# Regression eller "bedste rette linie"

Ole Witt-Hansen, Køge Gymnasium

**At man nu har fjernet tæppet fuldstændig under eleverne, hvad angår forståelsen af behandling af måleresultater er evident, men det er også lidt forstemmende, at adskillige lærere heller ikke kender den teoretiske baggrund for at anvende regression, mindste kvadraters metode, korrelationskoefficient og Chi-Square probability.**

I 2007 afsluttede jeg et fysik C niveau med at lave nogle øvelser med absorption af radioaktiv stråling. Jeg havde tænkt mig, som jeg har gjort det i mere end 30 år, at eleverne skulle tegne punkterne ind på enkeltlogaritmisk papir, aflæse halveringstykkelsen for at beregne den lineære absorptionskoefficient. Til min overraskelse anede eleverne hverken, hvad enkeltlogaritmisk papir var, eller hvad betydningen af en ret linie på et sådant papir indebar.

Til gengæld havde de lært, hvordan man bestemmer en regressionslinie i Excel eller ved at taste data ind på deres matematikcomputer. Dertil og ikke længere rakte deres forståelse af begrebet regression. Det er nu ikke første gang, jeg er sunket hen i tankefuldhed over den udvikling, der er sket med undervisningen i fysik i de senere år. Fra visuel empiri til IT-baseret "heksekunst".

Jeg har altid forklaret mine elever, at den eneste kurve, man med sikkerhed kan genkende, er en ret linie. Derfor anvender man de forskellige typer mm-papir til at identificere forskellige funktionelle sammenhænge for en række måledata. Når eleverne selv afsætter punkterne og selv konstaterer, at de ligger på en ret linie, så er der et enkelt begrebsmæssigt grundlag for at forstå, hvad måleresultaterne viser. (I mine fysiktimer går det stadig ud på at vise nogle lovmæssigheder – ikke at forsøge at falsificere lovmæssigheder, som har været kendt i mere end 100 år, som den fysikbog jeg har anvendt efter reformen ellers lægger op til).

Jeg har stort set opgivet at argumentere med eleverne, når de lidt forurettede hævder (og med fuld ret efter reformen), at matematikcomputerens 7 cifre for konstanterne i regressionsligninger er mere "præcise" end aflæsning fra en linie, de har tegnet.

Det er muligvis også tilfældet med præcisionsmålinger, der er korrigeret for systematiske og andre fejl. Det er imidlertid meget let at give eksempler på fysikøvelser, der indeholder systematiske fejl, eller hvor lineariteten ikke gælder for yderpunkterne, eller hvor der simpelthen er en fejlmåling.

Alle disse ting vil imidlertid være lette at afgøre efter punkternes beliggenhed, hvorefter man kan vælge, at se delvis bort fra sådanne punkter, når man tegner den bedste rette linie, mens en beregning af en regressionslinie ikke tager højde for sådanne afvigelser, og hvor en fejlmåling kan få overordentlig stor betydning.

Hvis jeg spørger eleverne, hvordan de kan vide, om punkterne følger en ret linie, så kan de godt finde på at henvise til en korrelationskoefficient eller for nylig har jeg minsandten hørt om *Chi-square probability*!

At man nu har fjernet tæppet fuldstændig under eleverne, hvad angår forståelsen af behandling af måleresultater er evident, men det er også lidt forstemmende, at adskillige lærere heller ikke kender den teoretiske baggrund for at anvende regression, mindste kvadraters metode, korrelationskoefficient og Chi-Square probability.

Som student fra 1964 er jeg i den situation, at jeg aldrig, hverken i gymnasiet eller på universitetet har modtaget undervisning i sandsynlighedsregning eller statistik. Det er måske grunden til, at jeg har undersøgt sagerne lidt dybere, end hvad gymnasielærebøgerne tilbyder.

## Mindste kvadraters metode

Mindste kvadraters metode er et specialtilfælde af "Method of maximum likelihood", som allerede har været anvendt af Gauss, og beskrevet udførligt af R.A. Fischer (1912).

