Spredningen på tendenslinjens hældning – en ukendt formel?

Kaj Ove Roland, Rødovre Gymnasium

Lineær regression kender vi alle sammen, og opgaver af denne type løses dagligt i det danske gymnasium. Typisk anvender man et værktøjsprogram, der ud over ligningen $y = a \cdot x + b$ for den bedste rette linje også serverer forklaringsgraden r^2.

Her vil jeg gerne reklamere for en formel, der synes at være aldeles ukendt, men som ikke desto mindre er særdeles nyttig, hvis man vil beregne spredningen (usikkerheden) på tendenslinjens hældning. I formlen indgår netop de størrelser, man typisk har til rådighed, dvs hældningen a, forklaringsgraden r^2 og antallet af datapunkter, n.

$$S_a = |a| \cdot \sqrt{\frac{1-r^2}{(n-2)\cdot r^2}}$$

Bevæbnet med denne formel bliver det dejligt nemt at nedskrive et 95 % konfidensinterval for hældningen, idet dette antager formen $[a - 1{,}96 \cdot S_a \,;\, a + 1{,}96 \cdot S_a]$.

En lignende formel kan udledes for spredningen (usikkerheden) på tendenslinjens skæring med andenaksen:

$$S_b = S_a \cdot \sqrt{\sigma_x{}^2 + \bar{x}^2}$$

Den er dog ikke helt så elegant, idet den kræver, at man først beregner spredningen σ_x og middelværdien $\bar{x}$ af datasættets x–værdier.

Formlerne herover kan jo udmærket anvendes uden at sætte sig ind i teorien bag, men for de interesserede bringes her en udledning af formlen for S_a.

Som sædvanlig antages, at vi har givet måleresultater (x_i, y_i), $i = 1, 2, \ldots, n$. Vi antager, at x–målingerne ikke er behæftet med nogen usikkerhed og at alle y–målinger er normalfordelt med samme spredning og med en middelværdi, der vokser lineært med x. Under disse antagelser finder man som bekendt, at det bedste estimat for a og b er givet ved mindste kvadraters metode, se formlerne (1) og (5) herunder, og at det bedste estimat for spredningen på y–målingerne er givet ved residualspredningen, se formel (7).

Vi starter derfor med at betragte formlen for tendenslinjens hældning, som man finder den ved mindste kvadraters metode:

$$a = \frac{n \cdot \sum x_i \cdot y_i - \sum x_i \cdot \sum y_i}{(\sum {x_i}^2) \cdot n - (\sum x_i)^2} \qquad (1)$$

hvor det i det følgende er underforstået, at alle summer løber fra 1 til n.

Det kan være nyttigt at indføre middelværdien og spredningen på x'erne:

$$\bar{x} = \frac{1}{n}\sum x_i \;\Leftrightarrow\; \sum x_i = n \cdot \bar{x} \qquad (2)$$

$$\begin{aligned} {\sigma_x}^2 &= \frac{1}{n}\sum (x_i - \bar{x})^2 = \frac{1}{n}\left(\sum x_i^2 - 2 \cdot \sum x_i \cdot \bar{x} + \sum \bar{x}^2\right) \\ &= \frac{1}{n}\left(\sum x_i^2 - 2 \cdot n \cdot \bar{x}^2 + n \cdot \bar{x}^2\right) = \frac{1}{n}\left(\sum x_i^2 - n \cdot \bar{x}^2\right) \;\Rightarrow \\ \sum x_i^2 &= n \cdot {\sigma_x}^2 + n \cdot \bar{x}^2 \end{aligned} \qquad (3)$$

Derved kan formel (1) omskrives til

$$a = \frac{\sum (x_i - \bar{x}) \cdot y_i}{n \cdot {\sigma_x}^2} \qquad (4)$$

mens b i tendenslinjens ligning kan beregnes som

$$b = \bar{y} - a \cdot \bar{x} \qquad (5)$$

Vi kan nu finde standardafvigelsen på a ved at bruge ophobningsloven

$$(S_a)^2 = \sum \left(\frac{\partial a}{\partial y_i} \cdot S_Y\right)^2 = \sum \left(\frac{x_i - \bar{x}}{n \cdot {\sigma_x}^2}\right)^2 \cdot {S_Y}^2 = \frac{{S_Y}^2}{n \cdot {\sigma_x}^2} \qquad (6)$$

