



Hoofdstuk 4

K Lineaire regressie

4.1 Spreidingsdiagram

4.2 Lineair verband

4.2.1 Correlatiecoëfficiënt

4.2.2 Regressierechte

4.2.3 Correlatie is geen causaliteit



Opdracht 1 bladzijde 150

In elk van de onderstaande voorbeelden beschouwen we een steekproef waarbij twee variabelen worden gemeten die betrekking hebben op dezelfde populatie.

- a** Het aantal consumpties met alcohol gedronken door een man tussen 20 en 60 jaar in de voorbije twee uren en zijn lichaamstemperatuur.
- b** De lengte en het gewicht van baby's geboren in december van dit jaar in alle Vlaamse ziekenhuizen.
- c** De leeftijden van huwelijkspartners in Vlaams Brabant bij een eerste huwelijk in 2017.
- d** Het aantal tewerkgestelde mannen en het aantal tewerkgestelde vrouwen in Brussel dit jaar.
- e** De examenresultaten voor wiskunde van een zesdejaarsklas en of er een voetbalwedstrijd uitgezonden werd op de avond voor het examen.

Beantwoord voor elk voorbeeld de onderstaande vragen.

- 1** Zijn de variabelen kwantitatief of kwalitatief?

We hernemen de omschrijvingen die in de tweede graad werden geleerd: "Kwantitatieve of numerieke variabelen nemen getalwaarden aan waarvoor het zinvol is een gemiddelde te berekenen of waarop andere bewerkingen kunnen worden uitgevoerd. Het zijn variabelen waarvan de waarde wordt verkregen door meten of tellen."; "Kwalitatieve of categorische variabelen verdelen de elementen in verschillende groepen of categorieën." Jaartallen zullen we in dit boek altijd als kwantitatieve variabele behandelen.

- a kwantitatief (2×)
 - b kwantitatief (2×)
 - c kwantitatief (2×)
 - d kwantitatief (2×)
 - e kwantitatief resp. kwalitatief
- 2** Denk je dat er een verband is tussen beide variabelen? Kun je spreken van oorzaak en gevolg?
- a Er is mogelijk een verband, waarbij de alcohol de oorzaak is van de verandering in lichaamstemperatuur.
 - b Er is zeker een verband. Maar er zijn talloze factoren die beide bepalen, zodat we niet meteen de ene grootheid als onmiddellijke oorzaak van de andere kunnen beschouwen.
 - c Ook hier is er een verband, maar de leeftijd van de ene partner is duidelijk niet de oorzaak van die van de andere.
 - d Er is wellicht een verband tussen beide, maar geen oorzakelijk verband.
 - e Zeker bij de groep van de voetballiefhebbers kan er een verband zijn, waarbij de voetbalwedstrijd de oorzaak is van de examenresultaten.

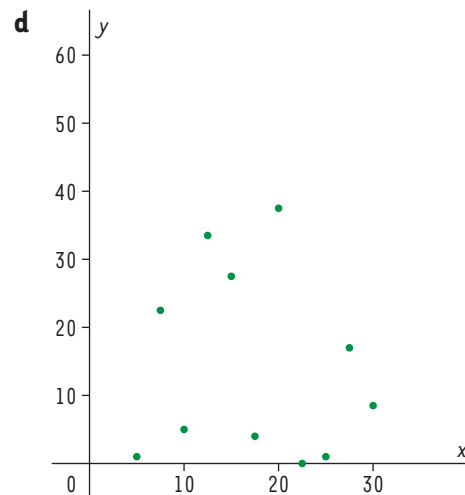
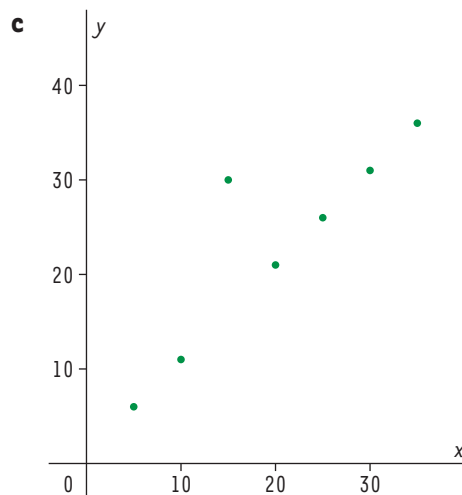
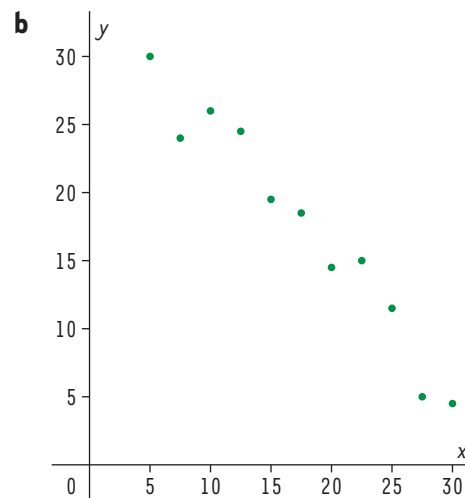
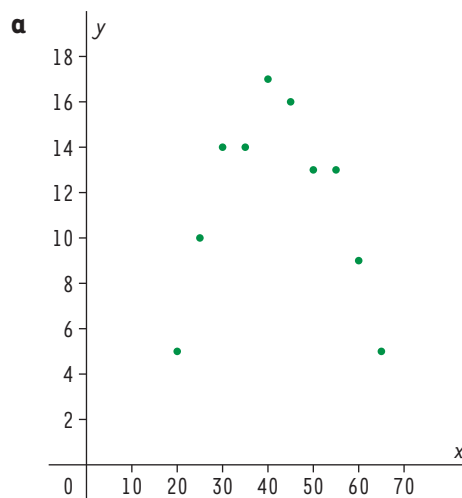
Opdracht 2 bladzijde 150

Hieronder vind je de grafische voorstelling van een aantal getallenkoppels (x, y) .

1 Bespreek telkens de globale vorm.

- a De puntenwolk heeft de vorm van een parabool.
- b De punten vormen een dalende rechte.
- c De punten vormen een stijgende rechte. Er is echter een uitschieter.
- d Hier is geen duidelijk patroon te bespeuren. De punten liggen lukraak in het eerste kwadrant.

2 Denk je dat er een verband bestaat tussen x en y ? Zo ja, bestaat er een rechte of een eenvoudige kromme lijn die de punten goed benadert?



- a Er is een vrij sterk kwadratisch verband tussen x en y .
- b Er is een lineair verband.
- c Als we de uitschieter buiten beschouwing laten, is er opnieuw sprake van een lineair verband tussen x en y .
- d Hier is er geen verband tussen x en y .

Opdracht 3 bladzijde 153

Denk je dat er correlatie is in de volgende gevallen? Indien ja, is de correlatie dan positief/negatief en sterk/zwak?

- 1 Het aantal ooievaars en het aantal geboorten vorig jaar in een bepaalde stad.

Geen correlatie, een eventueel verband berust louter op toeval.

- 2 De hoeveelheid pasta en de hoeveelheid aardappelen die jaarlijks verbruikt worden in een Vlaams gezin van 4 personen.

Vermoedelijk is er wel een verband: naarmate gezinnen van 4 personen meer aardappelen eten, zullen ze wellicht minder pasta eten en omgekeerd (ze zullen op dezelfde dag geen aardappelen en pasta tegelijkertijd eten). We vermoeden dus een negatief verband. Aangezien er nogal wat factoren zijn die beide grootheden bepalen, kunnen we eerder een relatief zwak verband verwachten.

Opdracht 4 bladzijde 154

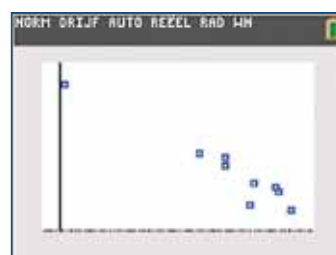
In de tabel vind je een aantal voedingsmiddelen met hun gehalte aan vetten en koolhydraten en hun energetische waarde uitgedrukt in kcal en kJ.

per 100 g	vetten (g)	koolhydraten o.a. suikers (g)	energetische waarde (kcal)	energetische waarde (kJ)
chocopasta Becel	35	52	552	2310
chocolade noten Côte d'Or	37	44	562	2351
zandkoekjes Barilla	18,7	68	475	1987
Prince koekjes chocolade LU	17	69	465	1946
koekjes TUC paprika	23	61	485	2029
ronde beschuiten	7,1	72,8	415	1736
mayonaise met eieren DL	81,1	1,6	743	3109
rozijnensterren	18	60	429	1795
chips paprika Lays	32	52	531	2222

- 1 Maak een spreidingsdiagram waarbij je de correlatie onderzoekt tussen de hoeveelheid koolhydraten en de energetische waarde in kcal. Zet de verklarende variabele op de horizontale as. Bespreek de vorm en de richting van de puntenwolk.