Lad $A$ og $B$ være to ligeberettigede "teorier" og lad $E$ være et eksperiment, som udføres: givet en af de to teorier. Eksperimentet skal afgøre, hvilken af de to teorier, der er den mest sandsynlige (likely).

Ud fra Bayes teorem:

$$P(A|E)P(E) = P(E|A)P(A) \text{ og}$$
$$P(B|E)P(E) = \mathrm{P}(E|B)P(B)$$

med $P(A) = P(B)$ ifølge antagelsen, finder man ved division af de to ligninger:

$$\frac{P(A|E)}{P(B|E)} = \frac{P(E|A)}{P(E|B)}$$

som blot udtrykker, at den mest sandsynlige teori givet eksperimentet, er den teori, som giver det mest sandsynlige udfald af eksperimentet – hvilket næppe er overraskende. Dette er grundlaget for "Principle of maximum likelihood".

Lad os antage, at vi har $n$ uafhængige observationer (målinger) $x_1$, $x_2$, … $x_n$ som har en sandsynlighedsfordeling $P(X,\alpha)$, som afhænger af en parameter $\alpha$, som vi ønsker at bestemme. Sandsynligheden for netop dette udfald af eksperimentet er produktet af sandsynlighederne for hver af de uafhængige målinger.

$$L(\alpha) = P(x_1,\alpha)\, P(x_2,\alpha) \ldots P(x_n,\alpha)$$

$L(\alpha)$ kaldes for Likelihood funktionen. Den "bedste" teori er – ifølge Bayes teorem – den værdi af α, som maksimerer $L(\alpha)$ ("Method of maximum likelihood").

I praksis tager man minus logaritmen til $L(\alpha)$ og bestemmer minimum.

$$-\ln(L(\alpha)) = -\ln(P(x_1,\alpha))-\ln(P(x_2,\alpha)) - \ldots -\ln(P(x_n,\alpha))$$

Antager vi nu *specielt,* at $x_1$, $x_2$, … $x_n$ er normalfordelte med spredning $\sigma_i$ og samme teoretiske middelværdi $\mu(\alpha)$, således at

$$P(x_i,\alpha) = \frac{1}{\sqrt{2\pi}\sigma_i}\, e^{-\frac{(x_i-\mu(\alpha))^2}{2\sigma_i^2}}$$

Udregner man nu $-\ln(L(\alpha))$ og undlader alle led, der ikke afhænger af $x_i$ eller $\alpha$, da de er uden betydning for bestemmelse af minimum for $-\ln(L(\alpha))$ finder man:

$$\frac{(x_1-\mu(\alpha))^2}{2\sigma_1^2} + \frac{(x_2-\mu(\alpha))^2}{2\sigma_2^2} + \ldots + \frac{(x_n-\mu(\alpha))^2}{2\sigma_n^2}$$

Hvis der er den samme teoretiske spredning $\sigma$ på alle $x_i$ (og kun da), får man det velkendte udtryk for $\chi^2$ ved at gange udtrykket igennem med $2\sigma^2$

$$\chi^2 = (x_1-\mu(\alpha))^2 + (x_2-\mu(\alpha))^2 + \ldots + (x_n-\mu(\alpha))^2$$

Ved at udregne minimum for denne funktion finder man den mest sandsynlige værdi for α. Dette gøres på sædvanlig vis ved at løse ligningen

$$\frac{\partial\chi^2}{\partial\alpha} = 0.$$

Disse resultater kan nemt generaliseres til flere parametre og flere variable.

### Maximum likelihood

*From Wikipedia, the free encyclopedia*

**Maximum likelihood estimation (MLE)** is a popular statistical method used to calculate the best way of fitting a mathematical model to some data. Modeling real world data by estimating maximum likelihood offers a way of tuning the free parameters of the model to provide an optimum fit.

The method was pioneered by geneticist and statistician Sir R. A. Fisher between 1912 and 1922. It has widespread applications in various fields, including:

- linear models and generalized linear models are commonly fit by maximum likelihood.
- econometrics and hypothesis testing in medical research.
- time-delay of arrival (TDOA) in acoustic detection.
- origin/destination and path-choice modeling in transport networks.

