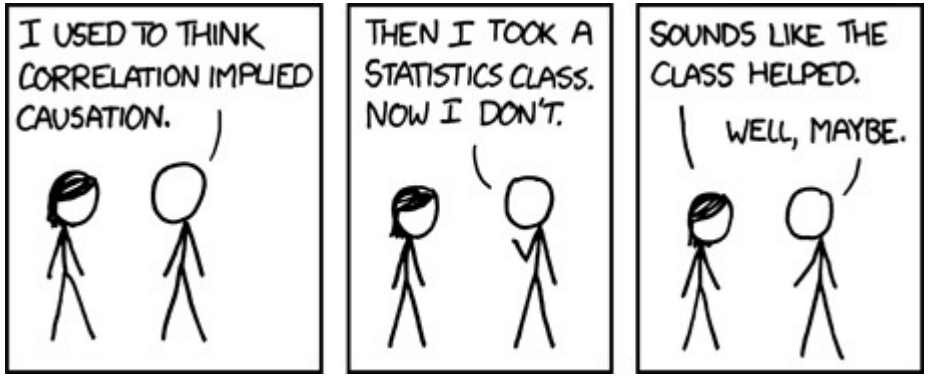
Data-Science 1

correlatie en regressie



KdG Karel de Grote Hogeschool

06/04/2023

Inhoud

- voorbeelden en causaliteit
- correlatie
  - Pearson
  - rangcorrelatie
- regressie
  - lineaire regressie
  - niet-lineaire regressie

