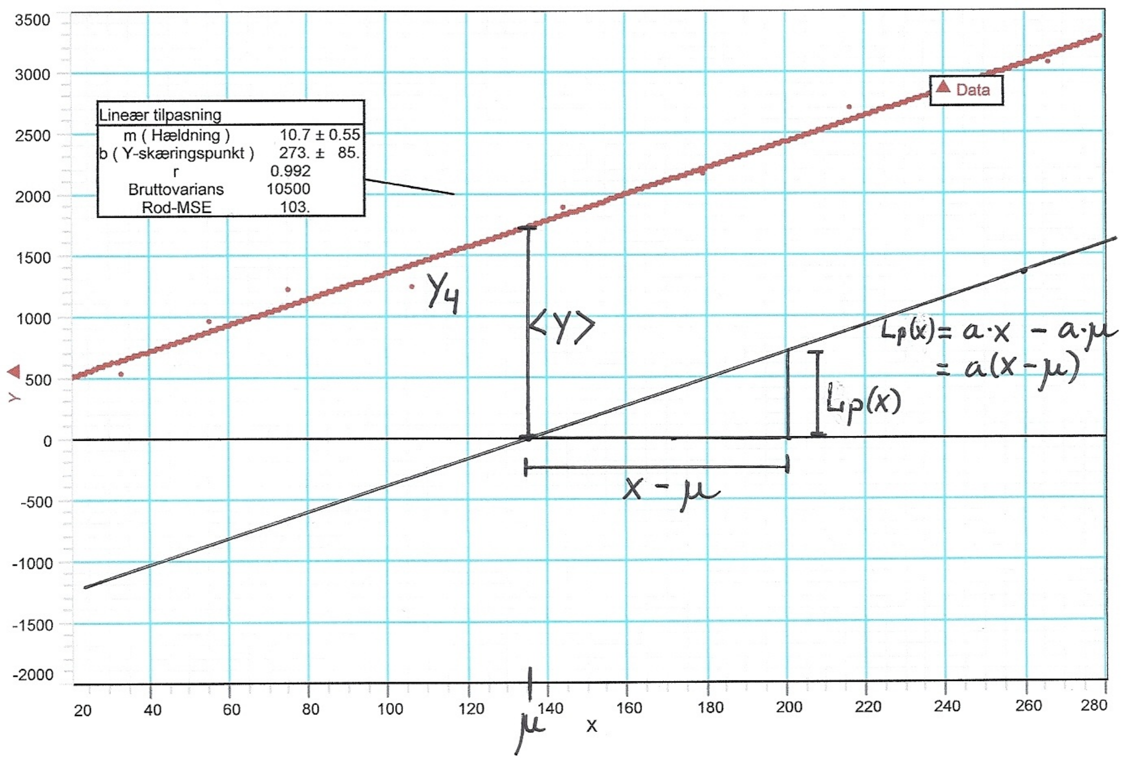
Mindste Kvadraters Rette Linje

Et nærmere kig på formlerne.

Jacob Nielsen[1]

Udseendet af formlen til beregning af hældningen af mindste kvadraters rette linje er ikke intuitivt klar. Uden at bevise formlen kan man dog komme et stykke i retning af en forståelse af dens gyldighed.



Figuren viser et datasæt og den tilhørende linje - $L_{MK}(x)$ bestemt ved mindste kvadraters metode. Datasættet kan ses nederst på næste side.

Desuden er linjen $L_p(x)$ indtegnet. Denne linje skærer x-aksen i (μ;0) og har samme hældning som $L_{MK}(x)$ - μ er middelværdien af datasættets x-værdier. Her bruges skrivemåderne:

$$\langle x \rangle = \mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

n er antallet af datapunkter.

[1] Datadrev\Matematik\Statistik og sandsynlighed\Mindste Kvadraters Rette Linje konstanterne 201009.wpd.

