Lineær regression og korrelation

Christian Damsgaard Jørgensen

Institut for Matematik og Datalogi
Syddansk Universitet, Odense

Uge 49

Introduktion

Vi vil analysere forholdet mellem to kvantitative variable, X og Y.

Hvad er lineær regressions- og korrelationsanalyse?

Teknikker baseret på tilpasning af en ret linje til data.

Data

Består af observationspar (X, Y).

To eksempler på data:

1. Amfetamin og fortæring af mad
2. Arsen i ris

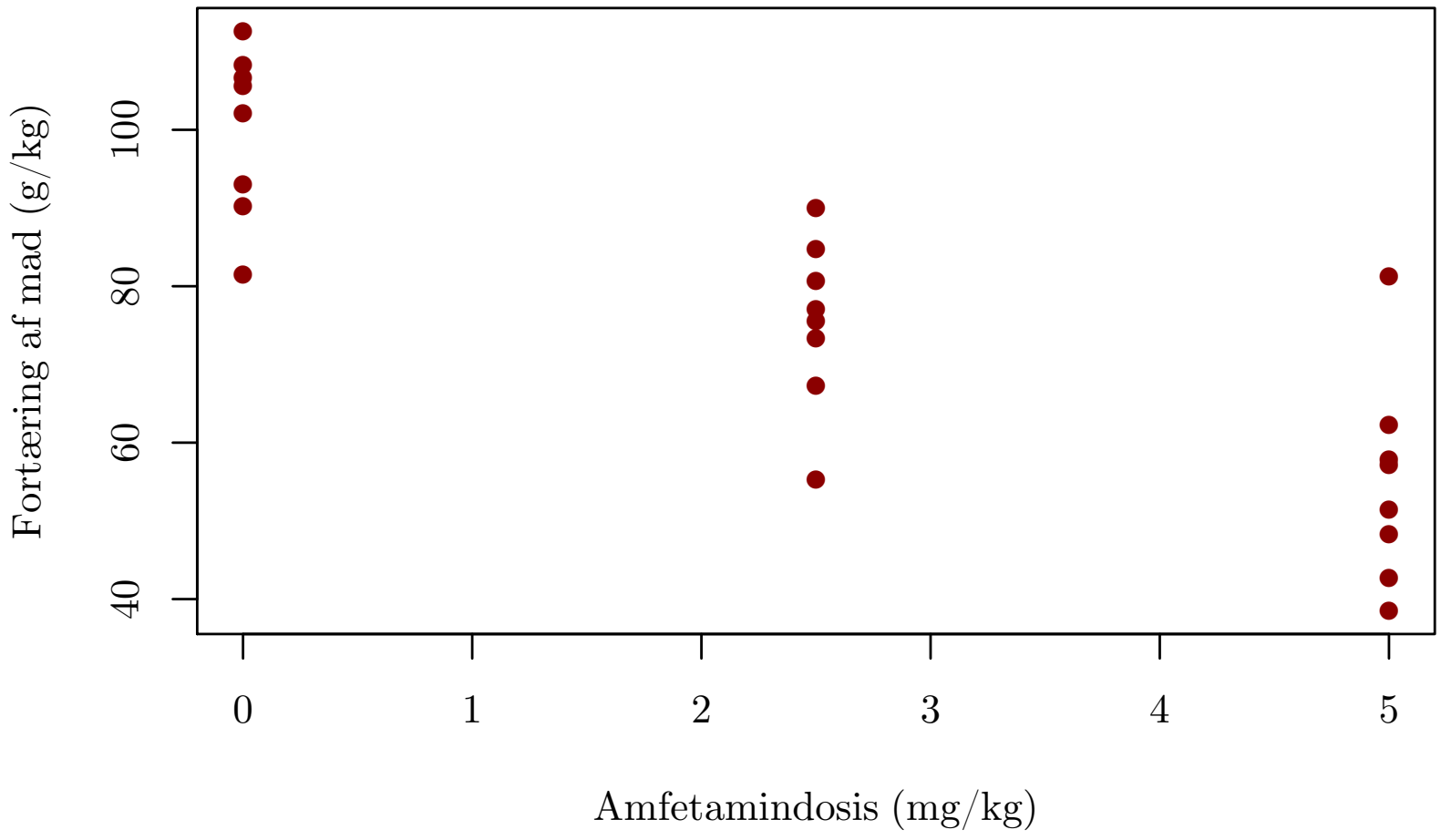
Eksempel 1: Amfetamin og fortæring af mad

```
# Indlæs data i R
data <- read.table("amphetamine.txt", header = TRUE)
# Se på de første seks observationer
head(data)
```

Output:

```
  FoodConsumption Dose
1           112.6    0
2           102.1    0
3            90.2    0
4            81.5    0
5           105.6    0
6            93.0    0
```

Eksempel 1: Amfetamin og fortæring af mad



Figur 1: Scatterplot af fortæring af mad mod amfetamindosis.