Voorbeelden

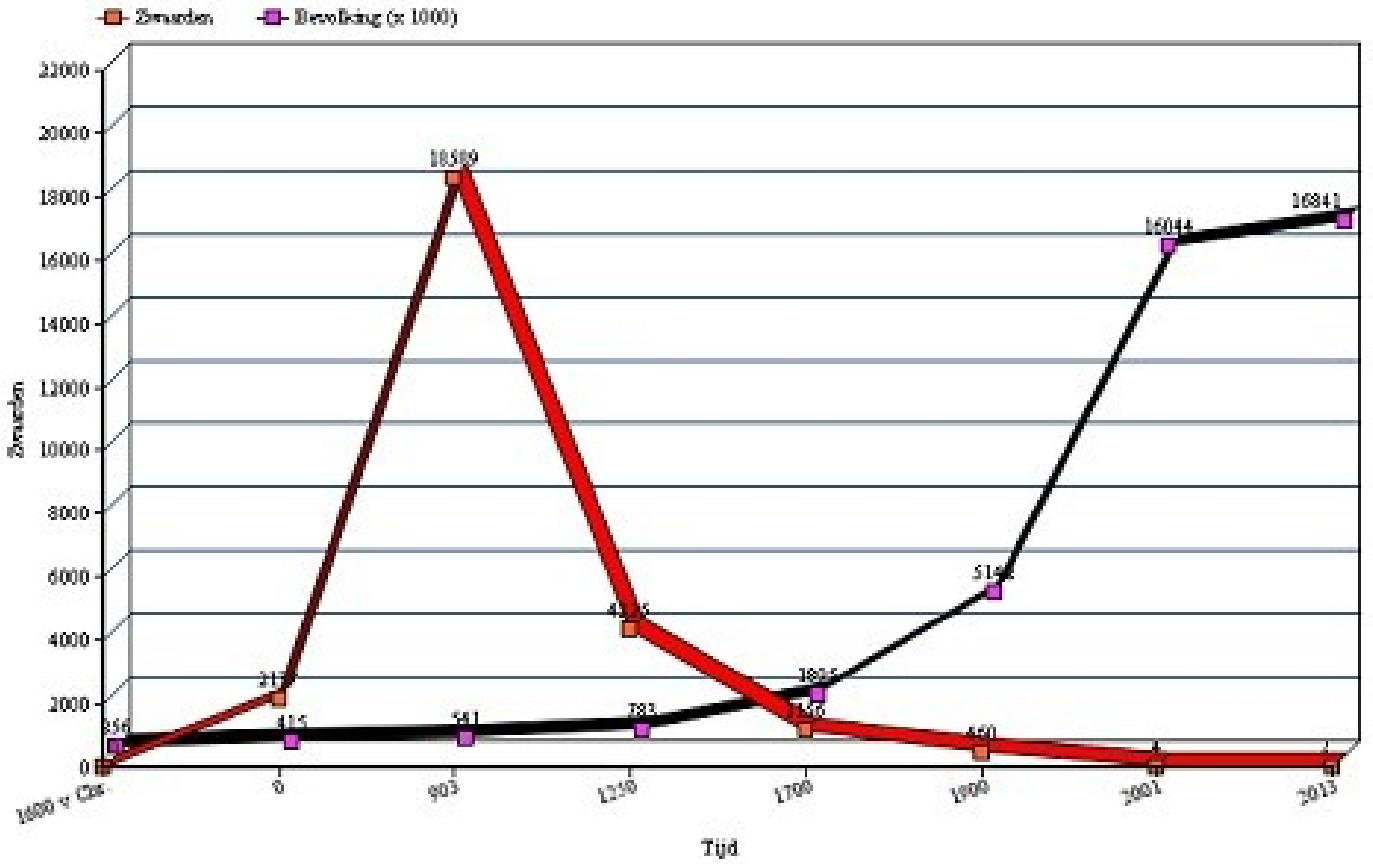
Correlatie

- verbanden zoeken tussen 2 variabelen (bivariate analyse)
  - als de ene var stijgt, dan stijgt ook de andere
  - als de ene var stijgt, dan daalt de andere
- soms is er een "causaal" verband
  - "onafhankelijke variabele" = oorzaak
  - "afhankelijke variabele" = gevolg

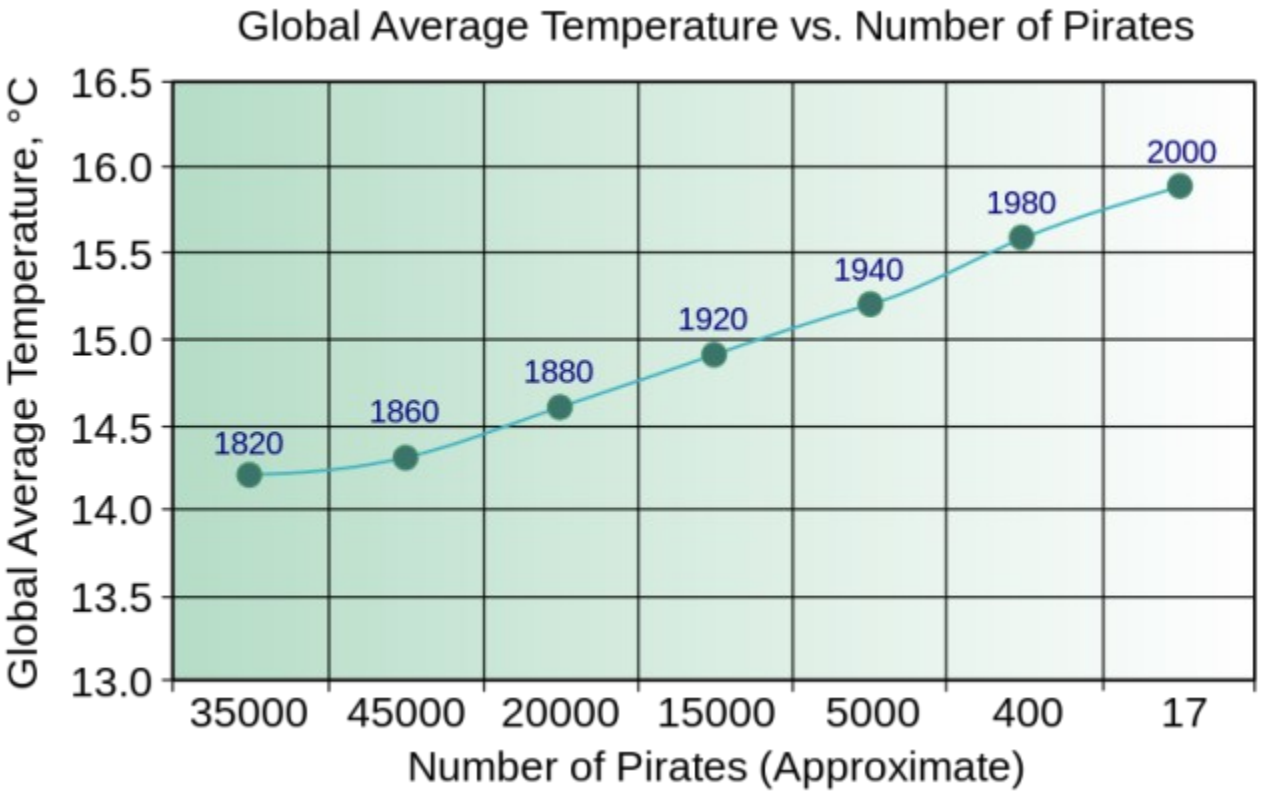
Opletten met causaliteit!