Indledningsvis mindes om, at hvis X er en stokastisk variable , og k er en konstat, så gælder følgende regler for beregning af middelværdier:

$$\langle k \cdot x \rangle = k \langle x \rangle$$
$$\langle x + k \rangle = \langle x \rangle + \langle k \rangle = \langle x \rangle + k$$
$$\langle x - \mu \rangle = \langle x \rangle - \langle \mu \rangle = \mu - \mu = 0$$

Med udgangspunkt i formlen til beregning af konstanten b hørende til $L_{MK}(x)$ fås:

$$b = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mu \cdot a) = \langle y_i - \mu \cdot a \rangle = \langle y_i \rangle - \langle \mu \cdot a \rangle = \langle y_i \rangle - \mu \cdot a$$
$$\Updownarrow$$
$$\langle y_i \rangle = \mu \cdot a + b$$

Den sidste linje udtrykket blot, at punktet (µ ; <y>) ligger på til $L_{MK}(x)$, hvilket jo lyder rimeligt.

Formlen til beregning af a er ikke helt så let at gennemskue, men vi kan komme et stykke hen ad vejen.
Ud fra figuren på forgående side får vi:

$$L_p(x) = L_{MK}(x) - \langle y \rangle$$
$$a = \frac{L_p(x)}{x - \mu} = \frac{L_p(x_i)}{x_i - \mu} = \frac{L_{MK}(x_i) - \langle y \rangle}{x_i - \mu} \approx \frac{y_i - \langle y \rangle}{x_i - \mu} =$$
$$\frac{(x_i - \mu)\cdot(y_i - \langle y \rangle)}{(x_i - \mu)^2} = \frac{(x_i - \mu)\cdot y_i}{(x_i - \mu)^2} - \frac{(x_i - \mu)\cdot \langle y \rangle}{(x_i - \mu)^2}$$

Første linje kommer af, at punktet (µ ; <y>) antages at ligger på til $L_{MK}(x)$. Det krøllede lighedstegn gælder, fordi målepunkterne antages at ligge tæt på $L_{MK}(x)$.

Redigerbare data, Data

X	Y
34.00	536.00
56.00	956.00
76.00	1209.00
107.00	1234.00
145.00	1875.00
217.00	2687.00
267.00	3056.00
180.00	2154.00

$$\mu = 135$$
$$\langle y \rangle = 1713$$

Ved at benytte de resultater, som vi allerede har, samt en ny antagelse, kan vi komme frem til formlen til beregning af a. Den ekstra antagelse er, at vi får en god tilnærmelse til middelværdien af en brøk - ved at tage middelværdi af tæller og nævner hver for sig. Dette gælder ikke i alle tilfælde.

$$\left\langle (x_i - \mu) \cdot \langle y \rangle \right\rangle = \langle x_i - \mu \rangle \cdot \langle y \rangle = 0$$

$$a_i \approx \frac{(x_i - \mu) \cdot y_i}{(x_i - \mu)^2} - \frac{(x_i - \mu) \cdot \langle y \rangle}{(x_i - \mu)^2}$$

$$\frac{\left\langle (x_i - \mu) \cdot y_i \right\rangle}{\left\langle (x_i - \mu)^2 \right\rangle} - \frac{\left\langle (x_i - \mu) \cdot \langle y \rangle \right\rangle}{\left\langle (x_i - \mu)^2 \right\rangle} = \frac{\left\langle (x_i - \mu) \cdot y_i \right\rangle}{\left\langle (x_i - \mu)^2 \right\rangle} - 0 =$$

$$\frac{\frac{1}{n} \sum_{i=1}^{i=n} (x_i - \mu) \cdot y_i}{\frac{1}{n} \sum_{i=1}^{i=n} (x_i - \mu)^2} = \frac{\sum_{i=1}^{i=n} (x_i - \mu) \cdot y_i}{\sum_{i=1}^{i=n} (x_i - \mu)^2} = a$$

Til sidst skal det understreges, at der ikke er ført bevis for formlerne til bestemmelse af parametrene i $L_{MK}(x)$. Men forhåbentlig har læseren fået en dybere forståelse af beregningerne.

Traditionelt bevises formlerne ved brug af differentialregning. Beviset kan findes i mange standardværker om statistik. Skal man også bevise formlerne for usikkerheden på a og b kræver det omfattende baggrunds viden om statistik.