### Halveringstid for et radioaktivt præparat

Lad os som eksempel se på det velkendte forsøg, hvor man ønsker at bestemme halveringstiden for et radioaktivt præparat. Den teoretiske værdi for aktiviteten $y$ er $y = ae^{-kt}$.

Nu er $N$ (antallet af sønderdelinger i et givet tidsrum) imidlertid ikke normalfordelt (det er Poissonfordelt med spredning $\sqrt{N}$), så spredningen $\sigma_i = \sigma_i(y_i)$ afhænger af aktiviteten.

Man kan muligvis se bort fra forskellene i normal- og Poissonfordelingen, men spredningen er under ingen omstændigheder den samme for alle målinger. Vi må derfor bestemme konstanterne $a$ og $k$ som minimum for:

$$f(a,k) = \frac{(y_1-ae^{-kt_1})^2}{2\sigma_1^2} + \frac{(y_2-ae^{-kt_1})^2}{2\sigma_2^2} + \ldots + \frac{(y_n-ae^{-kt_1})^2}{2\sigma_n^2}$$

Jeg tager det imidlertid som en selvfølge, at matematikcomputeren i alle tilfælde bestemmer konstanterne ved mindste kvadraters metode. Det har jeg såmænd heller ikke noget at indvende imod, men når man foregøgler eleverne, at det er mere "videnskabeligt" at anvende mindste kvadraters metode end selv at tegne en linie, så bygger det mere på "videnskabeligt krukkeri" end på kendskab til matematisk statistik.

## Lineær sammenhæng

Specielt vil vi se på det tilfælde, hvor vi har en stokastisk variabel $Y$, som ifølge teorien afhænger lineært af $x$. Der gælder således: $y_i = a \cdot x_i + b$. (Det antages, at der ikke er nogen spredning på $x$, hvilket langtfra altid er tilfældet). Den teoretiske middelværdi for $y_i$ bliver derfor $\bar{y}_i = a\bar{x}_i + b$. Vi vil også antage, at alle $y_i$-erne er normalfordelte og har samme spredning, (hvilket langtfra altid gælder), men som er forudsætningen for at kunne anvende mindste kvadraters metode.

Som beskrevet ovenfor fører "Principle of maximum likelihood" til, at vi skal finde minimum for kvadratsummen, som funktion af $a$ og $b$.

$$f(a,b) = (y_1-(ax_1+b))^2+(y_2-(ax_2+b))^2+...+(y_n-(ax_n+b))^2$$

eller med en mere kompakt skrivemåde:

$$f(a,b) = \sum_{i=1}^{n}(y_i - ax_i - b)^2$$

Ved at udregne $\frac{\partial f}{\partial a}$ og $\frac{\partial f}{\partial a}$ og sætte dem lig med nul, får man ligningerne:

$$\frac{\partial f}{\partial a} = \sum -2x_i(y_i - ax_i - b) = 0 \wedge \frac{\partial f}{\partial b} = \sum -2(y_i - ax_i - b) = 0$$
$$\Leftrightarrow$$
$$a\sum x_i^2 + b\sum x_i = \sum x_i y_i \wedge a\sum x_i + n \cdot b = \sum y_i$$

Af den sidste ligning følger:

$$b = \frac{1}{n}\sum y_i - a\frac{1}{n}\sum x_i = \bar{y} - a\bar{x},$$

hvor $\bar{y} = \frac{1}{n}\sum y_i$ og $\bar{x} = \frac{1}{n}\sum x_i$.