Op de TI-84 kunnen de gegevens ingeladen worden via het programma, zoals uitgelegd bij opdracht 2 van hoofdstuk 2.

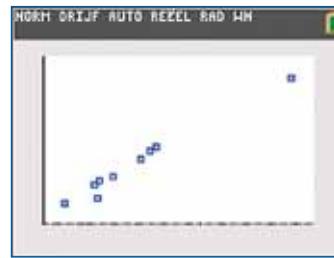
We zetten de verklarende variabele, de hoeveelheid koolhydraten, op de horizontale as (in XList) en de verklaarde variabele, de energetische waarde, op de verticale as (in YList).



De puntenwolk is dalend, de punten liggen dicht bij een dalende rechte. De correlatie is eerder sterk en negatief lineair.

- 2 Doe nu hetzelfde voor de vetten en de energetische waarde in kcal.

We zetten de verklarende variabele, de hoeveelheid vetten, op de horizontale as (in XList) en de verklaarde variabele, de energetische waarde, op de verticale as (in YList).



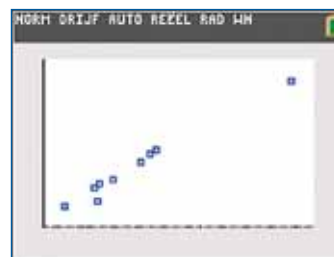
De puntenwolk is stijgend, de punten liggen dicht bij een stijgende rechte. De correlatie is eerder sterk en positief lineair.

- 3 Bij welk van beide is de correlatie het sterkst?

De correlatie tussen vetten en energetische waarde is sterker dan die tussen koolhydraten en energetische waarde.

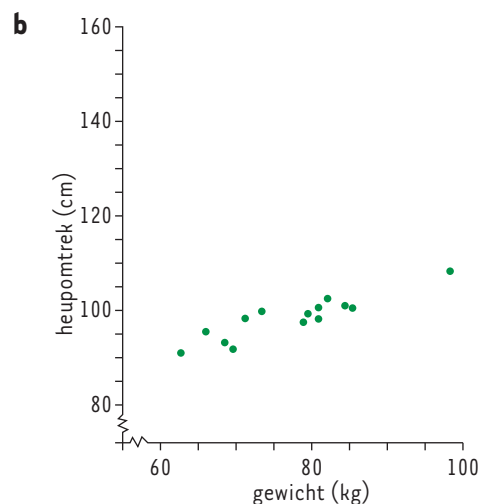
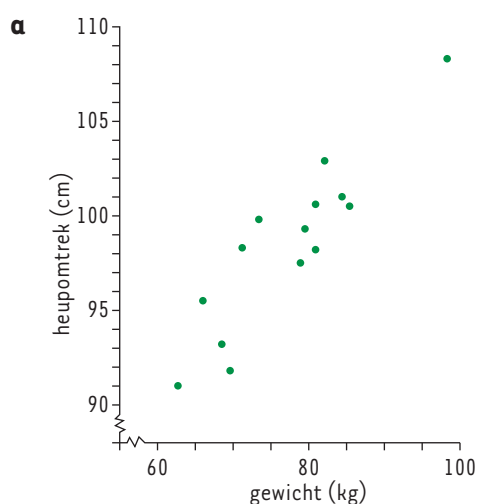
- 4 Zou je op vraag 3 een ander antwoord geven als je de energetische waarde uitdrukt in kJ in plaats van in kcal?

Neen, de eenheid heeft geen invloed op de correlatie. Ter controle:



Opdracht 5 bladzijde 154

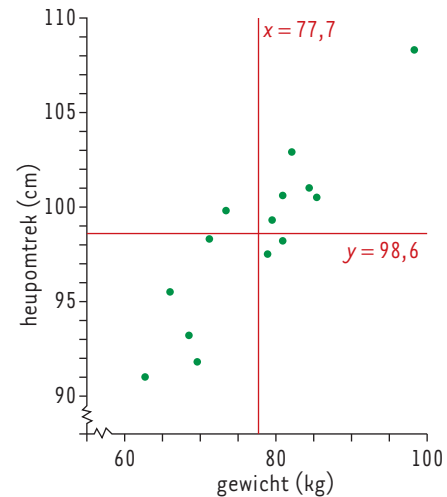
Zoë en Lucas willen de sterkte van de samenhang bepalen tussen het gewicht en de heupomtrek bij eenzelfde steekproef van 15 personen en maken elk een spreidingsdiagram uitgaande van dezelfde gegevens.



- 1 Zoë beweert dat er een matig tot sterke positieve samenhang is en Lucas dat de positieve samenhang zeer sterk is. Wie maakte welk spreidingsdiagram?

Het eerste spreidingsdiagram lijkt een minder sterke positieve samenhang te vertonen dan het tweede. Zoë maakte dus diagram a en Lucas diagram b.

- 2** Omdat grafische voorstellingen wat bedrieglijk kunnen zijn, willen we de mate van samenhang cijfermatig uitdrukken. Daartoe berekenen we de gemiddelde heupomtrek $\bar{y} = 98,6$ cm en het gemiddelde gewicht $\bar{x} = 77,7$ kg van de 15 personen. Dan tekenen we de horizontale rechte $y = 98,6$ en de verticale rechte $x = 77,7$. Deze rechten verdelen de puntenwolk in vier kwadranten. Vervolgens tellen we het aantal punten in elk kwadrant. De punten in de kwadranten I en III geven we een waarde +1 en die in II en IV een waarde -1. De som van deze waarden is een positief getal als de correlatie positief is en de meeste punten in het eerste en derde kwadrant liggen, zoals hier het geval is.



Bereken op deze manier de mate van samenhang.

In beide gevallen vinden we $6 - 1 + 5 - 2 = 8$.

- 3** Wat zijn volgens jou de plus- en minpunten van deze maat voor de correlatie?

plus: het getal is gemakkelijk te berekenen, louter door punten te tellen; het eindresultaat zal normaal gezien positief zijn bij positieve correlatie en negatief in het andere geval, zoals we van een correlatiegetal verwachten.

min: er wordt geen rekening gehouden met de concrete ligging van de punten in de kwadranten, enkel met het aantal punten.