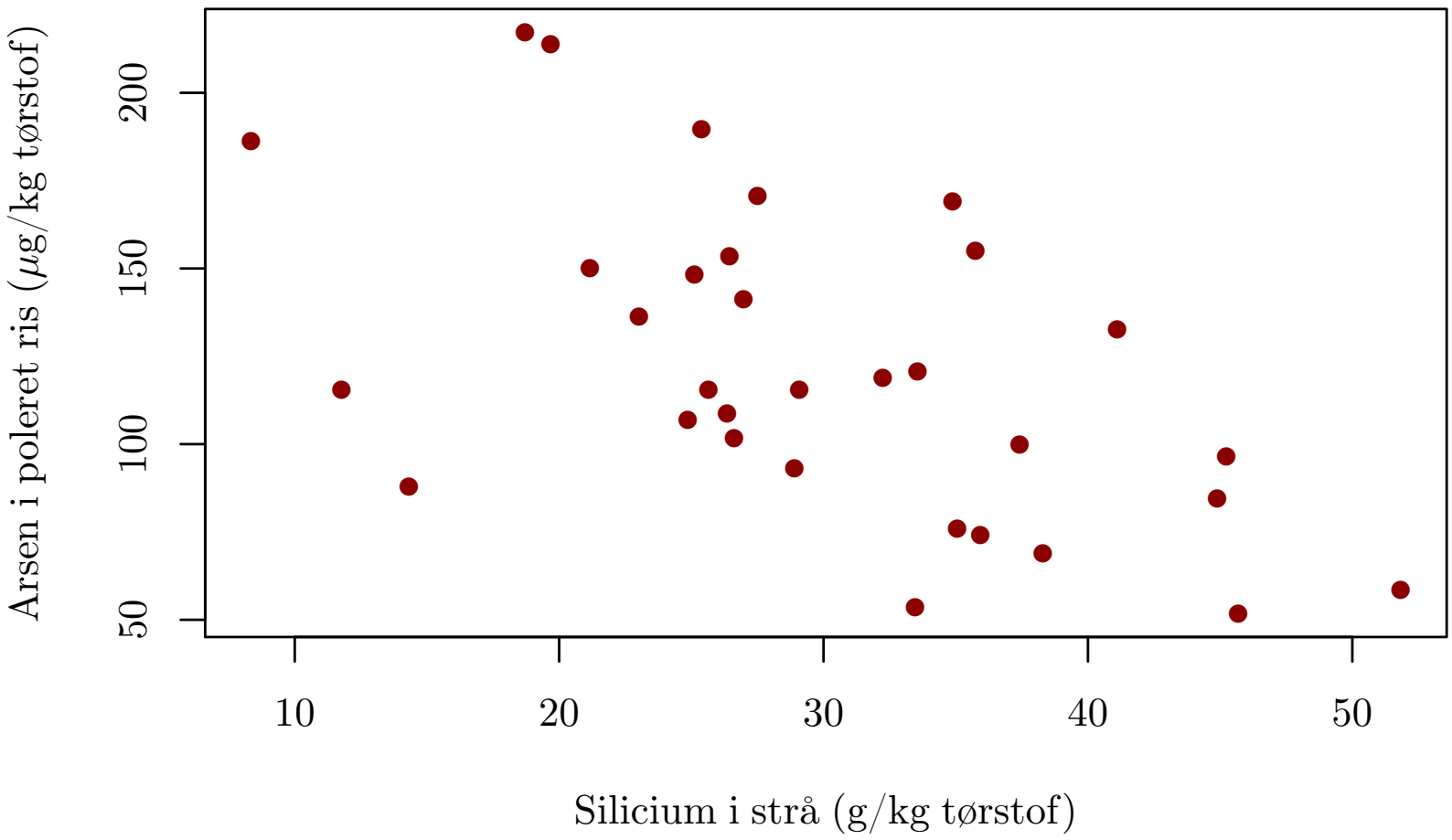
Eksempel 2: Arsen i ris

```
# Indlæs data i R
data <- read.table("rice.txt", header = TRUE)
# Se på de første seks observationer
head(data)
```