Ved at indsætte udtrykket for $b$ i den første ligning finder man så $a$.

$$a\sum x_i^2 + (\bar{y} - a\bar{x})\sum x_i = \sum x_i y_i \Leftrightarrow$$
$$a\left(\sum x_i^2 - n \cdot \bar{x}^2\right) = \sum x_i \cdot y_i - n \cdot \bar{x} \cdot \bar{y}$$

Man får da:

$$a = \frac{\sum x_i^2 - n\bar{x}^2}{\sum x_i \cdot y_i - n \cdot \bar{x} \cdot \bar{y}}$$

og

$$b = \frac{1}{n}\sum y_i - a\frac{1}{n}\sum x_i = \bar{y} - a\bar{x}$$

I forbindelse med lineær regression og matematikcomputere er begrebet korrelationskoefficient (desværre) dukket op i gymnasieundervisningen helt ned til c-niveau. Igen er det min fornemmelse, at det ikke er alle matematiklærere, der helt er klar over, hvorledes man udregner en korrelationskoefficient, og hvad betydningen af den er. Som det bliver understreget i [1], så er udregning af korrelationskoefficient (ligesom lineær regression) kun meningsfuld for variable, som er normalfordelte.

Indfører man to stokastiske variable $X$ og $Y$:

$$s_x^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2 = \frac{1}{n-1}\left(\sum x_i^2 - n\bar{x}^2\right)$$

$$s_y^2 = \frac{1}{n-1}\sum(y_i - \bar{y})^2 = \frac{1}{n-1}\left(\sum y_i^2 - n\bar{y}^2\right)$$

$$s_{xy} = \frac{1}{n-1}\sum(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1}\left(\sum x_i \cdot y_i - n \cdot \bar{x} \cdot \bar{y}\right)$$

er korrelationskoefficienten ifølge [1] defineret ved:

$$r_{xy} = \frac{s_{xy}}{s_x \cdot s_y} = \frac{\sum(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2} \cdot \sqrt{\sum(y_i - \bar{y})^2}}$$

Korrelationskoefficienten er (ifølge Cauchy Schwartz' ulighed) beliggende i intervallet [-1,1]. Ofte anvender man den numeriske værdi.

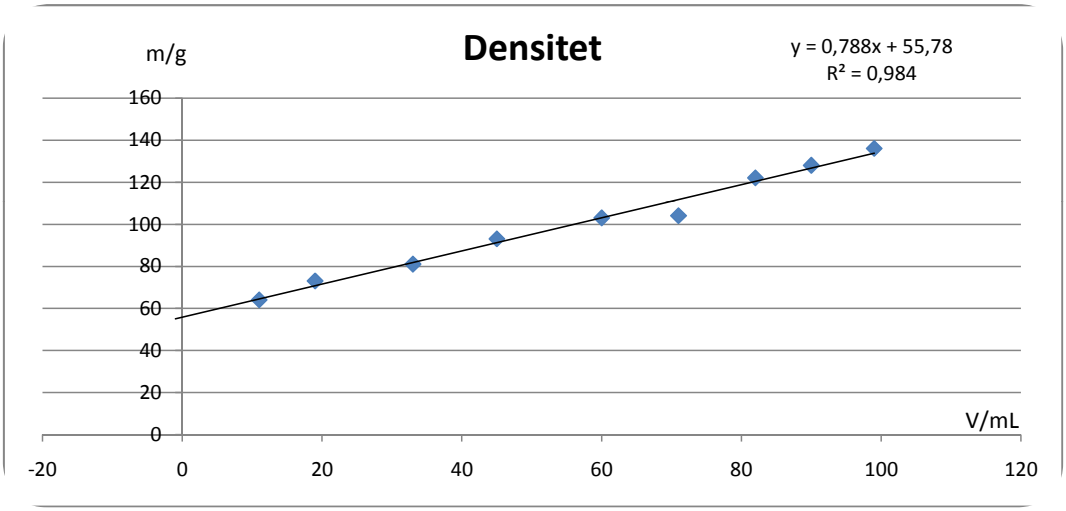
## Geometrisk tolkning

Opfatter man $(x_i - \bar{x})$ og $(y_i - \bar{y})$ som koordinaterne til to vektorer $\vec{a}$ og $\vec{b}$ i et $n$- dimensionalt metrisk rum, så ses det, at

$$r_{xy} = \cos(v) = \frac{\vec{a} \cdot \vec{b}}{|\vec{a}||\vec{b}|}$$

netop er cosinus til vinklen mellem de to vektorer $\vec{a}$ og $\vec{b}$.

