

Voortraject Statistiek

Prof. dr. Ellen Vandervieren

ellen.vandervieren@uantwerpen.be
CST, Venusstraat 35, bureau 107



Bivariate lineaire regressie

Waarneming	Gem temp (°C)	Aantal verkochte dozen ijs
1	17	45
2	22	55
3	24	63
4	28	80
5	19	51
6	20	50
7	19	45
8	22	59
9	23	61
10	26	71

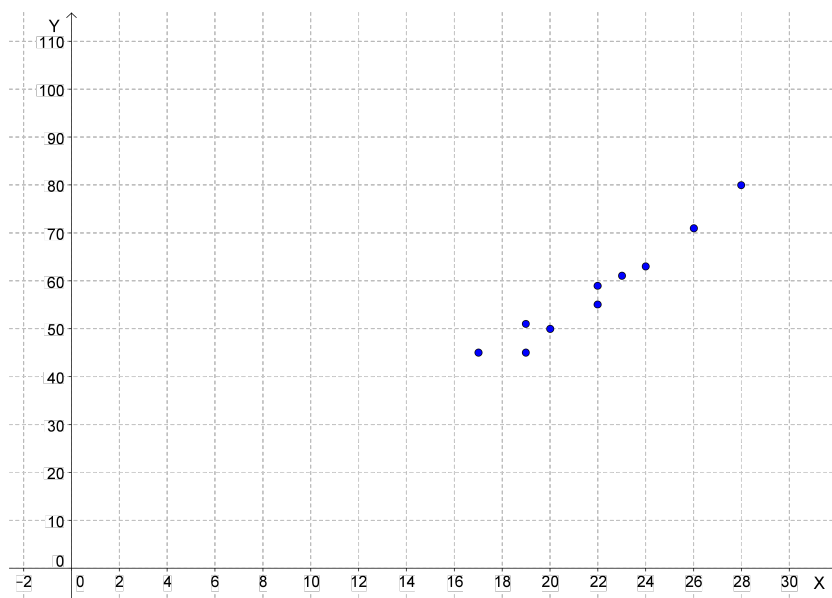
- ▶ Verband tussen temperatuur en aantal verkochte dozen ijs?
- ▶ Vermoeden: hoe warmer, hoe meer ijs er wordt verkocht
- ▶ Gem temperatuur = X 'onafhankelijke variabele'
- ▶ Aantal verk. dozen ijs = Y 'afhankelijke variabele'

⇒ Puntenwolk tekenen

1 / 21



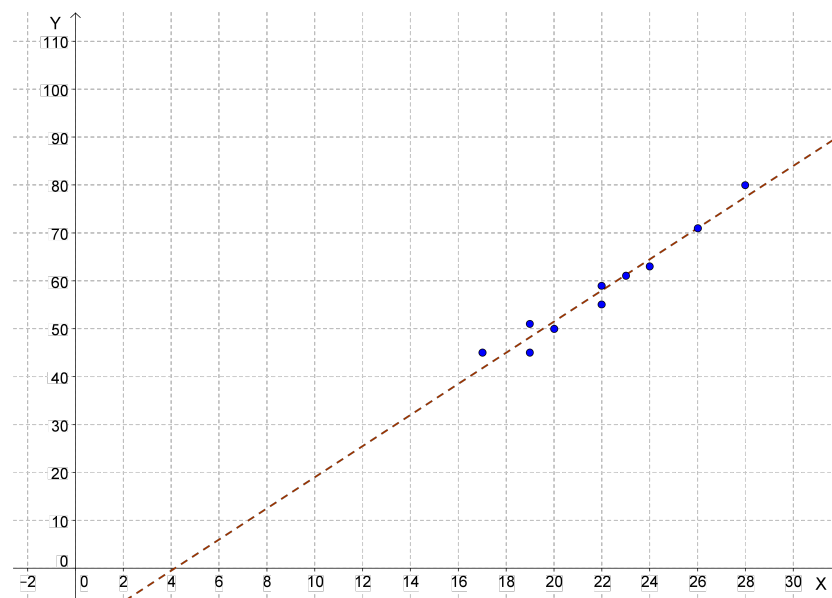
Spreidingsdiagram - Scatterplot



3 / 21



Spreidingsdiagram - Scatterplot



4 / 21



Spreidingsdiagram - Scatterplot

Conclusie:

- ▶ puntenwolk georiënteerd van linksonder naar rechtsboven
- ▶ hoe warmer, hoe meer dozen ijs worden verkocht
- ▶ punten liggen rond een stijgende rechte
- ▶ 'regressielijn':

$$E(y) = \beta_0 + \beta_1 x$$

OF

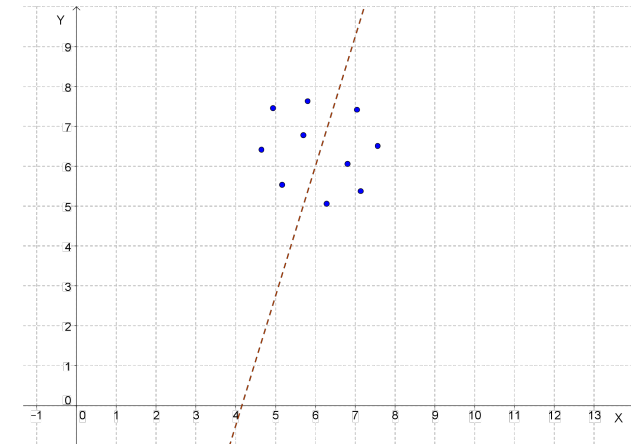
Geschatte aantal verkochte dozen ijs
 $= \beta_0 + \beta_1 \cdot \text{gem temp}$

- ▶ 'enkelvoudige lineaire regressie'

5 / 21



Wanneer lineaire regressie?

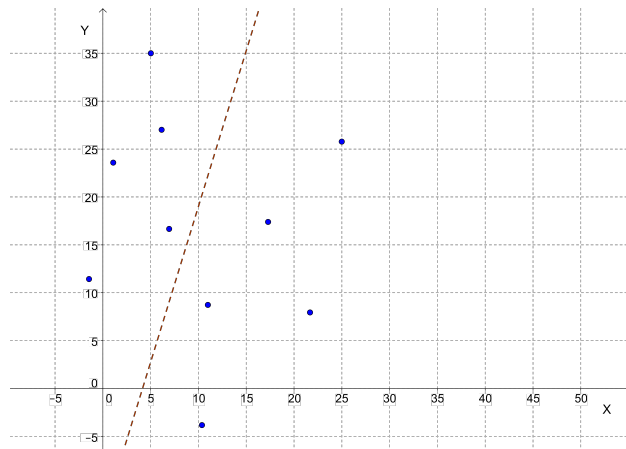


Geen lineair verband tussen X en Y
 $\Rightarrow$ lineaire regressie niet zinvol

6 / 21



Wanneer lineaire regressie?

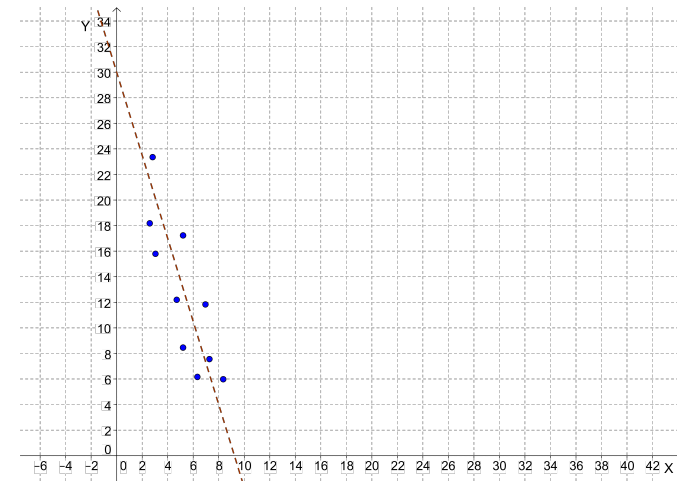


Geen sterk lineair verband tussen X en Y
 $\Rightarrow$ lineaire regressie niet zinvol

7 / 21



Wanneer lineaire regressie?



Sterk negatief lineair verband tussen X en Y
 $\Rightarrow$ lineaire regressie wel zinvol

8 / 21



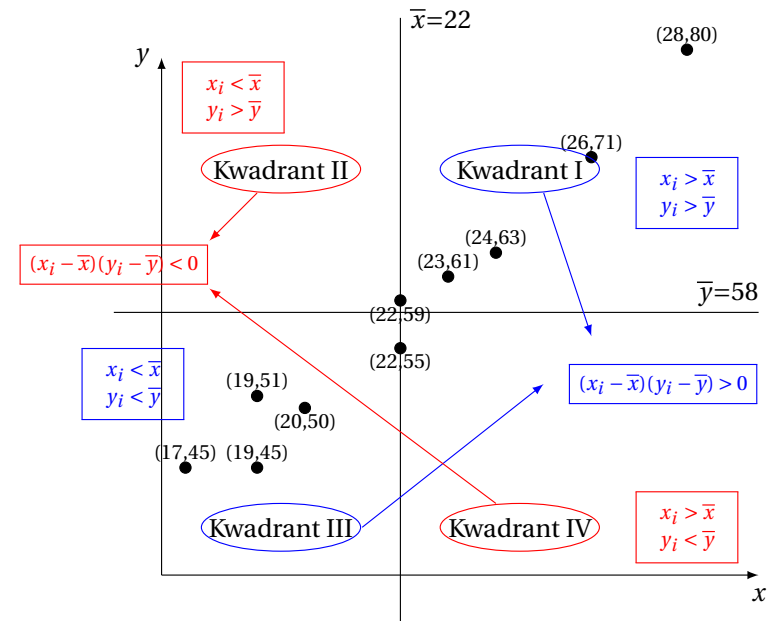
Wanneer lineaire regressie?