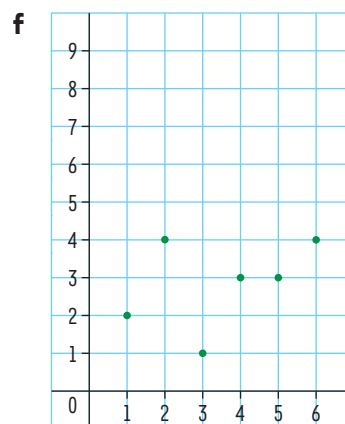
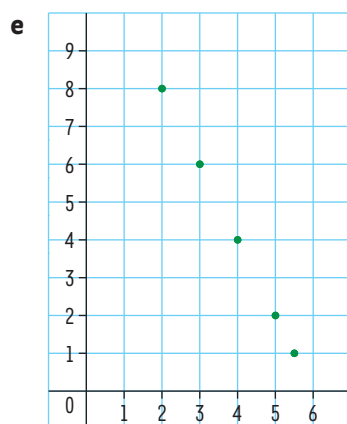
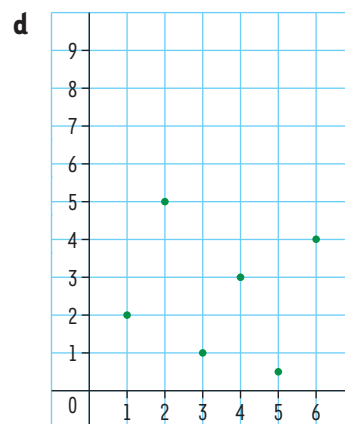
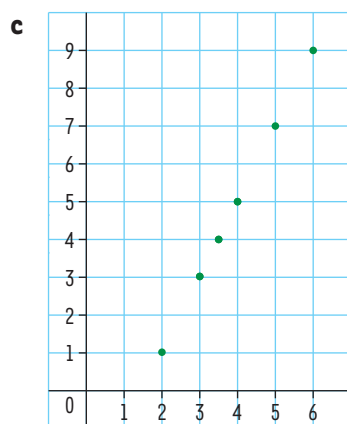
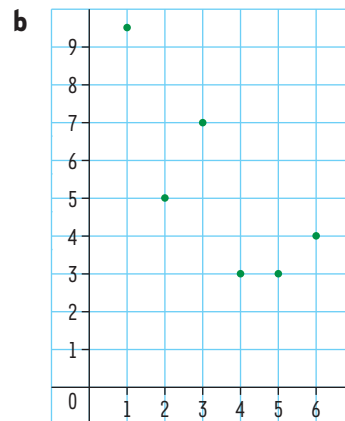
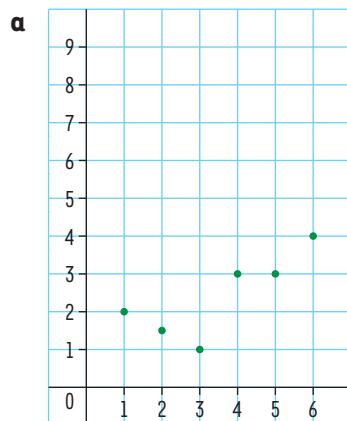
Opdracht 6 bladzijde 160

Denk je dat de correlatiecoëfficiënt gevoelig is voor uitschieters? Leg uit.

De formule van Pearson voor de correlatiecoëfficiënt maakt gebruik van het gemiddelde en de standaardafwijking van beide variabelen. Vermits beide gevoelig zijn voor uitschieters zal dit ook zo zijn voor de correlatiecoëfficiënt.

Opdracht 7 bladzijde 160

Plaats bij elke puntenwolk de gepaste correlatiecoëfficiënt. Kies uit de volgende waarden: -1 ; $-0,7$; $-0,05$; $0,4$; $0,79$; 1 .



De correlatiecoëfficiënt is positief bij een stijgende puntenwolk (fig. a, c, f). Bij fig. c liggen de punten op een rechte. Bij a is de samenhang iets sterker dan bij f. Bijgevolg:

a $0,79$

c 1

f $0,4$

De puntenwolk is dalend in fig. b en e. In fig. e liggen de punten op een rechte. Daarom:

b $-0,7$

e -1

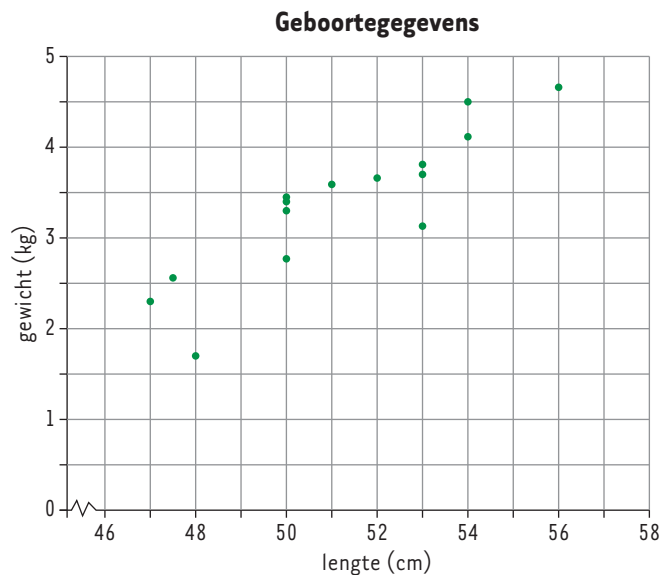
Bij fig. d is niet echt duidelijk of de correlatie positief dan wel negatief is. De punten liggen verspreid, de correlatiecoëfficiënt zal bijgevolg dicht bij 0 liggen. De overblijvende correlatiecoëfficiënt komt dus wel degelijk met figuur d overeen:

d $-0,05$

Opdracht 8 bladzijde 161

We onderzoeken de correlatie tussen de lengte en het geboortegewicht van een baby aan de hand van een steekproef van 15 lukraak gekozen baby's. Hieronder zie je de puntenwolk met de lengte op de horizontale as en het gewicht op de verticale.

lengte (cm)	gewicht (kg)
51	3,590
53	3,810
52	3,660
54	4,115
50	3,400
56	4,660
53	3,130
50	2,770
47	2,300
48	1,700
50	3,450
50	3,300
53	3,700
47,5	2,560
54	4,500



- 1 De punten op het spreidingsdiagram liggen duidelijk dicht bij een rechte. Schets een rechte die volgens jou de punten goed benadert.
- 2 Geef een vergelijking van de rechte die je getekend hebt.
- 3 Voorspel aan de hand van deze rechte het gewicht van een boreling van 49 cm.

De getekende rechten zullen hier lichtjes van elkaar verschillen, alsook de voorspelde waarde voor een boreling van 49 cm.

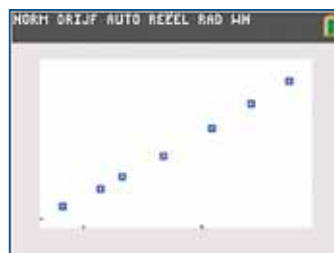
Opdracht 9 bladzijde 166

Lopers zetten meer stappen per seconde naarmate hun snelheid toeneemt. Hieronder vind je een tabel met het gemiddeld aantal stappen per seconde van vrouwelijke competitie-lopers bij verschillende snelheden.

snelheid (m/s)	4,83	5,15	5,33	5,68	6,09	6,42	6,74
stappen per s	3,05	3,12	3,17	3,25	3,36	3,46	3,55

- 1 Je wil het aantal stappen per seconde voorspellen uit de snelheid. Maak hiervoor een spreidingsdiagram.

We kiezen de snelheid als verklarende variabele.



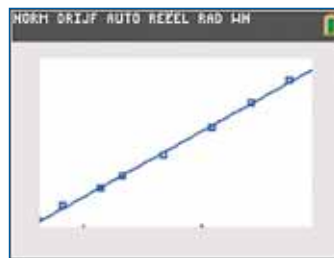
- 2 Beschrijf de correlatie en bepaal de correlatiecoëfficiënt.

De correlatie is sterk stijgend en lineair. De correlatiecoëfficiënt $r = 0,999$.



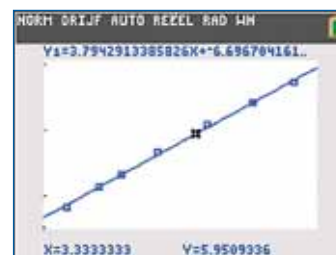
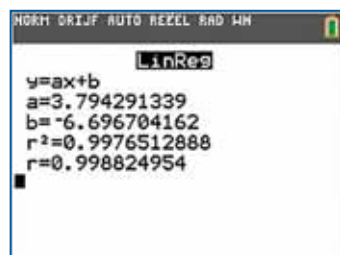
- 3 Bepaal een vergelijking van de regressielijn en plot ze.

Een vergelijking van de regressielijn is $y = 0,263x + 1,769$.



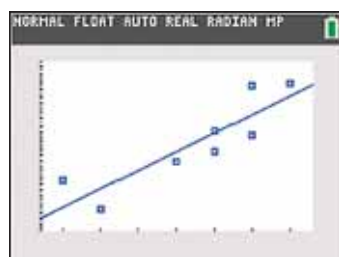
- 4 Een loopster telt 100 stappen gedurende 30 s. Kun je haar snelheid voorspellen?

Om haar snelheid te voorspellen, willen we het aantal stappen per seconde als verklarende variabele. We vullen deze waarde in de vergelijking van de nieuwe regressielijn in en vinden 5,95 m/s (zie schermafdruck rechts).



