Mindste Kvadraters Rette Linje

Ni[1]

I praksis støder man ofte på datasæt, hvor man forventer en lineær sammenhæng mellem to variable. Måler man for eksempel samhørende værdier af strømstyrke og spændingsfaldet over en resistor, forventer vi en lineær sammenhæng på grund af Ohms lov. Indtegnes målepunkterne i et koordinatsystem vil man imidlertid opdage, at punkterne kun tilnærmelsesvis falder på en ret linje.

Problemet er nu at finde den rette linje, der "passer bedst" med målepunkterne. Vi skal nedenfor præcisere, hvad der ligger i den noget upræcise formulering - "passer bedst".

I betegnelsen "mindst kvadraters rette linje" henvises til at regressions-linje er bestemt, så summen af kvadraterne på de "lodrette" afstande mellem målepunkterne og regressions-linjen bliver mindst mulig. Ved at anvende kvadratet på afvigelserne fra linjen opnår vi, at alle afvigelser bidrager positivt til summen, så afvigelser ikke helt eller delvis kan udligne hinanden.

Givet et sæt måledata, hvor vi forventer en lineær sammenhæng mellem x og y.

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

Mindste kvadraters rette linje minimerer summen af kvadratet på y-værdiernes afvigelse fra punkterne på en ret linje. Summen af kvadratet på afvigelserne er givet ved:

$$\sum_{j=1}^{n}[y_j-(ax_j+b)]^2$$

Konstanterne a og b er bestemt ved:

$$a=\frac{\sum_{j=1}^{n}y_j(x_j-\mu)}{\sum_{j=1}^{n}(x_j-\mu)^2}$$

$$b=\frac{1}{n}\sum_{j=1}^{n}[y_j-\mu a]$$

hvor μ er middelværdien af x-værdierne.

[1] Datadrev\Matematik\Sandsynklighed og Statistik 151009.wpd

Konstanterne a og b i forskriften: y = a*x +b kan naturligvis ikke bestemmes uden usikkerhed ud fra målepunkter, der ikke helt ligger på en ret linje. Vores bedste bud på usikkerhed på a og b er standard-afvigelserne S_a og S_b, der beregnes på følgende måde:

$$s_a^2 = \frac{s_y^2}{\sum_{j=1}^{n} (x_j - \mu)^2}$$

$$s_b^2 = \frac{1}{n} s_y^2 + \mu^2 s_a^2$$

hvor værdien af s_y^2 beregnes ud fra formlen:

$$s_y^2 = \frac{1}{n-2} \sum_{j=1}^{n} [y_j - (ax_j + b)]^2$$

Forudsætningen for gyldigheden af formlerne til beregning af a, b, S_a og S_b er følgende:

I *Spredningen på x-værdierne er mindre end spredningen på y-værdierne. Man vælger altså den mest nøjagtigt målte størrelse til x.*

II *For hvert x er y-værdien en normaltfordelt stokastisk variabel med middelværdi a*x+b og en spredning, der ikke afhænger af x.*

Man kan undersøge, om forudsætningerne er opfyldt, ved at gentage målingen af y for fastholdt x - og omvendt. Men den efterfølgende statistiske analyse er ret kompliceret.

I praksis går det ofte godt at anvende metoden uden at tænke nærmere over forudsætningerne, men det er vigtigt, at der altid foretages en "**visuel vurdering**". Det vil blot sige, at man indtegner regressionslinjen sammen med målepunkterne ,og overvejer, om linjen ligger fornuftigt i forhold til punkterne.

Man ser ofte en vurdering af regressionslinjens kvalitet ved hjælp af parameteren r^2. Denne vurdering er altid dårligere end en visuel vurdering og ofte direkte misvisende.

Regression kan med fordel foretages med fysik-programmet DATASTUDIO, hvis man har adgang til dette program; eller med regnearket MINKVAD, der udleveres ved henvendelse til undertegnede.