Output:

```
  StrawSi RiceAs
1    8.35 186.21
2   11.77 115.52
3   14.32  87.93
4   18.71 217.24
5   19.68 213.79
6   21.17 150.00
```

Eksempel 2: Arsen i ris



Figur 2: Scatterplot af arsenkonc. i ris mod siliciumkonc. i strå.

Korrelationskoefficienten

Antag nu, at vi har en stikprøve bestående af n observationspar, hvor hvert par repræsenterer målingerne af to variable, X og Y.

Styrken af den lineære sammenhæng

Hvis et scatterplot af Y versus X viser en generel lineær tendens, så vil det være oplagt at beskrive *styrken* af den lineære sammenhæng.

Korrelationskoefficienten

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x - \overline{x}}{s_x} \right) \left(\frac{y - \overline{y}}{s_y} \right)$$

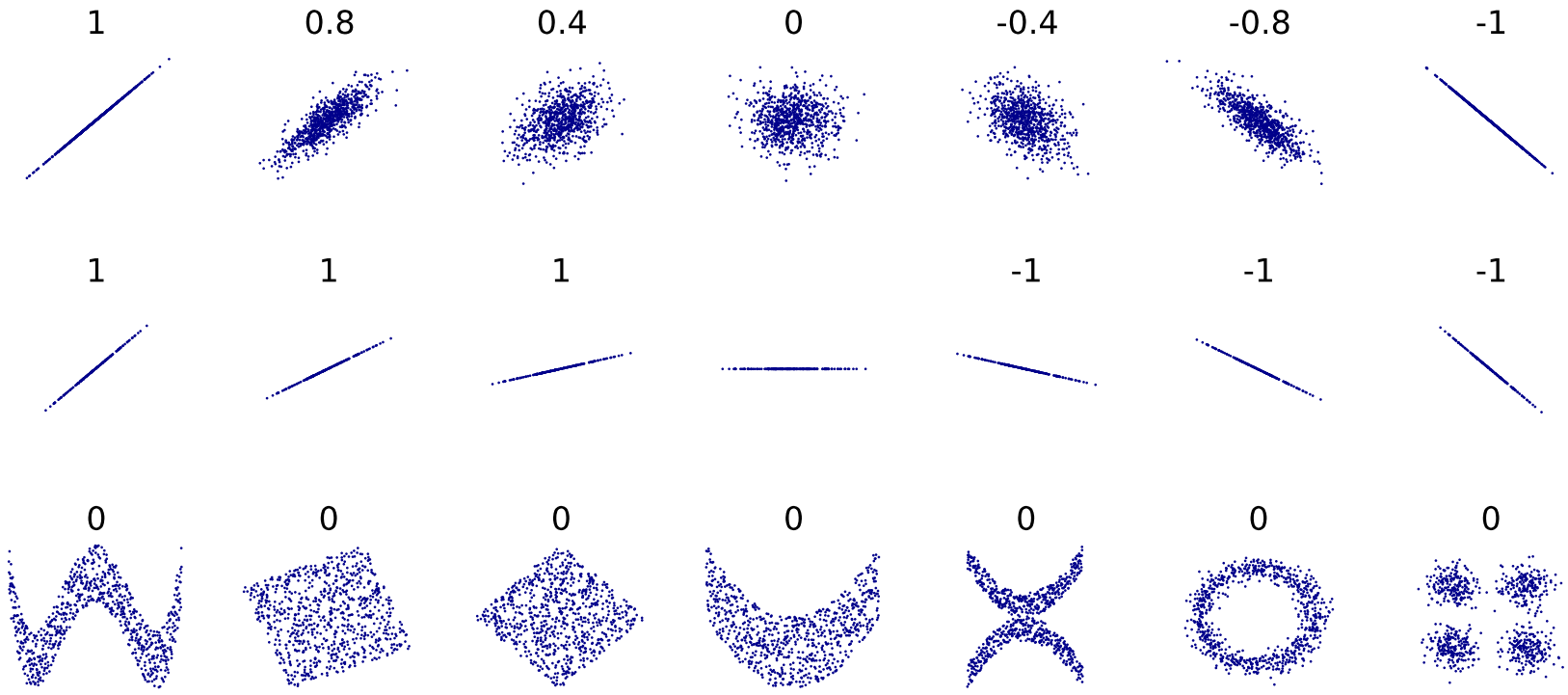
hvor

$$s_x = \sqrt{\frac{\sum_{i=1}^{n} (x - \overline{x})^2}{n-1}} \quad \text{og} \quad s_y = \sqrt{\frac{\sum_{i=1}^{n} (y - \overline{y})^2}{n-1}}$$

Ved indsættelse af udtrykkene for s_x og s_y kan dette omskrives til

$$r = \frac{\sum_{i=1}^{n} (x - \overline{x})(y - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x - \overline{x})^2 \sum_{i=1}^{n} (y - \overline{y})^2}}$$

Korrelationskoefficienten



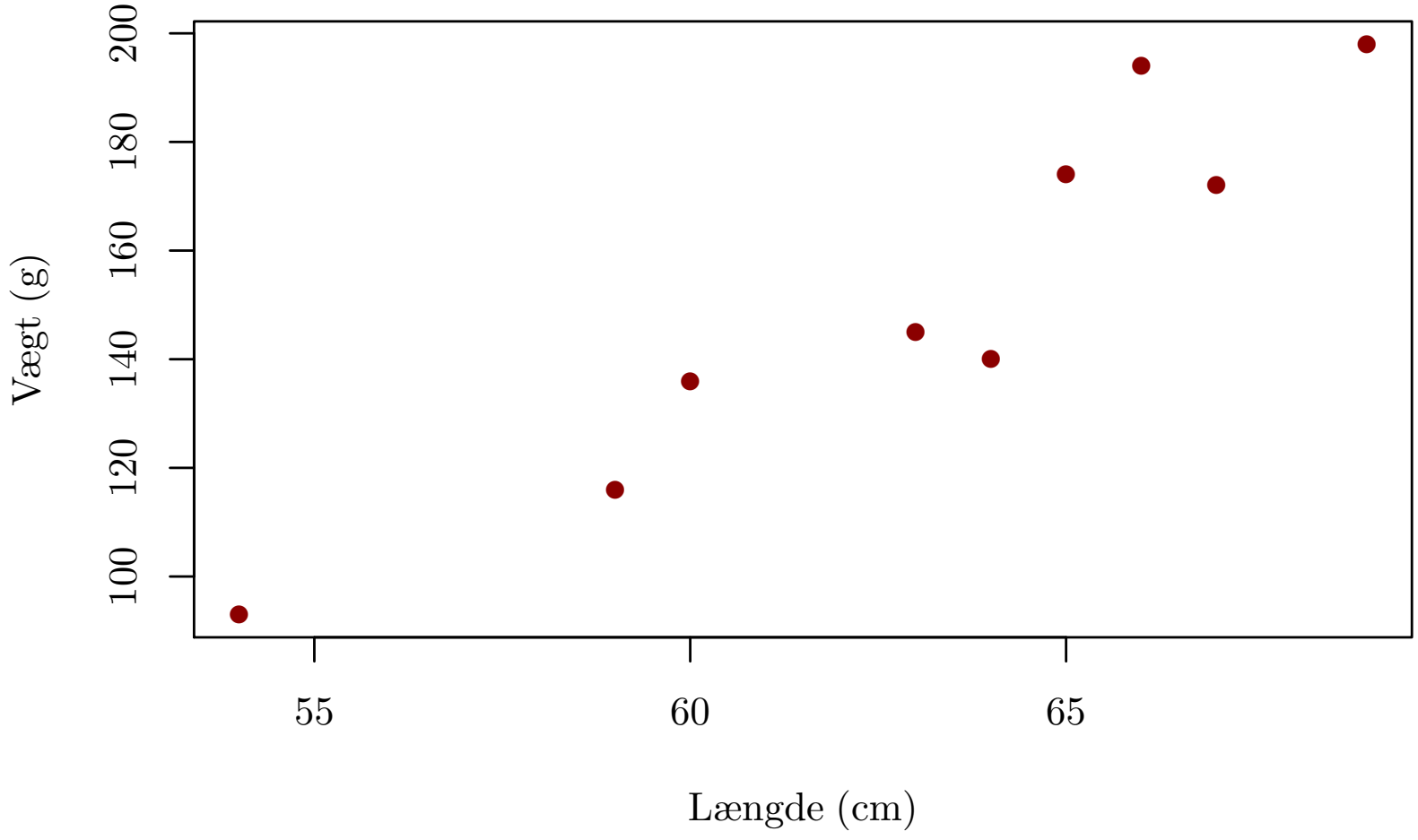
Eksempel 3: Slangers længde og vægt

```
# Indlæs data i R
data <- read.table("snakes.txt", header = TRUE)
# Se på de første seks observationer
head(data)
```