- ▶ Enkelvoudige lineaire regressie: indien er een sterk linear verband is tussen X en Y
 - ▶ punten volgen een stijgende rechte: **positief lineair verband**
 - ▶ punten volgen een dalende rechte: **negatief lineair verband**
- ▶ Maat nodig om de sterkte (en zin) van het lineaire verband tussen X en Y te meten: **correlatiecoëfficiënt**
- ▶ Formule correlatiecoëfficiënt opbouwen a.h.v. covariantie

9 / 21



Steekproefcovariantie



10 / 21



Steekproefcovariantie

- ▶ **Steekproefcovariantie** van waarnemingen $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$:

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

- ▶ **Interpretatie van steekproefcovariantie:**
 - ▶ $s_{xy} > 0$: meeste waarnemingen in kwadrant I en III → **positief lineair verband** tussen X en Y
 - ▶ $s_{xy} < 0$: meeste waarnemingen in kwadrant II en IV → **negatief lineair verband** tussen X en Y
 - ▶ $s_{xy} = 0$: waarnemingen liggen verspreid over alle kwadranten → **geen lineair verband** tussen X en Y

11 / 21



Steekproefcovariantie

- ▶ vb. 'Gem Temp' en 'Aantal verkochte dozen ijs':

$$\begin{aligned} s_{xy} &= \frac{1}{10-1} \sum_{i=1}^{10} (x_i - 22)(y_i - 58) \\ &= \frac{1}{9} [(17-22) \cdot (45-58) + \dots + (26-22) \cdot (71-58)] \\ &= \frac{338}{9} \\ &\approx 37.56^{\circ}\text{C} \rightarrow s_{xy} > 0: \text{positief lineair verband} \end{aligned}$$

- ▶ Hoe sterk is dit positief lineair verband?

12 / 21



Covariantie: nadeel

- ▶ Covariantie is afhankelijk van de gebruikte meeteenheden voor X en Y :
 - ▶ vb. 'Gem Temp' (°C) en 'Aantal verkochte dozen ijs': $s_{xy} = 37.56$ °C
 - ▶ vb. 'Gem Temp' (°C) en 'Aantal verkochte ijsjes' (elke doos bevat 10 ijsjes): $s_{xy} = 375.56$ °C
- ⇒ moeilijk te zeggen wanneer we een 'sterk' lineair verband hebben
- ▶ **Oplossing: covariantie dimensieloos maken door te standaardiseren: correlatiecoëfficiënt**

$$r_{xy} = \frac{s_{xy}}{s_x \cdot s_y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

13 / 21



Steekproefcorrelatiecoëfficiënt

$$r_{xy} = \frac{s_{xy}}{s_x \cdot s_y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

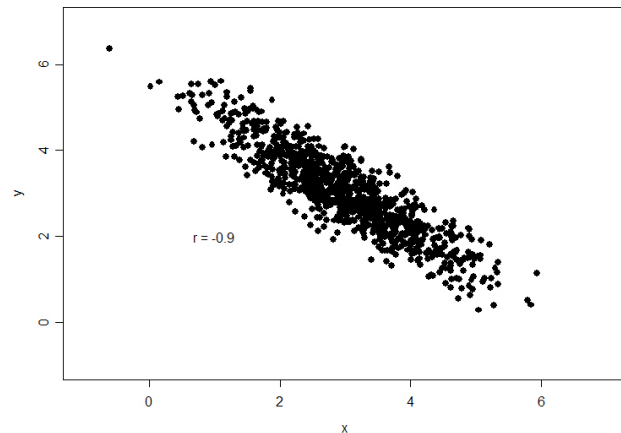
- ▶ is dimensieloos
- ▶ ligt steeds tussen -1 en +1:
 - ▶ $-1 \leq r_{xy} < 0$: negatief lineair verband
 - ▶ $r_{xy} = 0$: geen lineair verband
 - ▶ $0 < r_{xy} \leq 1$: positief lineair verband
- ▶ hoe groter de absolute waarde van r_{xy} , hoe sterker het lineair verband

