Lineær regression

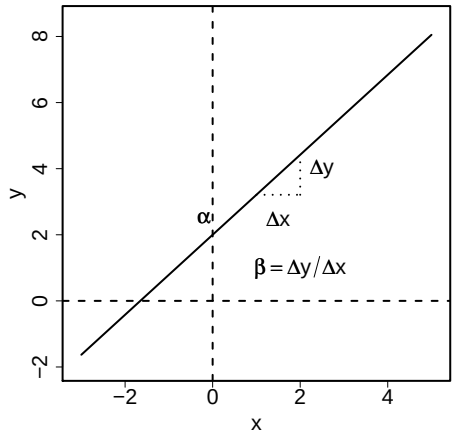
Claus Ekstrøm
E-mail: ekstrom@life.ku.dk

Program

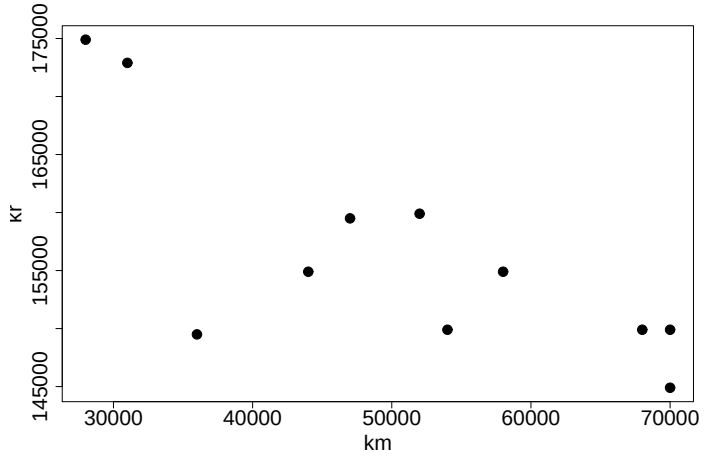
- Den rette linje
- Tilpasning af ret linje til data
  - Mindste kvadraters metode
- Modelkontrol
- Korrelationskoefficienten

Den rette linjes ligning

$$y = \alpha + \beta \cdot x$$

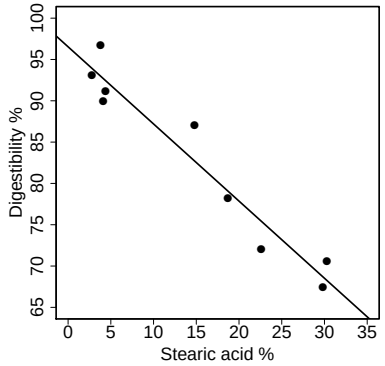


Eksempel — Claus' bil



Eksempel — Fordøjelighed

% stearinsyre	% fordøjelighed
29.8	67.5
30.3	70.6
22.6	72.0
18.7	78.2
14.8	87.0
4.1	89.9
4.4	91.2
2.8	93.1
3.8	96.7



Residualer

Antag, at vi har en linje med parameterestimater $\hat{\alpha}$ og $\hat{\beta}$. Så er modellen

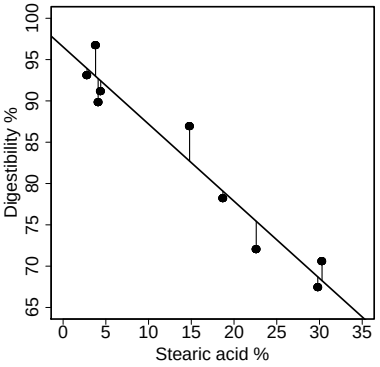
$$y = \hat{\alpha} + \hat{\beta} \cdot x$$

og modellen (linjen) fortæller, hvad vi vil forvente til en given x-værdi:

$$\hat{y}_i = \hat{\alpha} + \hat{\beta} \cdot x_i$$

Residualer er den afstand modellen "skyder forkert":

$$r_i = y_i - \hat{y}_i$$



Hvordan finder vi den bedste rette linje?

Vi har følgende krav:

- Residualerne er små
- Summen af residualerne skal være 0

Men! At summen skal være 0 er ikke nok.

Kvadrerede residualer

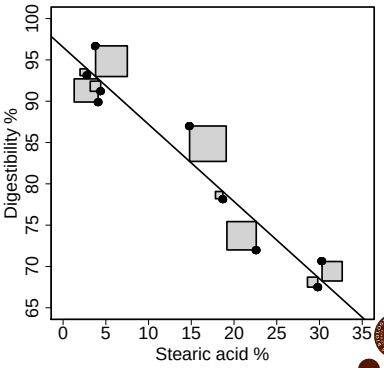
Summen af residualerne skal være 0:

$$\sum_i r_i = 0$$

Hvis vi *kvadrerer* residualerne vil en negativ værdi ikke kunne opveje en positiv værdi.

$$\sum_i r_i^2 = \sum_i (y_i - \alpha - \beta \cdot x_i)^2$$

Vi vil gerne have, at vores model skal være så tæt på de observerede data som muligt. Skal derfor finde α og β således at residualkvadratsummen minimeres!



Tilpasning af rette linjer

Bedste rette linje

Den "bedste" rette linje er den linje, der minimerer kvadratafvigelsessummen.

