Lineær regression

HELGE BENNEDSEN, helge_bennedsen@mail.dk

Vi har en hel masse punkter punkter (x, y) i et koordinatsystem og ønsker at bestemme den linje med ligningen $y = ax + b$, som ligger tættest på punkterne, og som går under navnet *tendenslinjen*.

Det vi gør, er at bestemme a og b således, at det samlede gennemsnit af kvadraterne på $(ax + b - y)$ bliver mindst mulig, dvs.

funktionen $G(a,b) = \overline{(ax+b-y)^2}$ skal være mindst mulig, hvilket kræver lidt differentialregning:

1) $$\frac{\partial G}{\partial a} = \overline{2\cdot(ax+b-y)\cdot x} = 2(a\overline{x^2} + b\overline{x} - \overline{xy}) = 0$$
$$\Leftrightarrow a\overline{x^2} + b\overline{x} = \overline{xy}$$

2) $$\frac{\partial G}{\partial b} = \overline{2\cdot(ax+b-y)\cdot 1} = 2(a\overline{x} + b - \overline{y}) = 0$$
$$\Leftrightarrow a\overline{x} + b = \overline{y}$$

Hvis vi så strikker videre på de to slutligninger får vi at:

$$a = \frac{\overline{x\cdot y} - \overline{x}\cdot\overline{y}}{\overline{x^2} - \left(\overline{x}\right)^2} \qquad b = \frac{\overline{y}\cdot\overline{x^2} - \overline{x}\cdot\overline{y\cdot x}}{\overline{x^2} - \left(\overline{x}\right)^2}$$

Hvis man nu indsætter de fundne værdier af a og b i ligningen $y = ax + b$, er det ganske nemt at påvise, at punktet med koordinatsættet $\left(\overline{x}, \overline{y}\right)$ ligger på den fundne tendenslinje.

At funktionen $G(a,b) = \overline{(ax+b-y)^2}$ har et minimum, er det nok rimeligt at argumentere for. Det ses ret nemt på funktionsudtrykket, at G ikke har noget maximum, da man kan vælge a og b så store, man vil, således at G vokser mod uendelig. På den anden side må G have mindst et lokalt minimum, da G ikke kan være negativ, og hvor dens partielle differentialkvotienter er lig med nul, hvilket vi har undersøgt i 1) og 2)

Lad os forestille os, at vi vil bestemme en tendenslinje på formlen $x = dy + e$ således, at en anden funktion (forudsat, at a er forskellig fra nul!)

$$K(d,e) = \overline{\left(dy+e-x\right)^2}$$

skal minimeres på tilsvarende måde som G, hvilket blot betyder, at vi i nogle tidligere fundne formler skal bytte om på x og y. Vi får følgende udtryk:

$$d = \frac{\overline{x\cdot y} - \overline{x}\cdot\overline{y}}{\overline{y^2} - \left(\overline{y}\right)^2} \qquad e = \frac{\overline{x}\cdot\overline{y^2} - \overline{y}\cdot\overline{y\cdot x}}{\overline{y^2} - \left(\overline{y}\right)^2}$$

Hvis alle punkter (x, y) ligger på den fundne tendenslinje, så ville der gælde, at $a\cdot d = 1$, hvilket ikke vil være tilfældet, såfremt blot et punkt ligger uden for tendenslinjerne. Men jo tættere $a\cdot d$ kommer på tallet 1, jo bedre er tendenslinjerne.