De mindste kvadraters metode

JENS CARSTENSEN, Frederiksberg

I LMFK–Bladet nr. 1 og 2 er lineær regression behandlet. Forfatterne gjorde sig overvejelser om, hvordan ligningen for den bedste rette linje, der tilnærmer n datapunkter i koordinatsystemet, kan bestemmes ved hjælp af de mindste kvadraters metode.

For læserne virker det måske noget tungt, men matematikken bag fastlæggelsen af den bedste rette linje kræver i virkeligheden intet andet end kendskab til formlen for andengradspolynomiets toppunkt. Teorien for ekstremumspunkter for funktioner af to variable er ganske overflødig.

Vi viser her hvordan sagen kan gribes an. Det er ikke nogen væsentlig indskrænkning at gennemføre regningerne med 3 punkter i stedet for med n punkter. Det letter skrivearbejdet og øger gennemskueligheden. Antag derfor, at vi har forelagt tre punkter

$$(x_1, y_1)\ ,\ (x_2, y_2)\ ,\ (x_3, y_3)$$

i koordinatsystemet. Vi ønsker ved hjælp af de mindste kvadraters metode at bestemme den bedste rette linje, der tilnærmer de tre punkter.

Vi indfører koordinaternes middeltal ved skrivemåden

$$\bar{x} = \tfrac{1}{3}(x_1 + x_2 + x_3) \quad , \quad \bar{y} = \tfrac{1}{3}(y_1 + y_2 + y_3)$$

Først viser vi

Sætning. Kvadratsummen

$$Q(a) = (y_1 - a)^2 + (y_2 - a)^2 + (y_3 - a)^2$$

er mindst, hvis a vælges som middeltallet af y_1, y_2 og y_3, dvs. hvis $a = \bar{y}$.

Bevis. Ved udregning får vi, at

$$\begin{aligned} Q(a) &= y_1^2 + y_2^2 + y_3^2 - 2a\cdot(y_1 + y_2 + y_3) + 3a^2 \\ &= 3a^2 - 2a\cdot 3\bar{y} + y_1^2 + y_2^2 + y_3^2 \end{aligned}$$

Dette er et andengradspolynomium i a, og efter formlen for andengradspolynomiets toppunkt antages mindsteværdien for

$$a = \frac{6\bar{y}}{2\cdot 3} = \bar{y}\,.$$

Den bedste rette linje

Den bedste rette linje, der tilnærmer de tre punkter, har ligningen

$$y = px + q$$

Vi kan skrive denne på formen

$$y = a + b(x - \bar{x})$$

ved at sætte $b = p$ og $q = a - b\bar{x}$, dvs. $a = q + b\bar{x}$.

Vi betegner de lodrette afstande fra de tre punkter til linjen med d_1, d_2 og d_3, dvs.

$$d_i = y_i - a - b(x_i - \bar{x})$$

Vi ønsker at finde konstanterne a og b, så kvadratsummen

$$\begin{aligned} s &= d_1^2 + d_2^2 + d_3^3 \\ &= \big(y_1 - a - b(x_1 - \bar{x})\big)^2 + \big(y_2 - a - b(x_2 - \bar{x})\big)^2 \\ &\qquad + \big(y_3 - a - b(x_3 - \bar{x})\big)^2 \end{aligned}$$

bliver så lille som mulig.

Bestemmelse af *a*

Ved udregning får vi

$$\begin{aligned} s &= (y_1 - a)^2 + (y_2 - a)^2 + (y_3 - a)^2 \\ &\quad + b^2\left((x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2\right) \\ &\quad - 2b\cdot\big((x_1 - \bar{x})(y_1 - a) + (x_2 - \bar{x})(y_2 - a) + (x_3 - \bar{x})(y_3 - a)\big) \end{aligned}$$

I sidste led udregner vi koefficienten til a:

$$(\bar{x} - x_1) + (\bar{x} - x_2) + (\bar{x} - x_3) = 3\bar{x} - (x_1 + x_2 + x_3) = 0$$

Vi betragter s som et andengradspolynomium i a. Den variable a, som vi ønsker at finde, optræder kun i de tre første led. Da s skal gøres så lille som muligt, skal summen af de tre første led altså gøres så lille som mulig. Efter ovenstående sætning sker dette ved at vælge $a = \bar{y}$.

Bestemmelse af *b*

Vi har nu, at

$$\begin{aligned} s &= (y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + (y_3 - \bar{y})^2 \\ &\quad + b^2\left((x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2\right) \\ &\quad - 2b\cdot\big((x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y}) + (x_3 - \bar{x})(y_3 - \bar{y})\big) \end{aligned}$$

For overskuelighedens skyld udregner vi delene i denne formel særskilt:

$$\begin{aligned} &(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + (y_3 - \bar{y})^2 \\ &= y_1^2 + y_2^2 + y_3^2 - 2\bar{y}\cdot(y_1 + y_2 + y_3) + 3\bar{y}^2 \\ &= y_1^2 + y_2^2 + y_3^2 - 2\bar{y}\cdot 3\bar{y} + 3\bar{y}^2 = y_1^2 + y_2^2 + y_3^2 - 3\bar{y}^2 \end{aligned}$$

og tilsvarende

$$(x_1-\bar{x})^2+(x_2-\bar{x})^2+(x_3-\bar{x})^2 = x_1^{\,2}+x_2^{\,2}+x_3^{\,2}-3\bar{x}^2$$