Output:

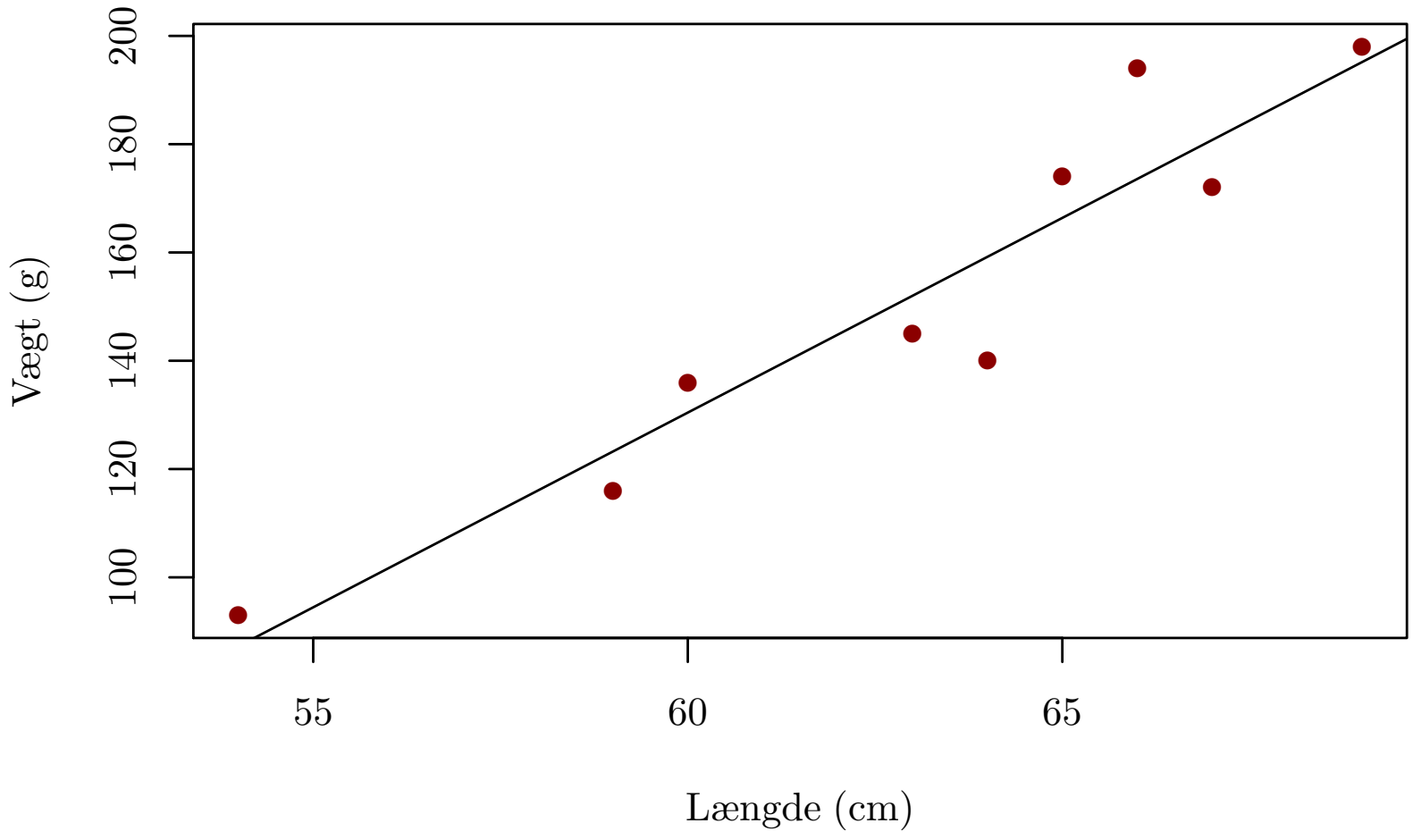
```
  Weight Length
1    136     60
2    198     69
3    194     66
4    140     64
5     93     54
6    172     67
```

Eksempel 3: Slangers længde og vægt



Figur 3: Kropslængde og vægt for ni slanger

Eksempel 3: Slangers længde og vægt



Figur 3: Kropslængde og vægt for ni slanger med regressionslinje.