- er is een verband tussen
  - het succes van twitter en de grootte van de Griekse staatsschuld
  - verkoop van ijsjes en de hoeveelheid zakkenrollers
  - de bevolkingsgrootte en het aantal zwaarddragers (https://speld.nl/2013/12/11/steeds-minder-mensen-met-zwaard-op-zak)

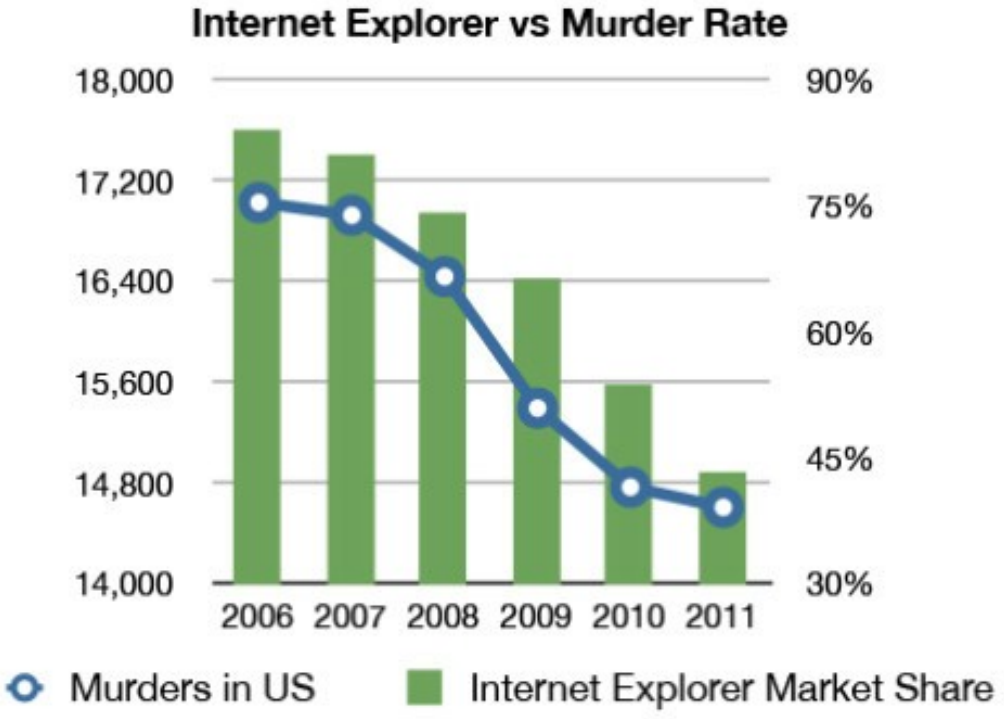
Zwaarddragers



Piraten


Global Average Temperature vs. Number of Pirates
Global Average Temperature, °C
16.5
16.0
15.5
15.0
14.5
14.0
13.5
13.0
1820
1860
1880
1920
1940
1980
2000
35000
45000
20000
15000
5000
400
17
Number of Pirates (Approximate)

