller i nSpire):

$$f(a, b) = 62a^2 + 4b^2 - 188a - 46b + 28ab + 147$$

For at finde den bedste værdi af a og b leder vi efter et minimum for funktionen f . For på den måde er summen af fejlkvadraterne jo blevet så lille som muligt; vi har fundet den rette linje, der bedst muligt tilnærmer data.

Fortsætter →

0) Start med at tegne $z(x, y) = f(x, y)$ i nSpire's 3D graf plot. For at få et pænt plot kan du højre-klikke og vælge område/zoom $x_{\min} = 0.5$, $x_{\max} = 1.5$, $y_{\min} = 0$, $y_{\max} = 4$, $z_{\min} = 0$, $z_{\max} = 2$. Vælg derefter igen område/zoom og vælg "højde/bredde" og sæt forholdene til $x: y: z$ til 6: 3: 2. Kan du se at $f(x, y)$ har et minimum? Kan du også se hvad (x, y) -koordinaterne til det lokale minimum er, sådan cirka?

1) Nu vil vi bestemme minimumspunktet vha. differentiation. Udregn (i hånden!) $f'_a(a, b)$ og $f'_b(a, b)$

2) Find de stationære punkter ved at løse ligningssystemet $f'_a(a, b) = 0$ og $f'_b(a, b) = 0$ (enten i hånden eller på nSpire). Du skal få $a = \frac{27}{26}$ og $b = \frac{55}{26}$. Check også at det passer med det punkt, du så i opgave 0).

3) Lad til slut nSpire eller Excel finde tendenslinjens ligning – og sammenlign så hældningskoefficient og akseskæring med hhv. $\frac{27}{26}$ og $\frac{55}{26}$ ☺ Passer det?

Konklusion: $a = \frac{27}{26}$ er altså den hældningskoefficient og $b = 55/26$ den y -skæring, der gør fejlkvadratet $f(a, b)$ mindst mulig. Og dermed er den fundne linje den bedste linje til at approksimere alle datapunkterne.

De næste spørgsmål er ikke obligatoriske!

4) Hvis vi ikke allerede havde set grafen, ved vi egentlig bare, at der i punktet $(a, b) = (\frac{27}{26}, \frac{55}{26})$ er enten et minimumspunkt eller et maksimumspunkt eller et saddelpunkt for fejlfunktionen $f(a, b)$. For at vise at det er et minimum som ønsket, får vi brug for Hesse-matricen

$$H = \begin{bmatrix} f''_{aa}(a, b) & f''_{ab}(a, b) \\ f''_{ba}(a, b) & f''_{bb}(a, b) \end{bmatrix}$$

Udregn (i hånden!) de fire dobbeltafledede $f''_{aa}(a, b)$, $f''_{bb}(a, b)$, $f''_{ab}(a, b)$, $f''_{ba}(a, b)$ og opskriv herefter Hesse-matricen udregnet i punktet $(a, b) = (\frac{27}{26}, \frac{55}{26})$.

Beregn nu determinanten af H . Passer fortegnet med et ekstremum – eller med et saddelpunkt?

5) For at se at ekstremumspunktet er et minimum, skal du finde egenverdierne. Find dem med nSpire's automatiske funktion eigvl(H) eller (hardcore) opskriv det karakteristiske polynomium

$$P(k) = \det(H - kE) = \det \begin{bmatrix} f''_{aa}(a, b) - k & f''_{ab}(a, b) \\ f''_{ba}(a, b) & f''_{bb}(a, b) - k \end{bmatrix}$$

og løs ligningen $P(k) = 0$ til at finde de to egenverdier for H . Du skal gerne få $k_1 = 1,6$ og $k_2 = 130,4$

6) Da begge egenverdierne er positive, så har funktionen f som ønsket et (lokalt) minimum i punktet $(a, b) = (\frac{27}{26}, \frac{55}{26})$.