Projekt 8.5 Formlen for lineær regression

(*Dette projekt er en fuldførelse af øvelse 8.11, s. 314-16 i grundbogen, hvor vi her udleder formlen for den lineære regression.*)

I 1880'erne indsamlede Francis Galton (1822-1911) et datasæt, der indeholdt sammenhørende målinger af fædres højder og deres førstefødte sønners højder som voksne. Der er målinger for i alt 952 par af fædre og sønner. Formålet med arbejdet var dengang at få en bedre indsigt i, hvordan egenskaber nedarves.

Historisk set var det analysen af dette datasæt, hvor du nedenfor ser de første 10 målinger (angivet i cm), der gav anledning til det umiddelbart besynderlige navn "regressionsanalyse". Regression er det modsatte af progression og betyder tilbageskridt. Konklusionen på undersøgelsen var nemlig, at særligt høje fædre i gennemsnit fik knap så høje sønner – og det opfattede Galton som et tilbageskridt! Men samtidig fik små fædre sønner, der også var små, men dog lidt højere.

Øvelse 8.11 Udfør lineær regression ved hjælp af et matematisk værktøjsprogram

a) Udfør et punktplot af datasættet ovenfor, og kommenter, om plottet giver anledning til at tro på, at der findes en bagvedliggende lineær sammenhæng.

b) Udfør lineær regression på datasættet. Du skal få en formel som: $y = 1.94x - 183.58$

c) Tegn residualplottet og kommenter på grundlag af det, om du stadig mener, der er tale om en lineær sammenhæng.

nummer	fars højde	sønnens højde
1	186,9	183,4
2	184,6	172,0
3	185,0	179,0
4	182,1	165,4
5	179,4	155,4
6	178,4	160,3
7	179,7	165,0
8	179,7	168,7
9	176,5	160,3
10	173,4	157,5

De første 10 ud af Galtons datasæt på i alt 952

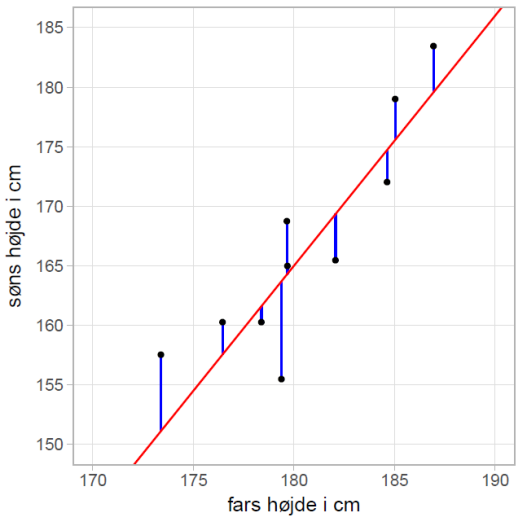
Vi har nu fået tegnet punktplottet sammen med regressionslinjen. At denne linje er "den bedste rette linje" beregnet ud fra "mindste kvadraters metode" betyder følgende, se figuren:

1. Vi måler **den lodrette afstand** mellem linjen og datapunkterne – det er længden af de blå linjestykker. Kalder vi fædrenes højder for $x_1, x_2, \ldots, x_{10}$ og sønnernes højder for $y_1, y_2, \ldots, y_{10}$, og er regressionslinjens forskrift $y = a \cdot x + b$, så er den lodrette afstand for det første datapunkt:

$$y_1 - (a \cdot x_1 + b) \qquad (*)$$

Overvej selv dette!

2. Vi ønsker ét samlet tal, der skal måle afstanden mellem linjen og datasættet. Læg mærke til, at nogle gange, fx ved punkt nr 3 og 4, vil formlen (*) give et negativt tal. Men hver afstand skal give et positivt bidrag. Derfor vælger vi at kvadrere disse tal og bagefter summere. (I videregående statistik argumenteres for denne fremgangsmåde).