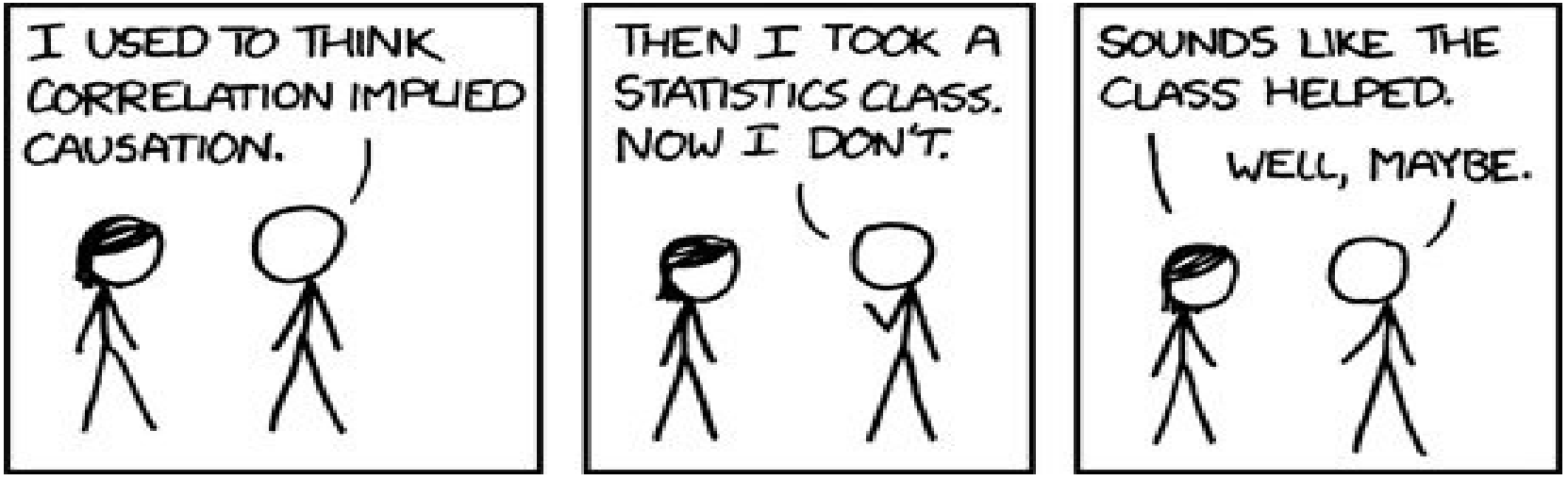
Internet exploder



Happiness


RH
Le (réseau) Socialez-vous ?
Les salariés heureux sont deux fois moins malades, six fois moins absents, 9 fois plus loyaux, 31% plus productifs et 55% plus créatifs.
(Source : Harvard / MIT)
@LeSocialezVous

Statistics class



Meer?