Eksempel 3: Slangers længde og vægt

	X	Y	$z_x = \frac{x-\overline{x}}{s_x}$	$z_y = \frac{y-\overline{y}}{s_y}$	$z_x z_y$
	60	136	$-0{,}65\ldots$	$-0{,}45\ldots$	$0{,}29\ldots$
	69	198	$1{,}29\ldots$	$1{,}30\ldots$	$1{,}68\ldots$
	66	194	$0{,}65\ldots$	$1{,}19\ldots$	$0{,}77\ldots$
	64	140	$0{,}22\ldots$	$-0{,}34\ldots$	$-0{,}07\ldots$
	54	93	$-1{,}94\ldots$	$-1{,}67\ldots$	$3{,}24\ldots$
	67	172	$0{,}86\ldots$	$0{,}57\ldots$	$0{,}49\ldots$
	59	116	$-0{,}86\ldots$	$1{,}02\ldots$	$0{,}88\ldots$
	65	174	$0{,}43\ldots$	$0{,}62\ldots$	$0{,}27\ldots$
	63	145	$0{,}00\ldots$	$-0{,}20\ldots$	$0{,}00\ldots$
sum	567	1368	0,00	0,00	7,55
gns.	63,00	152,00	0,00	0,00	
std.afv.	4,64	35,34	1,00	1,00	

Tabel 1: Standardiseret vægt, længde og deres produkter.

Eksempel 3: Slangers længde og vægt

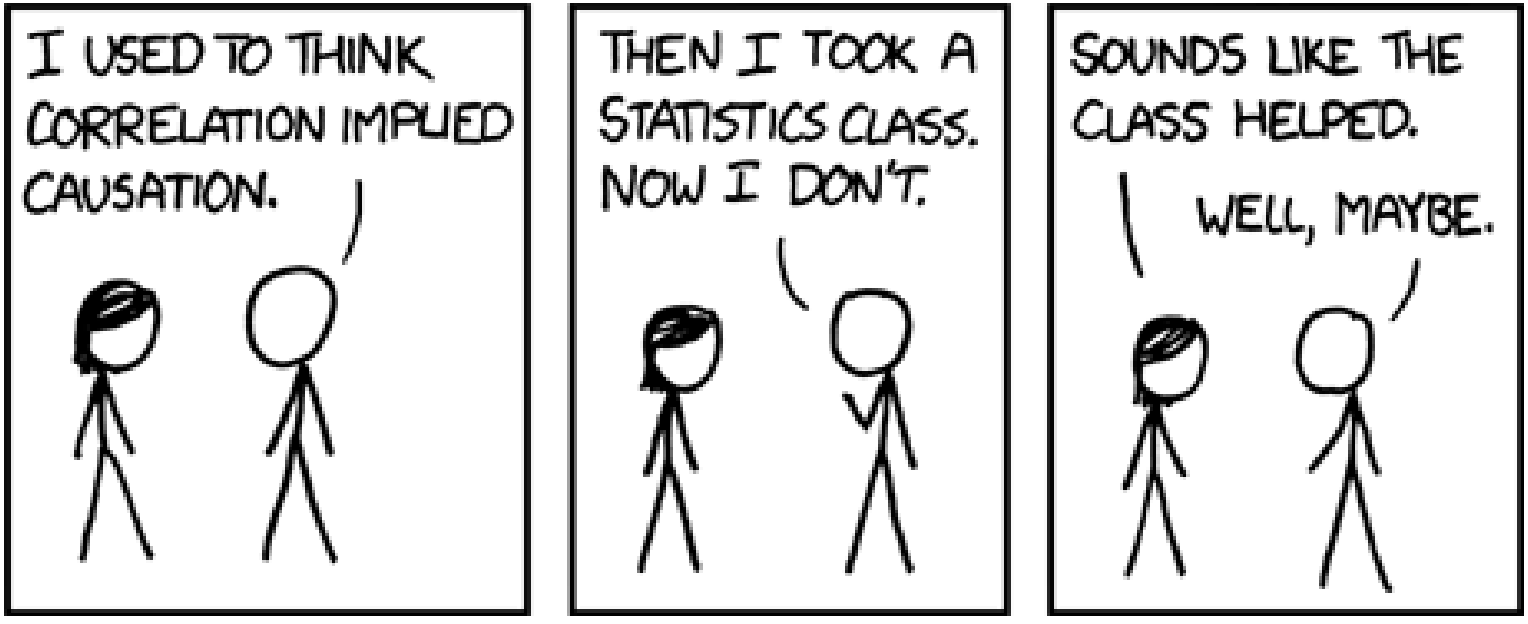
Korrelationskoefficienten (stikprøvekorrelationen) er omkring 0,94:

$$r = \frac{1}{9-1} \cdot 7{,}55 \approx 0{,}94$$

Korrelation og årsagssammenhæng

Korrelation medfører ikke årsagssammenhæng (kausalitet):

En observeret sammenhæng mellem to variable indikerer ikke nødvendigvis en årsagssammenhæng (kausalitet) mellem dem.



Kilde: xkcd.com/552

Se eksempler:
`http://www.tylervigen.com/spurious-correlations`

Linjens ligning

Ligningen for en ret linje kan skrives som

$$Y = b_0 + b_1 X$$

hvor b_0 er skæringen med y-aksen og b_1 er linjens hældning.

Hældningen b_1 er ændringsraten for Y med hensyn til X.

Den tilpassede regressionslinje for Y på X skrives

$$\widehat{y} = b_0 + b_1 x$$

da linjen kun giver estimerede eller prædikterede værdier.

$$b_1 = r\left(\frac{s_y}{s_x}\right) \quad \text{og} \quad b_0 = \overline{y} - b_1 \overline{x}$$

Residualer

For hver observation x_i er der en prædikteret Y-værdi

$$\widehat{y}_i = b_0 + b_1 x_i$$

Et *residual* e_i er forbundet med hvert observeret par (x_i, y_i):

$$e_i = y_i - \widehat{y}_i$$

Den residuale kvadratsum

$$\text{SS(resid)} = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$$

Mindste kvadrater-kriteriet

Den bedste rette linje minimerer den residuale kvadratsum.

Eksempel 2: Arsen i ris

Obs. #	x	y	$\widehat{y}$	$y - \widehat{y}$	$(y - \widehat{y})^2$
1	8,3	186,2	176,2...	10,0...	99,50...
2	11,8	115,5	167,6...	−52,1...	2716,00...
3	14,3	87,9	161,2...	−73,3...	5373,93...
4	18,7	217,2	150,2...	67,0...	4492,74...
5	19,7	213,8	147,8...	66,0...	4356,67...
⋮	⋮	⋮	⋮	⋮	⋮
28	41,1	132,8	94,0...	38,8...	1503,19...
29	45,2	96,6	83,6...	12,9...	167,11...
30	44,9	84,5	84,5...	0,0...	0,00...
31	45,7	51,7	82,5...	−30,8...	948,51...
32	51,8	58,6	67,1...	−8,5...	71,69...
Sum				0,0	41727,11

Tabel 2: Beregning af SS(resid).

Den residuale standardafvigelse

Hvor langt over eller under regressionslinjen ligger punkterne typisk?

$$s_e = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}} = \sqrt{\frac{\text{SS(resid)}}{n-2}}$$

En slags mål for vertikal afstand fra datapunkter til regressionslinje.