Opdracht 10 bladzijde 167

Een student voert een onderzoek uit naar de leesvaardigheid van kinderen. Hij noteert telkens de schoenmaat van de proefpersonen en hun resultaat in procent op een woordenschattest. Hij vindt de puntenkoppels (34, 49), (33, 60), (36, 67), (37, 71), (37, 79), (38; 77,5), (38, 96) en (39, 97). Na het tekenen van het spreidingsdiagram merkt hij dat er een lineair verband is tussen deze variabelen.



- 1 Euforisch besluit hij: "Kinderen met grote voeten zijn taalvaardiger". Wat denk je van deze uitspraak?

De schoenmaat wordt hier als verklarende variabele gebruikt. Met grotere schoenmaten komen hogere cijfers op de leestest overeen. Er wordt hier echter een foutief causaal verband vastgesteld omdat zowel schoenmaat als leesvaardigheid correleren met een derde ontbrekende factor: de leeftijd.

- 2 Hij denkt dat hij oorzaak en gevolg verwart en verbetert zichzelf: "Door beter te leren lezen, groeien je voeten". Wat denk je van deze uitspraak?

Als je de richting van de oorzakelijkheid omdraait, krijg je om dezelfde reden een non-sens-correlatie.

Opdracht 11 bladzijde 174

In de plantentuin van Meise en op verschillende plaatsen in Vlaanderen werden sensoren geplaatst, enerzijds in het open veld en anderzijds in een bosrijke omgeving. De impact van bomen op het klimaat wordt hiermee onderzocht. Deze sensoren sturen permanent gegevens over temperatuur, CO₂, luchtvochtigheid, fijnstof ... naar de website van Bos Online.



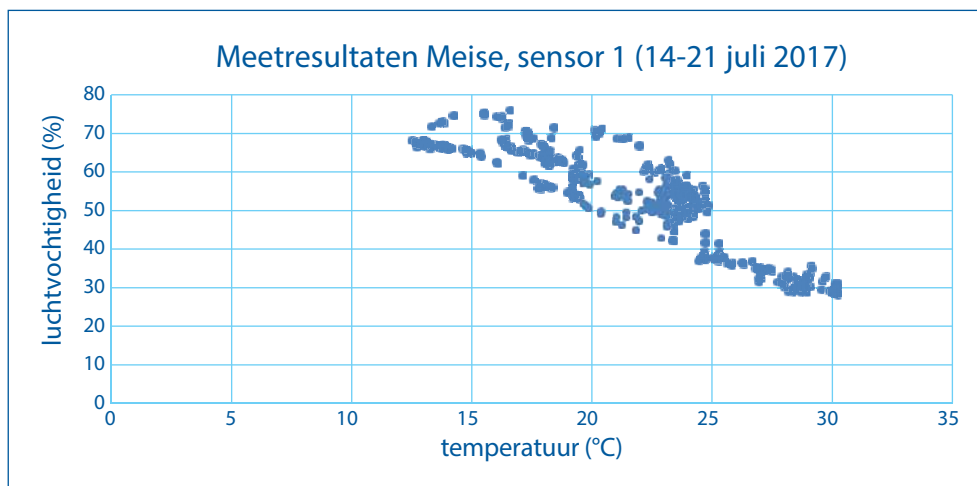
Voor deze opdracht gebruiken we een tabel met 551 temperatuursgegevens (°C) en gegevens van luchtvochtigheid (%), gemeten door sensor 1 op het open veld in Meise gedurende de week van 14 tot 21 juli 2017.

- 1 Denk je dat er een verband is tussen temperatuur en luchtvochtigheid? Beschrijf dit verband.

Velen zullen een verband tussen beide grootheden verwachten. Mogelijk verwachten sommigen dat hogere temperaturen meer verdamping opleveren en dus een hogere luchtvochtigheid. In dat geval zou er dus een positieve correlatie zijn. Anderen zullen mogelijk niet kunnen inschatten of de samenhang positief dan wel negatief is.

- 2 Bereken de correlatiecoëfficiënt.

We vinden: $r = -0,88753$. De correlatie blijkt dus negatief te zijn!



De verklaring voor deze negatieve correlatie is de volgende. Naarmate de temperatuur toeneemt, kan lucht meer waterdamp (bijvoorbeeld in g per m³) opnemen vóór er verzadiging optreedt. Verzadiging komt overeen met een (relatieve) luchtvochtigheid van 100 %. Bij toenemende temperatuur zal eenzelfde (absolute) luchtvochtigheid (in g per m³) overeenkomen met een afnemende relatieve luchtvochtigheid (in % van de hoeveelheid bij verzadiging). Dit is wat het spreidingsdiagram weergeeft.

Opdracht 12 bladzijde 174

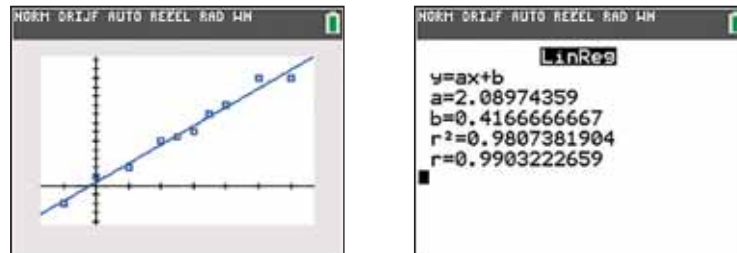
- 1 Plot met je rekentoestel of computer het spreidingsdiagram bij de volgende getallenreeksen.
- 2 Bereken en interpreteer de correlatiecoëfficiënt.
- 3 Zoek indien zinvol een vergelijking van de regressierechte en plot ze.

a

x	-1	0	1	2	2,5	3	3,5	4	5	6
y	-2	1	2	5	5,5	6	8	9	12	12

De puntenwolk is lineair en stijgend.

De correlatiecoëfficiënt is 0,990. Dit wijst op een sterke lineaire correlatie.

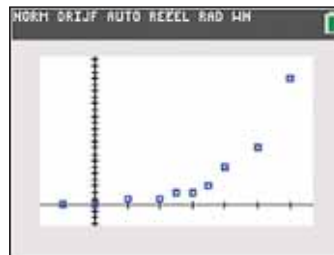


Een vergelijking van de regressierechte is $y = 2,090x + 0,417$.

b

x	-1	0	1	2	2,5	3	3,5	4	5	6
y	0,1	0,2	1	1	2	2	3	6	9	20

De vorm van de puntenwolk wijst op een exponentieel verband.

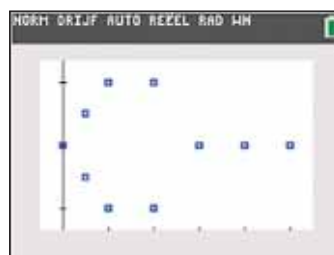


De correlatiecoëfficiënt is 0,814, maar bij een niet-lineair verband heeft dit geen betekenis.

c

x	0	0,5	0,5	1	1	2	2	3	4	5
y	2	1,5	2,5	1	3	1	3	2	2	2

De puntenwolk is verspreid.

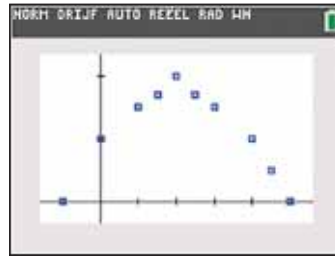


$r = 0,135$; gezien het niet-lineair verband kan hier geen zinvolle interpretatie aan gegeven worden.

d

x	-1	0	1	1,5	2	2,5	3	4	4,5	5
y	0	1	1,5	1,7	2	1,7	1,5	1	0,5	0

De puntenwolk verloopt eerst stijgend en dan dalend.



$r = -0,117$; geen zinvolle interpretatie.

Opdracht 13 bladzijde 175

Archaeopteryx is een van de oudste bekende vliegende dinosauriërs en wordt ook wel de Oervogel genoemd. Deze uitgestorven vogel leefde in het late Jura (ongeveer 150 miljoen jaar geleden) in het huidige Duitsland.

In 2017 waren in totaal twaalf skeletten opgegraven die nogal in grootte verschillen. Volgens sommige geleerden behoren die allemaal tot één soort: Archaeopteryx lithographica. Deze archeologen gaan ervan uit dat het om beenderen van niet-volggroeide vogels gaat en dat er een lineair verband bestaat tussen de afmetingen van bijvoorbeeld de femur (dijbeen) en de humerus (opperarmbeen). Maar er zijn ook archeologen die van mening zijn dat de skeletten niet tot één soort behoren.