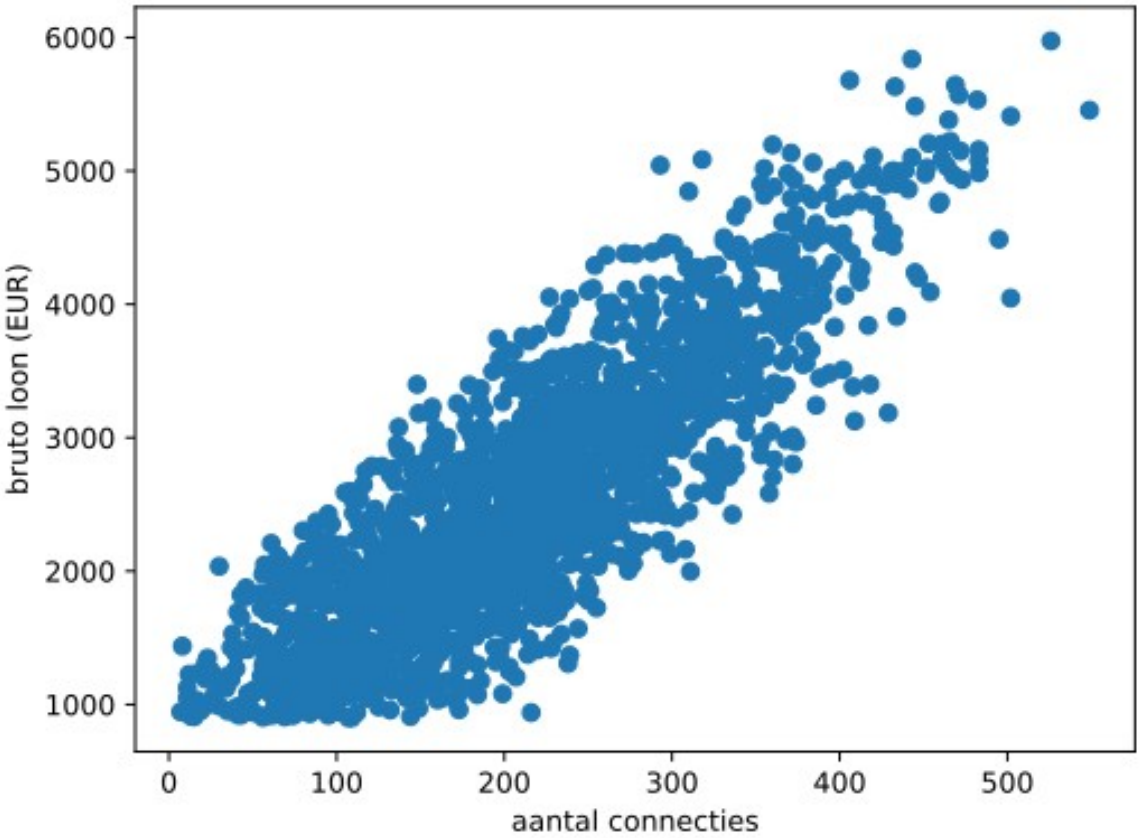
- http://tylervigen.com/spurious-correlations
- "the more things you study, the more likely it is that you’re going to ‘discover’ one of them is statistically significant"
- probleem in big-data: als je lang genoeg zoekt, zal je wel ergens een correlatie vinden...

Voorbeeld voor deze les

- is er een verband tussen het aantal connecties op LinkedIn en het inkomen?
- data in linkedIn.csv (2064 observaties)

loon	connecties
3252	304
2968	216
2976	159
...	...

Scatterplot (zie ook code)



Probleem met scatterplot

- hoe groot is dit verband?
- soms overlappen punten elkaar op de grafiek: moeilijk te zien