Hvis $|r_{xy}| = 1$, så er de to vektorer parallelle og følgelig er koordinatsættene $(x_i - \bar{x})$ og $(y_i - \bar{y})$ proportionale, så $y$ er en lineær funktion af $x$. Hvis derimod $|r_{xy}| = 0$, er de to vektorer ortogonale.




Den geometriske fortolkning af korrelationskoefficienten kan kaste lys over, at eleverne i Excel eller på deres matematikcomputer finder "fine" korrelationskoefficienter for måledata, der i grafisk afbildning på papir (og ikke på en gnidder-matematik-computerskærm) ser elendige ud.

Tag f.eks. $r_{xy}$ = 0,9 (kun 10% afvigelse fra 1) Dette giver en vinkel mellem vektorerne på 25,84°, hvilket man (i den ikke elektroniske geometri) næppe ville kalde udpræget parallelitet.

$r_{xy}$ = 0,95 giver 18,19° (som heller ikke vil bestå en ædruelighedsprøve om at følge en ret linie). Endelig $r_{xy}$ = 0,99 giver 8,10°, som vel må betragtes som acceptabelt.

En angivelse af en korrelationskoefficient på under 0,99 er ikke noget særlig godt kriterium for, at en række punkter ligger på en ret linie, mens en visuel vurdering i reglen er.

## Chi-square

Jeg synes, at det er meget problematisk at indføre teoretiske regressionslinier i gymnasieundervisningen, især efter den kraftige svækkelse af det teoretiske niveau efter reformen.

Men når man også indfører Chi-Square probablity, så synes jeg man har mistet jordforbindelsen til matematikundervisningen i gymnasiet. Nogle motoriske færdigheder på matematikcomputeren, kan eleverne vel lige så godt lære i biologi eller et andet fag, men jeg synes stadig, at man skal kunne forklare eleverne det, man lærer i gymnasiets matematikundervisning, og jeg synes også at det er et rimeligt krav (i gymnasiet), at lærerne selv forstår det, de underviser i.

Udtrykket for Chi-square sandsynlighedsfordelingen (kun relevant, hvis målingerne er normalfordelte) er udledt i alle de tre referencer (på 3 vidt forskellige måder). $N$ er antallet af frihedsgrader (antal uafhængige målinger), og den angiver sandsynligheden for at få en værdi, der ligger i intervallet $[\chi^2,\chi^2+d\chi^2]$. $\Gamma$ er Gamma (fakultetsfunktionen). For heltalligt $n$ er $\Gamma(n+1) = n!$

$$f(\chi^2) = \frac{1}{2^{\frac{N}{2}}\cdot\Gamma(\frac{N}{2})}\cdot e^{-\frac{\chi^2}{2}}\cdot(\chi^2)^{\frac{N}{2}-1}\cdot d\chi^2$$

Heraf følger fordelingsfunktionen.

$$F(t) = P(\chi^2>t) = \frac{1}{2^{\frac{N}{2}}\Gamma(\frac{N}{2})}\int_t^{\infty} e^{-\frac{\chi^2}{2}}\cdot(\chi^2)^{\frac{N}{2}-1}\cdot d\chi^2$$

Hvis man laver en "hypotesetest" med et signifikansniveau på 5%, så skal man altså kræve at $F(t) > 0{,}95$, så sandsynligheden for at få en bedre (mindre) $\chi^2$ højest er 5%.

Men jeg synes fortsat, at man skulle afholde sig fra at undervise i områder af matematikken, hvor en teoretisk forklaring på gymnasialt niveau er udelukket. ◊

---

(1) A. Hald: Statistical theory with engineering applications. Wiley 1952
(2) Harald Cramér: Mathematical methods og statistics. Princeton 1946 – 1974
(3) Jon Mathews & R.L. Walker: Mathematical Methods of physice. Benjamin 1970.