In 1990 publiceerden Marilyn Houck, Jacques Gauthier en Richard Strauss een artikel in Science, waar ze voor de vijf

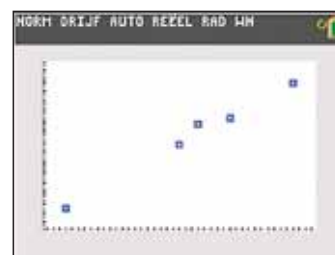
toen gekende fossielen hun standpunt verdedigden, onder andere aan de hand van een spreidingsdiagram van de data van de lengtes (in mm) van femur en humerus.

Kun je aan de hand van een spreidingsdiagram van de data hieronder het standpunt van Marilyn Houck en haar collega's te weten komen?



lengte femur (mm)	38	56	59	64	74
lengte humerus (mm)	41	63	70	72	84

Er is een sterke lineaire correlatie.



Marilyn Houck en haar collega's verdedigden op basis van deze sterke correlatie het standpunt dat het hier over dieren van eenzelfde soort Archaeopteryx lithographica gaat.

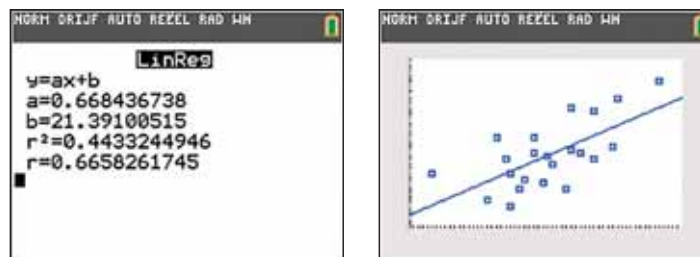
Opdracht 14 bladzijde 176

Mirte behaalde in juni op het examen wiskunde 71 %. Door een fout in de computer zijn haar punten op het kerstexamen echter gewist en ook haar examen is in het archief spoorloos verdwenen. Nu wil de leraar haar toch cijfers op het rapport geven. Hij vraagt aan de ouders of ze akkoord gaan dat hij dat doet door de punten van kerst van Mirte te voorspellen aan de hand van haar resultaat in juni. Hij zal hiervoor een regressierechte opstellen die het lineair verband modelleert tussen de examenresultaten van alle leerlingen van december en juni. De punten van de medeleerlingen op het kerst- en juni-examen zijn:

kerst (%)	53	71	74	67	87	57	68	74	93	67	62	51	69	57	84	83	59	70	62	60	65
juni (%)	57	84	67	80	85	74	70	59	94	61	45	62	77	64	75	80	69	75	62	65	71

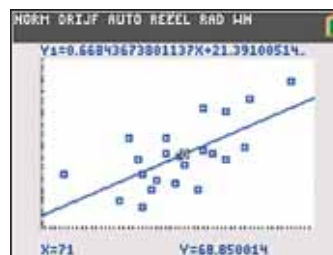
- 1 Zou jij hiermee akkoord gaan? Beargumenteer je antwoord.

Onderzoeken we de samenhang tussen beide reeksen resultaten, dan zien we dat de samenhang niet zeer sterk is: $r = 0,666$. De voorspelling van haar kerstresultaat op basis van haar punten in juni zal dus niet zo betrouwbaar zijn. Mirte zou dus gemakkelijk kunnen beweren dat ze met kerstmis veel hoger scoorde dan het regressiemodel voorspelt, bijvoorbeeld omdat ze toen harder had gestudeerd, toen minder 'pech' had met de vragen of een andere reden.



- 2 Welke punten zou de leerkracht geven mocht hij zijn methode toepassen?

We voeren een lineaire regressie uit en vermits we de punten van kerst willen voorspellen uit het juniresultaat zullen we de punten van juni gebruiken als verklarende variabele. We stellen een lineair model op: $y = 0,668 + 21,391$. Vervolgens vullen we 71, de punten van juni van Mirte, in de vergelijking van de regressielijn in. We vinden 69 (zie schermafdruck).



Opdracht 15 bladzijde 176

Francis Anscombe ontwierp vier voorbeelden om valkuilen bij lineaire regressie te illustreren.

I	
x	y
10,0	8,04
8,0	6,95
13,0	7,58
9,0	8,81
11,0	8,33
14,0	9,96
6,0	7,24
4,0	4,26
12,0	10,84
7,0	4,82
5,0	5,68

II	
x	y
10,0	9,14
8,0	8,14
13,0	8,74
9,0	8,77
11,0	9,26
14,0	8,10
6,0	6,13
4,0	3,10
12,0	9,13
7,0	7,26
5,0	4,74

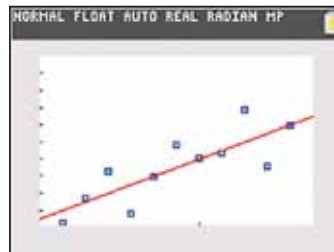
III	
x	y
10,0	7,46
8,0	6,77
13,0	12,74
9,0	7,11
11,0	7,81
14,0	8,84
6,0	6,08
4,0	5,39
12,0	8,15
7,0	6,42
5,0	5,73

IV	
x	y
8,0	6,58
8,0	5,76
8,0	7,71
8,0	8,84
8,0	8,47
8,0	7,04
8,0	5,25
19,0	12,50
8,0	5,56
8,0	7,91
8,0	6,89

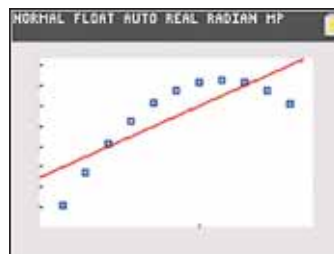
- 1 Bereken bij elk van de vier voorbeelden de correlatiecoëfficiënt en een vergelijking van de regressierechte.
- 2 Stel de gegevens en de regressierechte telkens grafisch voor en bespreek de kwaliteit van het lineair model.

Voor de vier situaties vind je: $r = 0,816$ en $y = 0,50x + 3,00$.

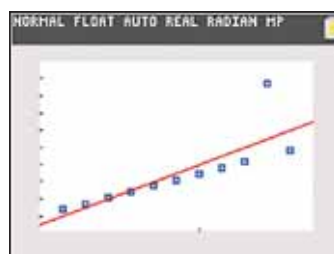
I Het lineaire model is een goede weergave van de relatie tussen x en y.



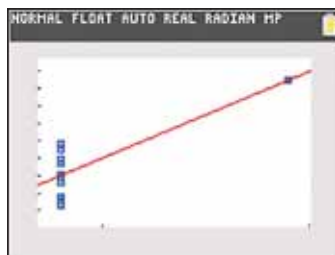
II Het lineaire model voldoet hier niet. De correlatie is wellicht kwadratisch en zeker niet lineair.



III Er is een sterk lineair verband tussen x en y, maar door die ene uitschieter levert de kleinste kwadratenmethode niet het model op dat het best dit lineair verband weergeeft. Het best passend lineair model zou beter opgesteld worden na het verwijderen van die uitschieter.



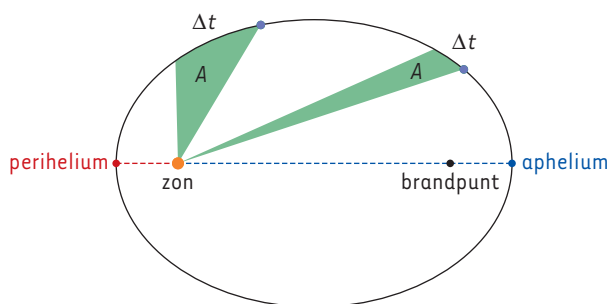
IV Het lineair model wordt bepaald door de ene afwijkende waarde bij 19 en geeft dus geen goed beeld van het verband tussen x en y . Als we die uitschieter weglaten, is de puntenwolk verticaal.



Opdracht 16 bladzijde 177

Johannes Kepler beschreef de beweging van de planeten in zijn boek *Mysterium cosmographicum* in drie wetten.