Pearson

2

Z-scores

- stel: 2 variabelen x en y
- zet eerst om naar Z-scores

$$z_i = \frac{x_i - \overline{x}}{s_x}$$

- gevolg?
- in python: zie code

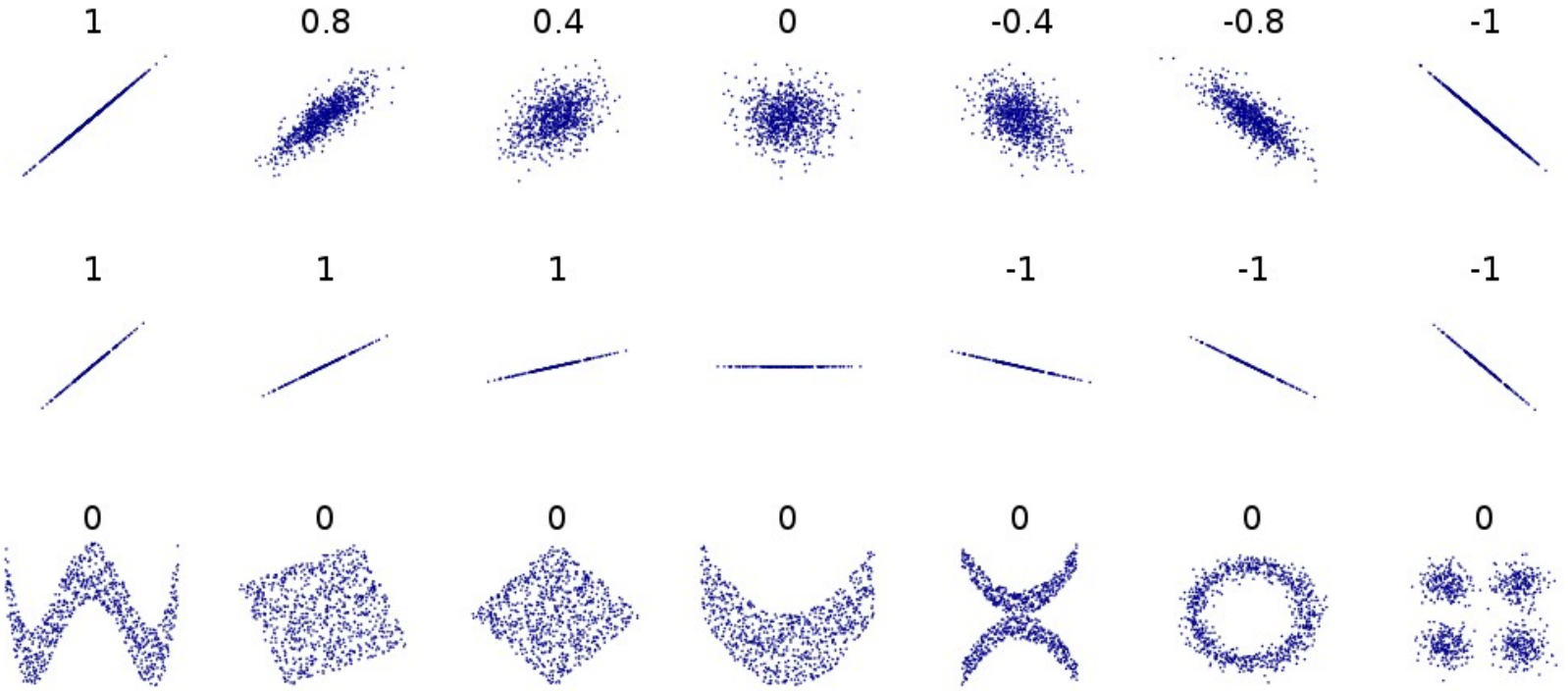
Correlatiecoëfficiënt Pearson

- vermenigvuldig beide reeksen Z-scores met elkaar en neem het gemiddelde

$$r_{xy} = \frac{1}{n} \sum_{i=1}^{n} z_{x_i} \cdot z_{y_i}$$

- resultaat: tussen -1 en 1
- in python: zie code

Resultaat



waarde van r_{xy}

- correlatie r_{xy} ligt altijd tussen -1 en 1
- als $r_{xy} < 0$ dan spreekt met van een negatief verband
- als $r_{xy} = 0$ dan is er geen verband
- als $r_{xy} > 0$ dan spreekt men van een positief verband

Interpretatie

- afhankelijk van de context
  - chemie: verband temperatuur en snelheid van een reactie
    - hoge correlatie verwacht (>0,99)
  - sociale wetenschappen: verband gezinssituatie en geluksgevoel
    - moeilijk om hoge correlatie te vinden (>0,4)
  - informatica: verband netwerktraffiek en aantal aanvallen van hackers
    - beetje tussenin (>0,6)

Afspraken

correlatie	betekenis
0	geen enkel lineair verband
0 tot 0,2	nauwelijks lineair verband
0,2 tot 0,4	zwak lineair verband
0,4 tot 0,6	redelijk lineair verband
0,6 tot 0,8	sterk lineair verband
0,8 tot 1	zeer sterk lineair verband

Rangcorrelatie

Meetniveau's

- vanaf welk niveau kan je correlatie berekenen?
- wat als de variabelen ordinaal zijn?
- voorbeeld: zet linkedIn om naar ordinaal
  - connecties: "weinig", "matig", "gemiddeld", "meer", "veel", "extreem veel"
  - loon: "klein", "modaal", "groot", "extreem"