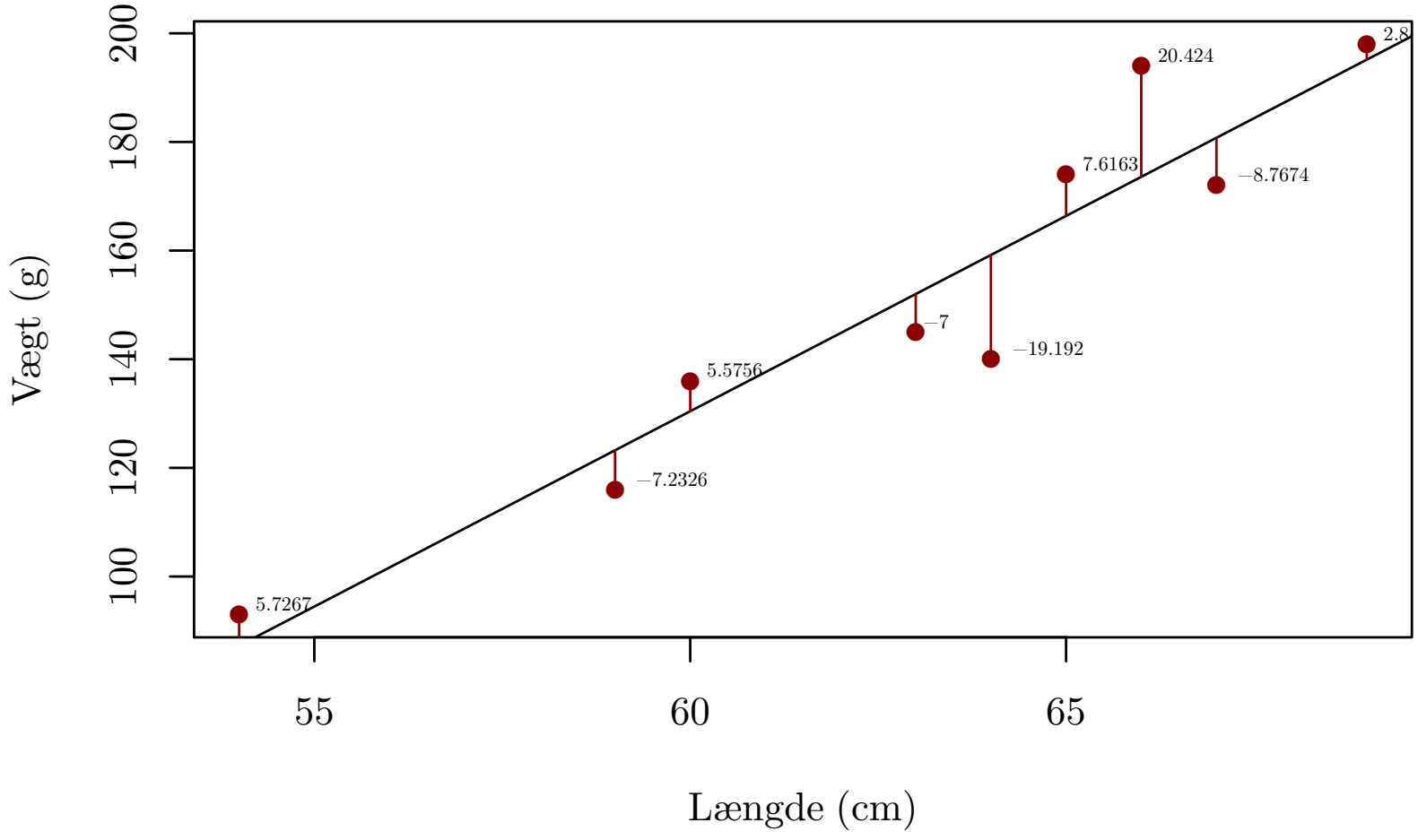
- Udledt fra residual kvadratsum
- Lettere at fortolke end residual kvadratsum

Beregning for eksempel 2 (arsen i ris) ved brug af tabel 2:

$$s_e = \sqrt{\frac{41727{,}11}{32-2}} = \sqrt{1390{,}90} = 37{,}30 \ \mu\text{g/kg}$$

Prædiktioner afviger med omkring 37,30 µg/kg i gennemsnit.

Eksempel 3: Slangers længde og vægt



Figur 4: Vægt versus længde med residualerne og et linjesegment (rødt), som betegner størrelsen af den residuale standardafvigelse.

Tilnærmet sammenhæng

Korrelationskoefficienten opfylder denne tilnærmede sammenhæng:

$$r^2 \approx \frac{s_y^2 - s_e^2}{s_y^2} = 1 - \frac{s_e^2}{s_y^2}$$

Litteratur


Samuels, Witmer, Schaffner.
Statistics for the Life Sciences (4. udgave).
Pearson Education, 2012.


Dalgaard.
Introductory Statistics with R (2. udgave).
Springer-Verlag New York, 2008.

Denne præsentation er baseret på eksempler fra Samuels et al.