- De eerste wet zegt dat alle planeten bewegen rond de zon in elliptische banen, met de zon in een van de brandpunten van de ellips.
- De perkenwet zegt dat in gelijke tijdsintervallen Δt gelijke oppervlakten A van de ellips worden beschreven.
- In de derde wet zegt Kepler dat het kwadraat van de omlooptijd P van een planeet rond de zon evenredig is met de derdemacht van de halve afstand van perihelium tot aphelium d (dit noemt men ook de gemiddelde afstand van de planeet tot de zon).



Voor enkele planeten van ons zonnestelsel vind je P en d in de volgende tabel.

planeet	omlooptijd rond de zon P (jaar)	gemiddelde afstand planeet-zon d (10^6 km)
Mercurius	0,24	57,91
Venus	0,62	108,21
Aarde	1	149,60
Mars	1,88	227,94
Jupiter	11,86	778,41
Saturnus	29,46	1426,73
Uranus	84,02	2870,97
Neptunus	164,80	4498,25

Uit de derde wet van Kepler volgt dat $P^2 = k \cdot d^3$, met k een evenredigheids-constante die voor alle planeten van ons zonnestelsel dezelfde is. Aangezien het verband tussen P en d niet lineair is, kunnen we k niet via lineaire regressie bepalen.

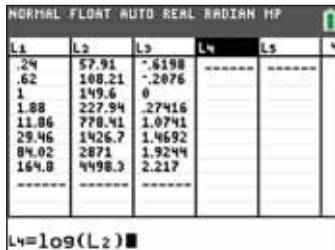
- Door van beide leden van $P^2 = k \cdot d^3$ een logaritme te nemen, bijvoorbeeld de Briggse logaritme, kan de formule omgevormd worden tot een eerstegraadsverband tussen $\log P$ en $\log d$. Schrijf $\log P$ als $a \log d + b$, met a en b constanten.

$$\log P^2 = \log(kd^3) \Leftrightarrow 2 \log P = \log k + 3 \log d \Leftrightarrow \log P = \frac{\log k + 3 \log d}{2}$$

$$\text{Hieruit volgt: } a = \frac{3}{2} \text{ en } b = \frac{\log k}{2}.$$

- 2 Bepaal de coëfficiënten a en b uit de vorige deelvraag met behulp van lineaire regressie. Ga grafisch de samenhang na.

Op de TI-84 kan gemakkelijk van alle gegevens in een lijst de logaritme in één keer berekend worden. Om bijvoorbeeld van de gegevens in L_2 de logaritme in L_4 te plaatsen, volstaat het om in de hoofding van lijst L_4 de opdracht **log(L₂)** in te voeren (zie linkse schermafdruck hieronder, waarin ook al $L_3 = \log(L_1)$). De lineaire regressie wordt vervolgens toegepast op L_3 en L_4 . We vinden: $a = 1,500$ en $b = -3,261$; de samenhang blijkt bovendien zeer sterk te zijn.



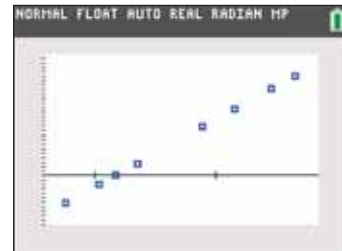
L1	L2	L3	L4	L5
.24	57.91	1.76198		
.62	108.21	2.03376		
1	149.6	2.17616		
1.88	227.94	2.35741		
11.86	778.41	2.89141		
29.46	1426.7	3.15392		
84.02	2871	3.45844		
164.8	4498.3	3.6517		

$L4 = \log(L2)$



LinReg

$y = ax + b$
 $a = 1.49966122$
 $b = -3.261362006$
 $r^2 = .9999982491$
 $r = .9999991245$

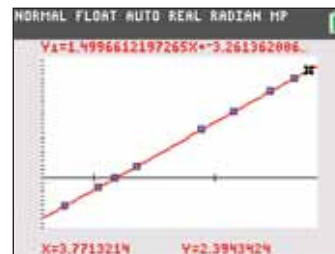
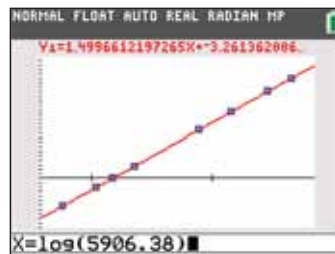


- 3 Leid hieruit een goede schatting van de evenredigheidsconstante k af.

$$b = \frac{\log k}{2} \text{ dus } k = 10^{2b}; \text{ aangezien } b = -3,261, \text{ volgt hieruit dat } k = 3,006 \cdot 10^{-7}.$$

- 4 Welke omlooptijd voorspel je voor Pluto, een dwergplaneet op een gemiddelde afstand van $5906,38 \cdot 10^6$ km van de zon? (Vergelijk met de werkelijke omlooptijd van 247,94 jaar.)

We zoeken de omlooptijd die hoort bij $d = 5906,38 \cdot 10^6$ km. In de tabel wordt de afstand in 10^6 km uitgedrukt en we maakten een regressiemodel met $\log d$. We zoeken de bijhorende y -waarde van $\log(5906,38) = 3,771$ en vinden $\log P = 2,3943424$ (zie schermafdruck). Hieruit volgt: $P = 10^{2,3943424} = 247,9376$.

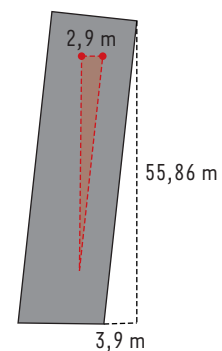


De omlooptijd van Pluto is volgens ons model 247,94 jaar. Dit komt precies overeen met de werkelijke omlooptijd.

Opdracht 17 bladzijde 178

De klokkentoren van de kathedraal van Pisa is beter bekend als de scheve Toren van Pisa. Reeds na het afwerken van de tweede verdieping, rond 1178, zorgden een te zwakke ondergrond en onaangepaste fundering ervoor dat de toren scheef begon te staan. Desondanks werd de bouw verdergezet. In de jaren 1960 was de toren al vervaarlijk aan het overhellen (zie foto) en werd gevreesd dat de 14 500 ton zware constructie zou omvallen. Er werd een team van experts samengesteld om de helling te verminderen en de toren te stabiliseren.

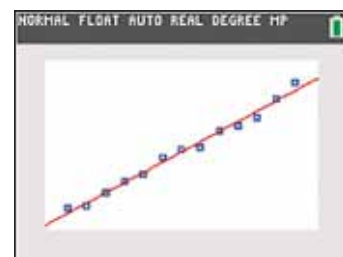
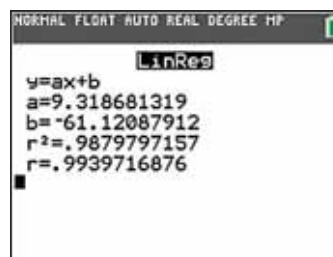
In de tabel hieronder is de verklarende variabele het aantal jaren ten opzichte van 1900. De responsvariabele is de 'overhelling': dit is de horizontale afstand tussen een punt op de toren en de positie van datzelfde punt mocht de toren verticaal staan. Om ervoor te zorgen dat deze overhelling niet afhankelijk is van de hoogte van de toren, werden de metingen uitgevoerd tussen twee vaste niveaus. In de tabel is de overhelling uitgedrukt in tienden van een mm ten opzichte van 2,9 m. De overhelling in 1975 was 2,9642 m, zodat in de tabel 642 staat.



jaren na 1900	75	76	77	78	79	80	81	82	83	84	85	86	87
overhelling	642	644	656	667	673	688	696	698	713	717	725	742	757

- 1 Bepaal een vergelijking van de kleinste kwadratenrechte en onderzoek grafisch of dit een goed model voor de gegevens oplevert.

Het spreidingsdiagram vertoont het patroon van een stijgende rechte. Er is een sterk positieve correlatie: $r = 0,994$. Een lineair model past goed bij de gegevens.



De regressierechte heeft vergelijking $y = 9,319x - 61,121$.

- 2 De ingenieurs die de Toren van Pisa bestudeerden waren echter niet geïnteresseerd in gegevens van het verleden, maar wensten te voorspellen wat er in de toekomst met de toren zou gebeuren. Wat vonden de ingenieurs in 1988 wanneer ze hun model extrapoleerden naar 1995?