Rangnummers

- zet ordinale variabele om naar rangnummers
  - kleinste waarde krijgt waarde 1
  - grootste waarde krijgt waarde n
  - als er dubbels zijn: krijgen dezelfde waarde (meestal gemiddelde van alle rank-nummers)
- in python: zie code

Spearman

- deze gebruikt de formule van Pearson op de rangnummers
  - probleem?
- functie in python: zie code

Kendall: principe

- zet score=0
- kijk naar alle combinaties x_i en x_j
  - als $x_i > x_j$ en $y_i > y_j$, dan score++
  - als $x_i < x_j$ en $y_i < y_j$, dan score++
  - als $x_i > x_j$ en $y_i < y_j$, dan score--
  - als $x_i < x_j$ en $y_i > y_j$, dan score--
- deel score door het aantal mogelijke combinaties

Kendall (simple)

	x	y
0	klein	middel
1	middel	groot
2	groot	enorm
3	enorm	groot

i	j	$x_i \rightarrow x_j$	$y_i \rightarrow y_j$	score
0	1	groter	groter	1
0	2	groter	groter	1
0	3	groter	groter	1
1	2	groter	groter	1
1	3	groter	gelijk	0
2	3	groter	kleiner	-1

$$\Rightarrow corr = \frac{1+1+1+1+0-1}{6} = 0{,}5$$

in Python: zie code

4

(lineaire) Regressie

Regressie

- stel: er is een hoge correlatie
- dus...: punten liggen dicht tegen een lijn
- welke lijn? --> regressie
- de lijn laat toe om voorspellingen te maken
  - volgorde van de variabelen is nu wel belangrijk: onafhankelijke en afhankelijke variabelen (x en y)

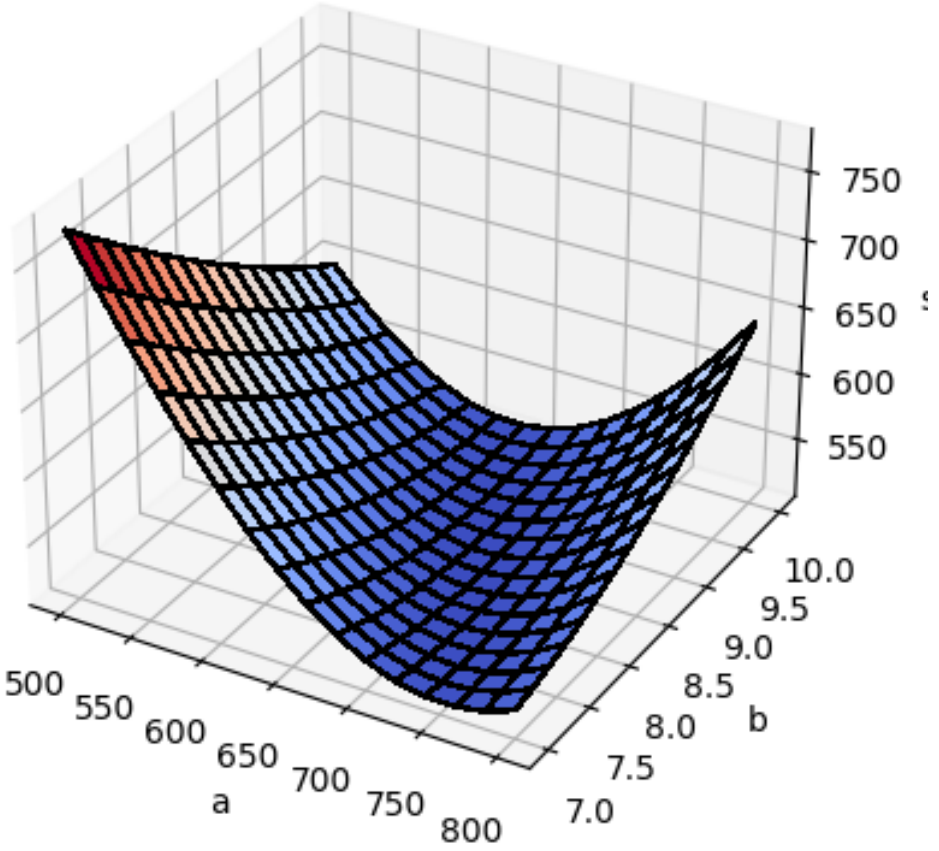
Lineaire regressie

- "lineair" = rechte lijn
- vergelijking: y = a + b * x
- gevraagd: wat is a en b zodat de lijn zo goed mogelijk door de punten gaat?
- als we a en b kennen, kunnen we meten hoe goed de lijn door de punten gaat met:

```
se = math.sqrt((((a+b*x)-y)**2).mean())
```

- doet je dit aan iets denken?