3. Det samlede udtryk for afstanden mellem linjen og datasættet er derfor:

$$S = (y_1 - (a \cdot x_1 + b))^2 + (y_2 - (a \cdot x_2 + b))^2 + \ldots + (y_{10} - (a \cdot x_{10} + b))^2$$

Bemærk: Der ligger i det foregående nogle subjektive valg, der ikke kan begrundes alene ud fra en matematisk argumentation.
- For det første vælger vi at måle afstanden lodret. Den mindste afstand fra et punkt til en linje findes jo som den vinkelrette afstand. Dette giver imidlertid nogle ganske klodsede udtryk, som det er svært at regne på. Endvidere er det svært at generalisere metoden til fx polynomisk regression. Så vi *vælger* at se på den lodrette afstand. Det giver ikke alene en fordel mht. udregninger, men betyder også, at vi ser på afstand fra den empiriske værdi til modelværdien.
- For det andet vælger vi at se på kvadratet på afstanden og ikke kun afstanden. Det er også et valg, der bl.a. er bestemt af, at afstand måles som den numeriske værdi, og denne størrelse er svær at håndtere vi store udtryk. Men det betyder, at afvigere kan komme til at betyde uforholdsmæssigt meget, da afstandene kvadreres.
Mindste kvadraters metode giver en række beregningsmæssige fordele, men det er vigtigt at huske, at bag en beregning af en regressionsmodel ligger sådanne subjektive beslutninger

4. Udtrykket i punkt 3 er opskrevet med tanke på den forskrift, vi får ved at udføre regression. Men udtrykket kan skrives op med *ethvert* lineært udtryk, $y = a \cdot x + b$. Blandt alle disse uendeligt mange udtryk for *S*, vælger vi nu den linje, der giver *det mindste tal*. Det er præcis her, differentialregningen kommer på banen. *S* kan betragtes som en funktion af de to variable *a* og *b*, $S(a,b)$, og de værdier af *a* og *b*, der giver den mindste kvadratsum, findes som et minimumssted for funktionen $S(a,b)$. Funktioner af to variable behandles i bog 3, men det er principielt samme metode, som vi kender fra funktioner af én variabel:

- Vi differentierer først mht. *b* og løser ligningen: $S'_b(a,b) = 0$
- Vi differentierer dernæst mht. *a* og løser ligningen $S'_a(a,b) = 0$

Resultatet af det hele kan sammenfattes i denne sætning, som vi beviser nedenfor

Sætning 2: Formlen for den lineære regressionslinje
Lad der være givet et datasæt bestående af n punkter, (x_i, y_i), hvor $i = 1..n$.
Vi indfører betegnelserne $\overline{x}$ og $\overline{y}$ for middelværdierne (gennemsnittene):

$$\overline{x} = \frac{1}{n} \cdot (x_1 + x_2 + .. + x_n) = \frac{1}{n} \cdot \sum_1^n x_i, \quad \text{og} \quad \overline{y} = \frac{1}{n} \cdot (y_1 + y_2 + .. + y_n) = \frac{1}{n} \cdot \sum_1^n y_i$$

Koefficienterne i regressionslinjen (*den bedste rette linje*) betegnes $\hat{a}$ og $\hat{b}$ og beregnes således:

$$\hat{a} = \frac{\sum_{i=1}^n (y_i - \overline{y}) \cdot (x_i - \overline{x})}{\sum_{i=1}^n (x_i - \overline{x})^2} \qquad \hat{b} = \overline{y} - \hat{a} \cdot \overline{x}$$

Bevis.
Las os nu sige, vi har *n* datapunkter. Vi opskriver udtrykket for *S*:

$$S = \left(y_1 - (a \cdot x_1 + b)\right)^2 + \left(y_2 - (a \cdot x_2 + b)\right)^2 + ... + \left(y_n - (a \cdot x_n + b)\right)^2 = \sum_{k=1}^n \left(y_k - (a \cdot x_k + b)\right)^2$$

Sumtegnet er en kompakt og overskuelig måde at skrive lange udtryk på. Men man skal lige vænne sig tiol det, så vi vil veksle.
Først vil vi gange parenteserne ud. Den generelle parentes udregnes:

$\left(y_k - (a \cdot x_k + b)\right)^2$	Benyt kvadratsætningen på de to led:
$= {y_k}^2 + (a \cdot x_k + b)^2 - 2 \cdot y_k \cdot (a \cdot x_k + b)$	Benyt kvadratsætningen på $(a \cdot x_k + b)^2$:
$= {y_k}^2 + \left((a \cdot x_k)^2 + b^2 + 2 \cdot a \cdot x_k \cdot b\right) - 2 \cdot y_k \cdot (a \cdot x_k + b)$	Hæv de første og gange ind i den sidste parentes
$= {y_k}^2 + a^2 \cdot {x_k}^2 + b^2 + 2 \cdot a \cdot x_k \cdot b - 2 \cdot y_k \cdot a \cdot x_k - 2 \cdot y_k \cdot b$	(*)

Vi skal bestemme *a* og *b* så dette udtryk minimeres. Læg mærke til, at *x'*erne og *y'*erne er konstanter. Det er jo de givne datapunkter. Det betyder, at udtrykket kan ses som en funktion af to variable, *a*, og *b*. Grafen for funktionen $S(a,b)$ er en flade der ligger over $(a,b)-$planen, hvor *a* og *b* er de uafhængige variable afsat ud af henholdsvis 1. og 2. aksen. Vi leder efter et minimum for denne flade.

Men det er klart, at findes et sådant minimum, hvor fladen er "nede i en bølgedal", så er dette minimumspunkt også minimum for den funktion af *a*, hvis graf vi får ved at lægge et lodret snit gennem minimumspunktet og parallel med *a*-aksen. Og tilsvarende er det minimumspunkt for den funktion af *b* hvis graf vi får ved at lægge et lodret snit gennem minimumspunktet og parallel med *b*-aksen.

Derfor vil vi nu bestemme først *b* ved at holde *a* fast (dvs vi ser på snitkurven parallel med *b*-aksen). Og derefter *a* efter samme opskrift.

Vi differentierer udtrykket (*) mht *b*:

$$\frac{d}{db}\left(y_k{}^2+a^2\cdot x_k{}^2+b^2+2\cdot a\cdot x_k\cdot b-2\cdot y_k\cdot a\cdot x_k-2\cdot y_k\cdot b\right)$$
$$=2\cdot b+2\cdot a\cdot x_k-2\cdot y_k$$

Det var det k'te led, og nu summerer vi og sætter lig med 0 – overvej nøje, at vi kan gøre det, ved fx at skrive det ud som en stor sum:

$\sum_{k=1}^{n}\left(2\cdot b+2\cdot a\cdot x_k-2\cdot y_k\right)=0$	Vi summerer de tre led hver for sig, og rykker det sidste over:
$\sum_{k=1}^{n}2\cdot b+\sum_{k=1}^{n}2\cdot a\cdot x_k=\sum_{k=1}^{n}2\cdot y_k$	Vi forkorter med 2:
$\sum_{k=1}^{n}b+\sum_{k=1}^{n}a\cdot x_k=\sum_{k=1}^{n}y_k$	Vi sætter *a* udenfor sumtegn, dvs. parentes, og tæller antal *b*:
$n\cdot b+a\cdot\sum_{k=1}^{n}x_k=\sum_{k=1}^{n}y_k$	Vi dividerer med*:*
$b+a\cdot\frac{1}{n}\sum_{k=1}^{n}x_k=\frac{1}{n}\sum_{k=1}^{n}y_k$	Vi anvender definitionen på gennemsnit, som kaldes $\overline{x}$ og $\overline{y}$:
$b+a\cdot\overline{x}=\overline{y}$	Vi isolerer *b*:
$b=\overline{y}-a\cdot\overline{x}$	Det er formlen for *b* i sætningen.

Derefter vil vi nu bestemme *a* ved at holde *b* fast (dvs. vi ser på snitkurven parallel med *a*-aksen):
Men først indsættes den *b*-værdi, som vi nu kender i udtrykket (*)

$S(a) = {y_k}^2 + a^2 \cdot {x_k}^2 + (\overline{y} - a \cdot \overline{x})^2 + 2 \cdot a \cdot x_k \cdot (\overline{y} - a \cdot \overline{x}) - 2 \cdot y_k \cdot a \cdot x_k - 2 \cdot y_k \cdot (\overline{y} - a \cdot \overline{x})$ Gang parenteserne ud:

$S(a) = {y_k}^2 + a^2 \cdot {x_k}^2 + \overline{y}^2 + a^2 \cdot \overline{x}^2 - 2 \cdot \overline{y} \cdot a \cdot \overline{x} + 2 \cdot a \cdot x_k \cdot \overline{y} - 2 \cdot a^2 \cdot x_k \cdot \overline{x} - 2 \cdot y_k \cdot a \cdot x_k - 2 \cdot y_k \cdot \overline{y} + 2 \cdot y_k \cdot a \cdot \overline{x}$

Dette udtryk differentieres:

$\frac{d}{da}\left({y_k}^2 + a^2 \cdot {x_k}^2 + \overline{y}^2 + a^2 \cdot \overline{x}^2 - 2 \cdot \overline{y} \cdot a \cdot \overline{x} + 2 \cdot a \cdot x_k \cdot \overline{y} - 2 \cdot a^2 \cdot x_k \cdot \overline{x} - 2 \cdot y_k \cdot a \cdot x_k - 2 \cdot y_k \cdot \overline{y} + 2 \cdot y_k \cdot a \cdot \overline{x} \right)$

$= 2 \cdot a \cdot {x_k}^2 + 2 \cdot a \cdot \overline{x}^2 - 2 \cdot \overline{y} \cdot \overline{x} + 2 \cdot x_k \cdot \overline{y} - 4 \cdot a \cdot x_k \cdot \overline{x} - 2 \cdot y_k \cdot x_k + 2 \cdot y_k \cdot \overline{x}$ Saml *a*-leddene

$= 2 \cdot a \cdot ({x_k}^2 + \overline{x}^2 - 2 \cdot x_k \cdot \overline{x}) - 2 \cdot \overline{y} \cdot \overline{x} + 2 \cdot x_k \cdot \overline{y} - 2 \cdot y_k \cdot x_k + 2 \cdot y_k \cdot \overline{x}$ Anvend en kvadratsætning:

$= 2 \cdot a \cdot (x_k - \overline{x})^2 - 2 \cdot \overline{y} \cdot \overline{x} + 2 \cdot x_k \cdot \overline{y} - 2 \cdot y_k \cdot x_k + 2 \cdot y_k \cdot \overline{x}$ Sæt henh. $-2 \cdot \overline{y}$ og $+2 \cdot y_k$ uden for parentes

$= 2 \cdot a \cdot (x_k - \overline{x})^2 - 2 \cdot \overline{y} \cdot (\overline{x} - x_k) + 2 \cdot y_k \cdot (\overline{x} - x_k)$ Kontroller! – og sæt nu $(\overline{x} - x_k)$ udenfor

$= 2 \cdot a \cdot (x_k - \overline{x})^2 - 2 \cdot (\overline{x} - x_k) \cdot (\overline{y} - y_k)$

Det var det k'te led, og nu summerer vi og sætter lig med 0.

$\sum_{k=1}^{n} \left(2 \cdot a \cdot (x_k - \overline{x})^2 - 2 \cdot (\overline{x} - x_k) \cdot (\overline{y} - y_k) \right) = 0$ Summen splittes op og vi rykker over:

$\sum_{k=1}^{n} 2 \cdot a \cdot (x_k - \overline{x})^2 = \sum_{k=1}^{n} 2 \cdot (\overline{x} - x_k) \cdot (\overline{y} - y_k)$ Forkort med 2 og sæt *a* uden for:

$a \cdot \sum_{k=1}^{n} (x_k - \overline{x})^2 = \sum_{k=1}^{n} (\overline{x} - x_k) \cdot (\overline{y} - y_k)$ Isoler a:

$a = \frac{\sum_{k=1}^{n} (\overline{x} - x_k) \cdot (\overline{y} - y_k)}{\sum_{k=1}^{n} (x_k - \overline{x})^2}$ Det er formlen for *a* i sætningen

Øvelse 8.12 Udfør lineær regression ved at beregne koefficienterne
Anvend sætning 2 til at beregne formlen for den lineære regression på de 10 datapunkter, og sammenlign med resultatet i øvelse 8.11. Det er lettest at opstille beregningen i et regneark, hvor dataværdierne er skrevet ind i to kolonner.