Endelig er

$$\begin{aligned}&(x_1-\bar{x})(y_1-\bar{y})+(x_2-\bar{x})(y_2-\bar{y})+(x_3-\bar{x})(y_3-\bar{y})\\&= x_1y_1+x_2y_2+x_3y_3-\bar{y}(x_1+x_2+x_3)-\bar{x}(y_1+y_2+y_3)+3\bar{x}\cdot\bar{y}\\&= x_1y_1+x_2y_2+x_3y_3-\bar{y}\cdot 3\bar{x}-\bar{x}\cdot 3\bar{y}+3\bar{x}\cdot\bar{y}\\&= x_1y_1+x_2y_2+x_3y_3-3\bar{x}\cdot\bar{y}\end{aligned}$$

Dermed har vi fundet følgende udtryk for kvadratsummen s:

$$\begin{aligned}s = {}&(x_1^{\,2}+x_2^{\,2}+x_3^{\,2}-3\bar{x}^2)b^2-2(x_1y_1+x_2y_2+x_3y_3-3\bar{x}\,\bar{y})b\\&+y_1^{\,2}+y_2^{\,2}+y_3^{\,2}-3\bar{y}^2\end{aligned}$$

Dette er et andengradspolynomium i b, og efter formlen for andengradspolynomiets toppunkt antages den mindste værdi for

$$b=\frac{2(x_1y_1+x_2y_2+x_3y_3-3\bar{x}\,\bar{y})}{2(x_1^{\,2}+x_2^{\,2}+x_3^{\,2}-3\bar{x}^2)}=\frac{x_1y_1+x_2y_2+x_3y_3-3\bar{x}\,\bar{y}}{x_1^{\,2}+x_2^{\,2}+x_3^{\,2}-3\bar{x}^2}$$

Dermed er både a og b fundet.

Den linje, der tilnærmer de tre punkter bedst (nemlig ved, at summen af kvadraterne på de lodrette afstande fra linjen til punkterne er mindst) har altså ligningen

$$y=a+b(x-\bar{x})$$

med de fundne værdier for a og b. Læg her mærke til, at linjen går gennem punktet med koordinaterne $(\bar{x},\bar{y})$. Dette koordinatsæt passer nemlig i linjens ligning.

Hvis vi i stedet for 3 punkter har forelagt n punkter med koordinaterne

$$(x_1, y_1)\ ,\ (x_2, y_2)\ ,\ \ldots\ ,\ (x_n, y_n)$$

får vi åbenbart følgende formler for a og b:

$$a=\bar{y}=\tfrac{1}{n}\left(y_1+y_2+\cdots+y_n\right)$$

$$b=\frac{x_1y_1+x_2y_2+\cdots+x_ny_n-n\bar{x}\,\bar{y}}{x_1^{\,2}+x_2^{\,2}+\cdots+x_n^{\,2}-n\bar{x}^2}$$

Eksempel. Hvis de tre punkter har koordinaterne

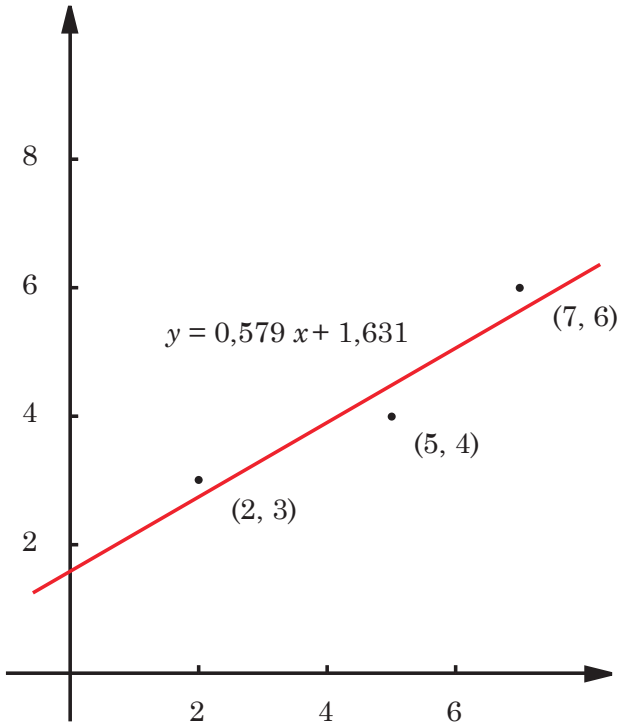
$(x_1, y_1) = (2, 3)$, $(x_2, y_2) = (5, 4)$, $(x_3, y_3) = (7, 6)$

får vi

$$\bar{x}=\tfrac{1}{3}(2+5+7)=\tfrac{14}{3}\quad,\quad \bar{y}=a=\tfrac{1}{3}(3+4+6)=\tfrac{13}{3}$$

hvoraf

$$b=\frac{6+20+42-3\cdot\frac{14}{3}\cdot\frac{13}{3}}{4+25+49-3\cdot\left(\frac{14}{3}\right)^2}=\frac{68-14\cdot\frac{13}{3}}{78-\frac{196}{3}}=\frac{11}{19}$$



Altså får den bedste rette linje (regressionslinjen) ligningen

$$y=\frac{13}{3}+\frac{11}{19}\left(x-\frac{14}{3}\right)\quad\Leftrightarrow\quad y=\frac{11}{19}x+\frac{31}{19}$$

Hvis man indtaster de tre punkters koordinater i et matematikprogram, får man netop dette (i decimaltal).

Henvisninger

Jens Carstensen: *Matematisk statistik* (Systime, 1983)

Douglas Downing & Jeff Clark: *Statistics the Easy Way* (Barron's Educational Series, 1989)

Carl Winsløw: *Fodgængerversion af lineær regression* (LMFK–Bladet 1, 2015)

Svend Erik Morsing: *Mere om lineær regression* (LMFK–Bladet 2, 2015)

Ann E. Watkins, Richard L. Scheaffer & George W. Cobb: *Statistics in Action* (Key Curriculum Press, 2004)