Estimater:

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \quad (1)$$

$$\hat{\alpha} = \bar{y}-\hat{\beta}\cdot\bar{x}. \quad (2)$$

Eksempel

i	x	y	$(x_i-\bar{x})$	$(y_i-\bar{y})$	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$	$(x_i-\bar{x})(y_i-\bar{y})$
1	29.8	67.5	15.211	-15.411	231.378	237.502	-234.420
2	30.3	70.6	15.711	-12.311	246.839	151.563	-193.421
3	22.6	72.0	8.011	-10.911	64.178	119.052	-87.410
4	18.7	78.2	4.111	-4.711	16.901	22.195	-19.368
5	14.8	87.0	0.211	4.089	0.045	16.719	0.863
6	4.1	89.9	-10.489	6.989	110.017	48.845	-73.306
7	4.4	91.2	-10.189	8.289	103.813	68.706	-84.455
8	2.8	93.1	-11.789	10.189	138.978	103.813	-120.116
9	3.8	96.7	-10.789	13.789	116.400	190.133	-148.767
Sum	131.3	746.2	0.000	0.000	1028.549	958.529	-960.399

Mindste kvadraters metode

Find α og β så kvadratsummen minimeres

$$\sum_i r_i^2 = \sum_i (y_i-\alpha-\beta\cdot x_i)^2$$

Kan findes ved at løse ligningssystemet

$$\frac{\partial f}{\partial \alpha} = \sum_{i=1}^{n}\frac{\partial}{\partial\alpha}(y_i-\alpha-\beta\cdot x_i)^2 = \sum_{i=1}^{n}2(y_i-\alpha-\beta\cdot x_i)\cdot(-1)$$

$$= -2\cdot(y_\bullet - n\alpha - \beta x_\bullet) = 0 \quad (3)$$

$$\frac{\partial f}{\partial \beta} = \sum_{i=1}^{n}2(y_i-\alpha-\beta\cdot x_i)\cdot(-x_i) = 0 \quad (4)$$

Modelkontrol

- Kvantitative variable
- Lineær sammenhæng?
- Indflydelsesrige observationer
- x på y eller y på x.
- Ekstra- og interpolation.

Transformationer

Andemad

Days	Leaves	Days	Leaves
0	100	7	918
1	127	8	1406
2	171	9	2150
3	233	10	2800
4	323	11	4140
5	452	12	5760
6	654	13	8250

Hvis data har en anden struktur end en ret linje giver det ikke mening at lave lineær regression. Kan dog nogle gange fikse det, hvis man kan transformere data så de kommer på lineær form. Exponential growth model for populationsstørrelse til tid t:

$$f(t) = c \cdot \exp(b \cdot t).$$

Tag logaritmer på begge sider

$$\log(f(t)) = \underbrace{\log c}_{\alpha} + \underbrace{b}_{\beta} \cdot t.$$

Korrelationskoefficienten

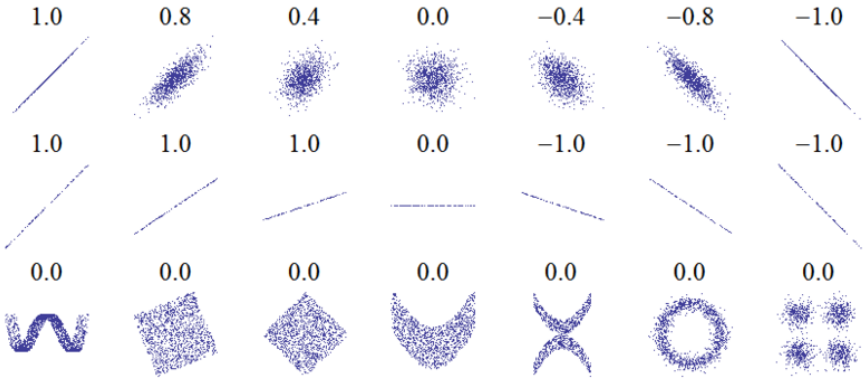
Korrelationskoefficienten beskriver den *lineære* sammenhæng mellem X og Y og er givet ved

$$\hat{\rho} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(\sum_i (x_i - \bar{x})^2)(\sum_i (y_i - \bar{y})^2)}}.$$

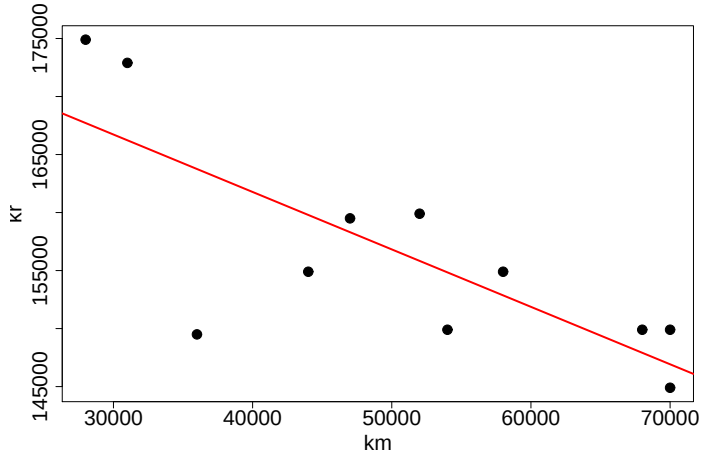
Korrelationskoefficienten kan fortolkes som den lineære hældning man vil opnå, hvis man "standardiserer" X og Y. Korrelationskoefficienten er

- 0 hvis der ikke er nogen information omkring Y i X
- 1 hvis observationerne ligger perfekt på en linje med positiv hældning
- -1 hvis observationerne ligger perfekt på en linje med negativ hældning

Korrelationskoefficienten — eksempel



Eksempel — Claus' bil



Dagens hovedpunkter

- Lineær regression
  - Fortolkning, modellering, fortolkning af parametre, modelkontrol
- Residualer
- Mindste kvadraters metode
- Korrelationskoefficienten
  - Fortolkning, anvendelighed
- Transformation af data