14 / 21



Steekproefcorrelatiecoëfficiënt

$$r_{xy} = -0.9$$

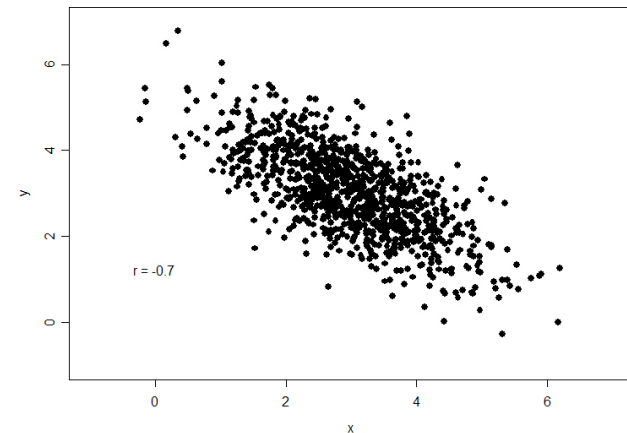


15 / 21



Steekproefcorrelatiecoëfficiënt

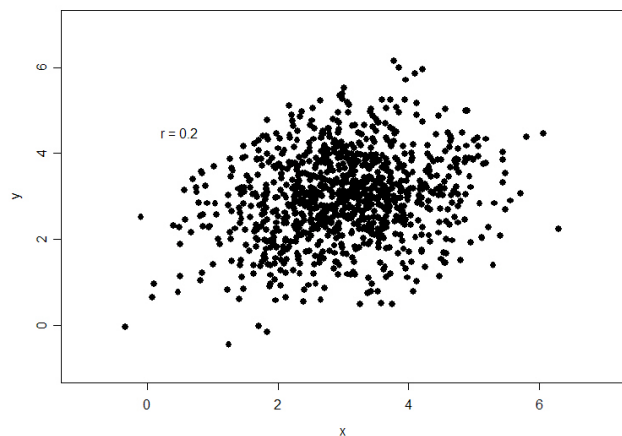
$$r_{xy} = -0.7$$



16 / 21



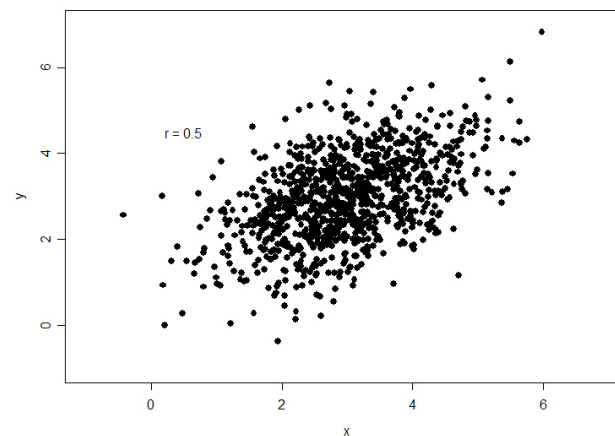
$$r_{xy} = 0.2$$



17 / 21



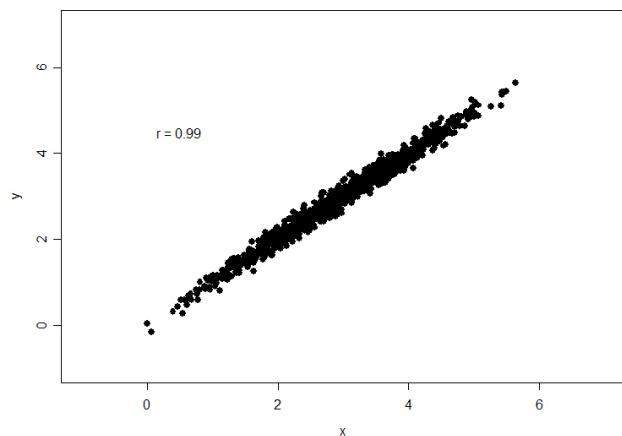
$$r_{xy} = 0.5$$



18 / 21



$$r_{xy} = 0.99$$



19 / 21



- ▶ vb. 'Gem Temp' (°C) en 'Aantal verkochte dozen ijs':

$$r_{xy} = \frac{s_{xy}}{s_x \cdot s_y} = \frac{37.56}{3.40 \cdot 11.29} \approx 0.98$$

- ▶ vb. 'Gem Temp' (°C) en 'Aantal verkochte ijsjes':

$$r_{xy} = \frac{s_{xy}}{s_x \cdot s_y} = \frac{375.56}{3.40 \cdot 112.94} \approx 0.98$$

- ▶ r_{xy} ligt dichtbij 1
 $\Rightarrow$ sterk positief lineair verband tussen X en Y
 $\Rightarrow$ enkelvoudige lineaire regressie is zinvol

20 / 21



- ▶ **Kleinste kwadraten criterium toepassen:**

$$\begin{cases} \beta_0 = -13.5 \\ \beta_1 = 3.25 \end{cases}$$

- ▶ $E(y) = -13.5 + 3.25 \cdot x$
- ▶ Geschatte aantal verkochte dozen ijs
= $-13.5 + 3.25 \cdot \text{gemiddelde temperatuur}$