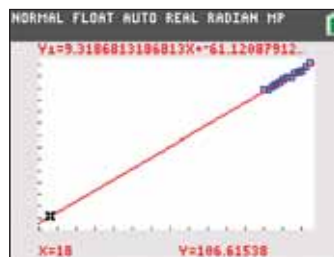
We zoeken de overhelling in 1995, dus bij $x = 95$, en vinden $y = 824$ (zie schermafdruk; bemerk dat de vensterinstellingen eerst aangepast moeten worden). Dit komt overeen met een overhelling van 2,9824 m.



Op de grafiek zien we dat de waarde ver buiten de punten ligt die voor het lineaire model gebruikt werden. Deze voorspelling dient dus met de nodige voorzichtigheid gebruikt te worden. Ze zal enkel betrouwbaar zijn indien de huidige trend zich onveranderd voortzet.

- 3 In 1918 was de overhelling 2,9071 m. Vergelijk deze opgemeten waarde met de waarde die het lineair model voorspelt en geef mogelijke verklaringen indien er een grote afwijking is.

Op dezelfde manier als hierboven vinden we, na het aanpassen van de vensterinstellingen, voor $x = 18$ een y -waarde van 107 of dus een overhelling van 2,9107 m.



Deze waarde wijkt 3,6 mm af van de toen opgemeten waarde. Een verklaring zou de nauwkeurigheid van de meting kunnen zijn. Een andere verklaring is dat het gebruikte model alleen betrouwbaar is bij interpolatie tussen 75 en 87. De waarde 18 ligt hier ver buiten.

Opdracht 18 bladzijde 179

Een belangrijke oorzaak van de globale opwarming van de aarde is ongetwijfeld de stijging van het gehalte van CO_2 in onze atmosfeer. Het Mauna Loa Observatory in Hawaii wordt beschouwd als een van de beste locaties om de luchtkwaliteit te meten, omdat mogelijke lokale beïnvloeding door vegetatie of menselijke activiteit op de CO_2 -concentratie er minimaal is. Omwille van de gunstige ligging, de continue monitoring en zorgvuldige selectie van de data beschouwt men deze gegevens als een betrouwbare indicator voor de concentratie van CO_2 in de middelste lagen van de troposfeer.

De eerste tabel toont de maandelijkse gemiddelde koolstofdioxideconcentraties, opgemeten in het Mauna Lao observatorium in 2016. De data geven de droge luchtfractie, gedefinieerd als het aantal moleculen CO_2 op het totaal aantal moleculen in de lucht, nadat de waterdamp werd verwijderd. Deze grootheid wordt uitgedrukt in ppm (parts per million). Zo wordt 0,000 400 in de tabel genoteerd als 400.

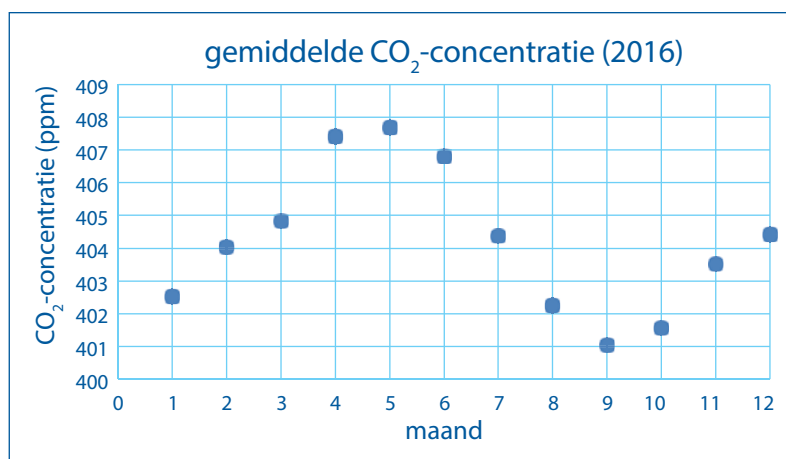
maand (2016)	1	2	3	4	5	6
CO ₂ (ppm)	402,52	404,04	404,83	407,42	407,70	406,81
maand (2016)	7	8	9	10	11	12
CO ₂ (ppm)	404,39	402,25	401,03	401,57	403,53	404,42

De tweede tabel bevat de jaarlijkse gemiddelde koolstofdioxideconcentraties van 1959 tot 2016. De jaartallen zijn geteld sinds 1950; zo komt 9 in de tabel overeen met 1959.

jaar	CO ₂ (ppm)	jaar	CO ₂ (ppm)	jaar	CO ₂ (ppm)	jaar	CO ₂ (ppm)	jaar	CO ₂ (ppm)	jaar	CO ₂ (ppm)
9	315,97	19	324,62	29	336,84	39	353,12	49	368,38	59	387,43
10	316,91	20	325,68	30	338,75	40	354,39	50	369,55	60	389,90
11	317,64	21	326,32	31	340,11	41	355,61	51	371,14	61	391,65
12	318,45	22	327,45	32	341,45	42	356,45	52	373,28	62	393,85
13	318,99	23	329,68	33	343,05	43	357,10	53	375,80	63	396,52
14	319,62	24	330,18	34	344,65	44	358,83	54	377,52	64	398,65
15	320,04	25	331,11	35	346,12	45	360,82	55	379,80	65	400,83
16	321,38	26	332,04	36	347,42	46	362,61	56	381,90	66	404,21
17	322,16	27	333,83	37	349,19	47	363,73	57	383,79		
18	323,04	28	335,40	38	351,57	48	366,70	58	385,60		

- 1 Onderzoek met behulp van een spreidingsdiagram het verband tussen het maandelijkse gemiddelde CO₂-gehalte en de maand, in 2016.

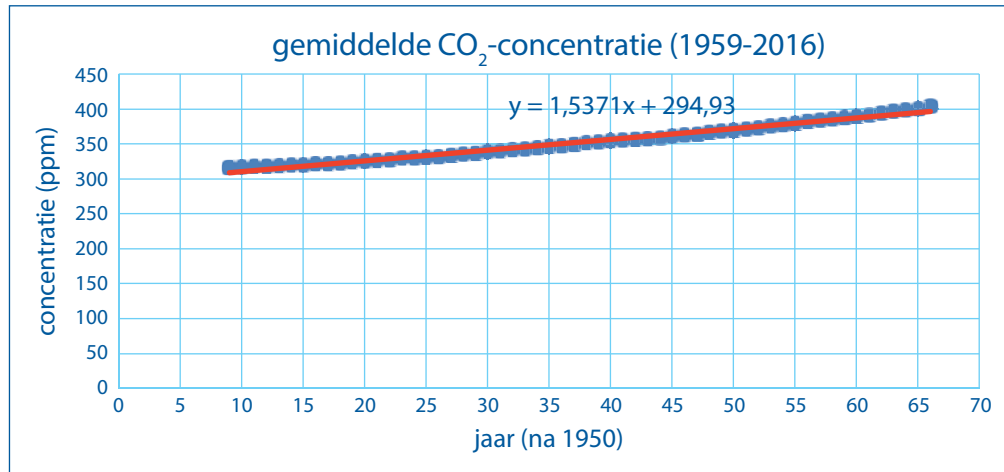
In 2016 is er geen lineair verband tussen het maandelijkse gemiddelde CO₂-gehalte en de maand. Het patroon is eerder sinusoidaal. Dit heeft te maken met seizoensgebonden veranderingen.



- 2 Maak een spreidingsdiagram van de jaarlijkse gemiddelde CO₂-concentraties en de tijd, van 1959 tot 2016. Is het verband lineair? Zo ja, wat is de correlatiecoëfficiënt?

De jaarlijkse gemiddelde concentratie tussen 1959 en 2016 lijkt wel lineair te verlopen.