Her er det bedste estimat for spredningen på y–værdierne givet ved

$$\begin{aligned} {S_Y}^2 &= \frac{1}{n-2} \cdot \sum \left(y_i - a \cdot x_i - b\right)^2 \\ &= \frac{1}{n-2} \cdot \sum \left(y_i - \bar{y} - a \cdot (x_i - \bar{x})\right)^2 \end{aligned} \qquad (7)$$

mens Pearsons korrelationskoefficient er givet ved

$$\begin{aligned} r &= \frac{\sum x_i \cdot y_i - \bar{x} \cdot \sum y_i}{\sqrt{\left(\sum {x_i}^2 - \frac{(\sum x_i)^2}{n}\right) \cdot \left(\sum {y_j}^2 - \frac{(\sum y_j)^2}{n}\right)}} \\ &= \frac{\sum (x_i - \bar{x}) \cdot y_i}{\sqrt{n} \cdot \sigma_x \cdot \sqrt{\sum {y_j}^2 - n \cdot \bar{y}^2}} \end{aligned} \qquad (8)$$

Den grundlæggende observation er nu, at der i formlerne (4) og (8) kun optræder to y–udtryk, nemlig $\sum (x_i - \bar{x}) \cdot y_i$ og $\sum {y_j}^2 - n \cdot \bar{y}^2$, og vi kan derfor erstatte disse to variable med de to variable a og r. Af formel (4) fås

$$\sum (x_i - \bar{x}) \cdot y_i = a \cdot n \cdot {\sigma_x}^2 \qquad (9)$$

og formel (8) giver ved kvadrering og anvendelse af (9):

$$\sum {y_j}^2 - n \cdot \bar{y}^2 = \frac{a^2 \cdot n^2 \cdot {\sigma_x}^4}{n \cdot {\sigma_x}^2 \cdot r^2} = \frac{a^2 \cdot n \cdot {\sigma_x}^2}{r^2} \qquad (10)$$

Vi kan nu omskrive formlen (7) ved først at bruge kvadratsætningen og dernæst anvende (9) og (10) til at fjerne alle explicitte y–udtryk:

$$\begin{aligned} {S_Y}^2 &= \frac{1}{n-2}\sum \left(y_i - \bar{y} - a \cdot (x_i - \bar{x})\right)^2 \\ &= \frac{1}{n-2}\left(\begin{array}{r} \sum (y_i - \bar{y})^2 - 2a \cdot \sum (x_i - \bar{x}) \cdot (y_i - \bar{y}) \\ + a^2 \cdot \sum (x_i - \bar{x})^2 \end{array}\right) \\ &= \frac{1}{n-2}\left(\begin{array}{r} \sum {y_i}^2 - n \cdot \bar{y}^2 - 2a \cdot \sum (x_i - \bar{x}) \cdot y_i \\ + 2a \cdot \sum (x_i - \bar{x}) \cdot \bar{y} + a^2 \cdot n \cdot {\sigma_x}^2 \end{array}\right) \\ &= \frac{1}{n-2}\left(\frac{a^2 \cdot n \cdot {\sigma_x}^2}{r^2} - 2a \cdot a \cdot n \cdot {\sigma_x}^2 + 0 + a^2 \cdot n \cdot {\sigma_x}^2\right) \end{aligned}$$

hvilket kan reduceres til

$${S_Y}^2 = \frac{n}{n-2} \cdot {\sigma_x}^2 \cdot a^2 \cdot \frac{1 - r^2}{r^2} \qquad (11)$$

Endelig kan (11) indsættes i (6), hvorved den omtalte formel fremkommer

$${S_a}^2 = \frac{1}{n-2} \cdot a^2 \cdot \frac{1 - r^2}{r^2} \quad \Leftrightarrow \quad S_a = |a| \cdot \sqrt{\frac{1 - r^2}{(n-2) \cdot r^2}}$$

På lignende vis kan man udlede en formel for spredningen på b. Man finder som tidligere nævnt

$$S_b = S_a \cdot \sqrt{{\sigma_x}^2 + \bar{x}^2}\ .$$