Op zoek naar a en b...



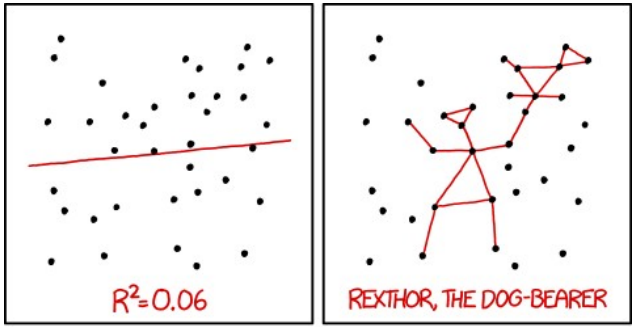
Waarden van a en b

- oplossing:

$$b = r_{xy} \cdot \frac{s_y}{s_x} \qquad a = \overline{y} - b \cdot \overline{x}$$

- a = "intercept" (hoogte van de lijn)
- b = "slope" of "richtingscoëfficiënt"
- in python: zie code

Verklaarde variantie



- de waarde van s_e kan ook geschreven worden als:

  $$s_e = s_y \sqrt{1 - r_{xy}^2}$$

- na kwadrateren geeft dit: $s_e^2 = s_y^2 (1 - r_{xy}^2)$
- als R^2 gelijk is aan 0, dan is de variantie op de fout gelijk aan de variantie op de y-waarden. De y-waarden kunnen dan totaal niet voorspeld worden door het model.
- als de R^2 gelijk is aan 1, dan is er geen variantie op de fout. De y-waarden worden dan perfect voorspeld door het model.
- algemeen: R^2 geeft aan in hoeverre het model de waarde van y kan voorspellen
- men noemt dit: "de verklaarde variantie"
- in python: zie code

Niet-lineaire regressie

Niet-lineaire regressie

- soms kan je geen recht lijn door de punten tekenen, maar wel een andere:
  - kwadratisch: $y = a + b \cdot x + c \cdot x^2$
  - kubisch: $y = a + b \cdot x + c \cdot x^2 + d \cdot x^3$
  - exponentieel: $y = e^{a + b \cdot x}$
  - logaritmisch: $y = a + b \cdot \log(x)$
  - ...
- zoek telkens a, b, c, d zodat s_e zo klein mogelijk is

correlatie en regressie in python

- zie code

Oefeningen

Oefeningen huiswerk

- wat is de Pearson correlatie tussen schoenmaat en lengte?
- wat betekent deze waarde?
- is er een verschil als je de uitschieters eerst verwijdert? Hoe doe je dit?
- kan je ook de Kendall correlatie berekenen? Wat is de waarde?
- heeft het zin om een regressielijn te bepalen?
- zijn er nog andere correlaties te vinden in de enquête? Hint: gebruik de methode corr() van het hele dataframe.
- Kijk eens naar de correlatie tussen opwarming en zakgeld. Wat betekent dat? Heeft het zin om hier een regressielijn te bepalen?

Oefeningen huiswerk

- bereken een lineaire regressie die de schoenmaat voorspelt adh van de lengte
  - verwijder eerst de uitschieters
  - wat is de vergelijking van de rechte?
- welke schoenmaat voorspel je voor iemand van 180cm groot?
- wat is de gemiddelde fout op de voorspelling? Is dit veel of weinig?
- wat is de verklaarde variantie? Wat betekent dit?

Oefeningen

- Samenhang
  - Sociale media vs punten
  - Batterijen
  - Stress
  - Smartphones