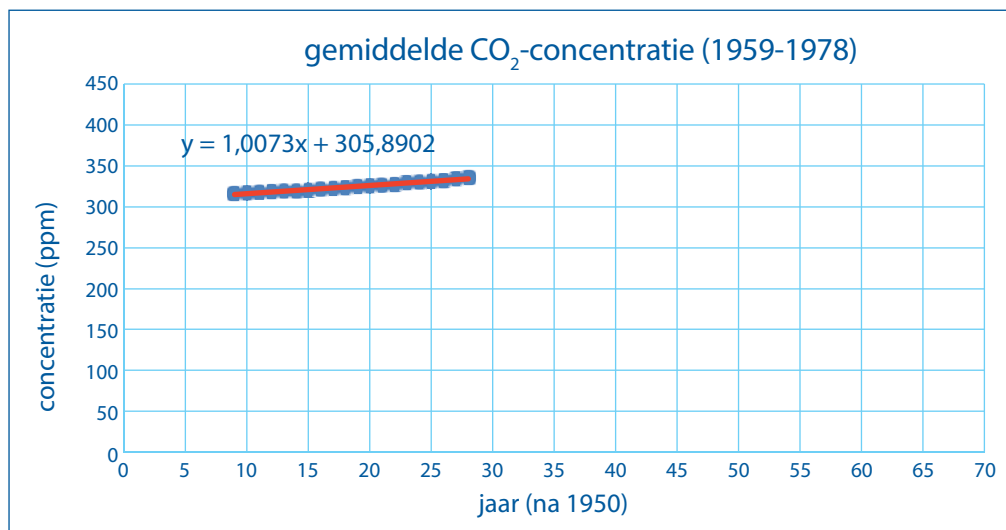
De aangroei snelheid lezen we af als richtingscoëfficiënt van de rechte. Ze is voor die periode gelijk aan 1,5371 ppm/jaar.

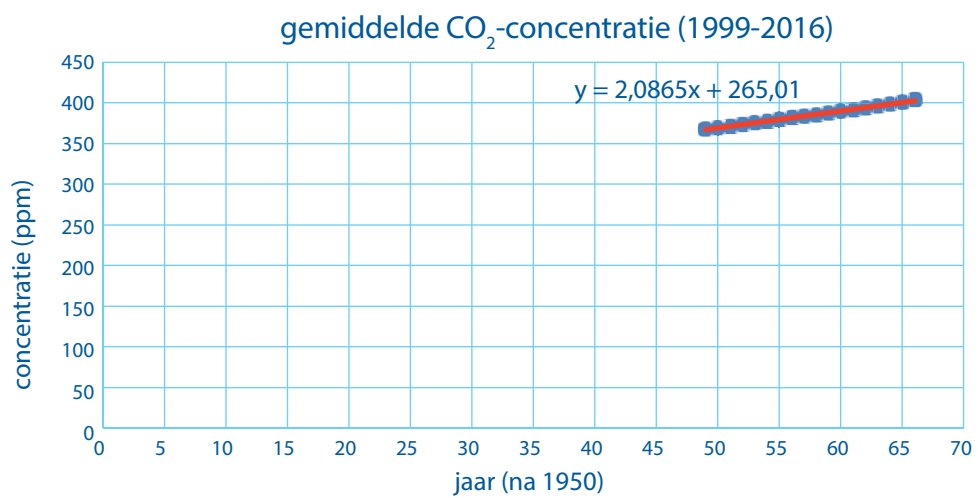
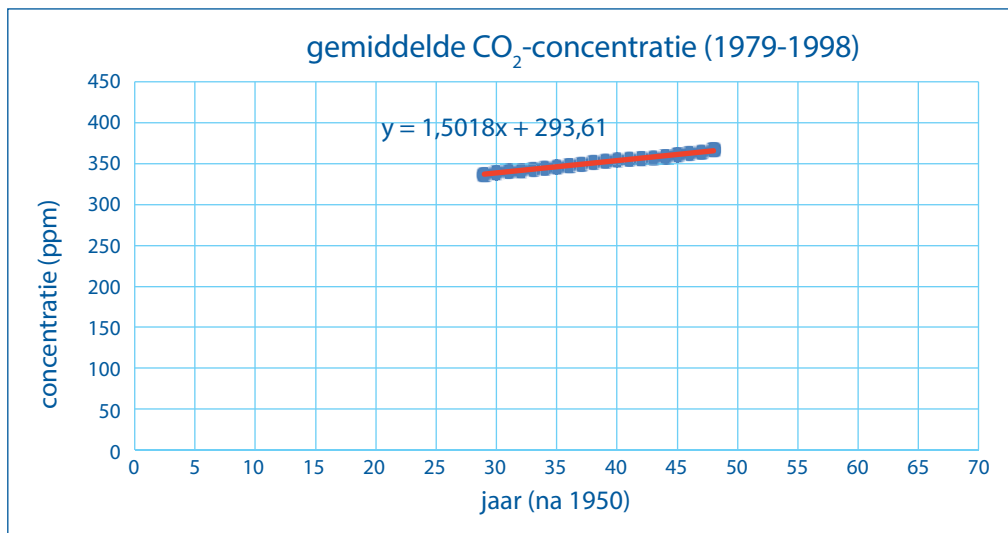


- 3 Uit een lineair model van de gegevens kun je de aangroei snelheid van het CO₂-gehalte halen. We delen de gegevens op in drie perioden 1959-1978, 1979-1998 en 1999-2016. Als het lineaire model goed is, zal de aangroei snelheid in de verschillende periodes ongeveer hetzelfde zijn. Vergelijk de aangroei snelheid in de drie periodes onderling en over de hele periode. Wat kun je hieruit besluiten?

De aangroei snelheid in de verschillende deelperiodes neemt duidelijk toe:

- van 1959 tot 1978: gemiddeld 1,0073 ppm/jaar;
- van 1978 tot 1998: gemiddeld 1,5018 ppm/jaar;
- van 1999 tot 2016: gemiddeld 2,0865 ppm/jaar.





Daaruit blijkt dat de concentratietoename versneld verloopt en dat het lineair model niet goed is. Vermoedelijk is een exponentieel model beter.

Opdracht 19 bladzijde 180

Het gezin Janssens wil de verwarmingskosten van hun huis optimaliseren en noteert daartoe een heel jaar hun energieverbruik, uitgedrukt in kWh. (Een kilowattuur, afgekort als kWh, is de energie die verbruikt wordt als men een machine met een vermogen van 1 kW gedurende 1 uur laat werken.)

Hun verbruik zal niet alleen bepaald worden door het aantal dagen dat ze moeten verwarmen, maar ook door hoe koud het dan is. Bij berekeningen in verband met energiegebruik in gebouwen wordt een grootheid gebruikt die beide aspecten combineert: het aantal verwarmingsgraaddagen (HDD, van *heating degree days*).

Hierbij wordt de buitentemperatuur vergeleken met een basistemperatuur, dit is de buitentemperatuur waarboven een gebouw geen verwarming nodig heeft. In België is de basistemperatuur 15,5 °C.

Voorbeeld HDD-berekening voor een bepaalde maand

We vergelijken op elke dag van de maand de opgemeten temperatuur met de basistemperatuur.

- 1ste dag: 14°, dus 1,5° onder de basistemperatuur: $1,5^\circ \cdot 1 \text{ dag} = 1,5 \text{ HDD}$.
- 2de dag: 13,5°, dus 2° onder de basistemperatuur: $2^\circ \cdot 1 \text{ dag} = 2 \text{ HDD}$.
- 3de dag: 16°, dus boven de basistemperatuur: $0^\circ \cdot 1 \text{ dag} = 0 \text{ HDD}$. Vermits de temperatuur hoger is dan de basistemperatuur moet er immers niet verwarmd worden.

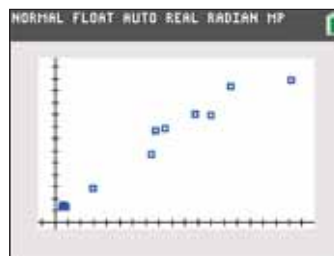
Op die manier wordt het aantal HDD van elke dag van die maand berekend. Het totaal (1,5 + 2 + 0 + ... HDD) geeft de waarde van die maand. Zo komt men bijvoorbeeld aan 21 HDD voor de maand augustus in de tabel hieronder.

De familie Janssens zoekt het aantal HDD voor het voorbije jaar op en vergelijkt dit met hun energieverbruik.

maand	8	9	10	11	12	1	2	3	4	5	6	7
HDD	21	17	185	299	337	454	269	194	211	74	14	13
kWh	131	143	573	888	1135	1189	897	765	778	279	134	125

- 1 Toon met een spreidingsdiagram aan dat er een positieve lineaire correlatie is tussen het gemiddeld aantal verwarmingsgraaddagen per maand en het maandelijks energieverbruik.

De puntenwolk toont een stijgende rechte, er is dus een positieve lineaire correlatie, met $r = 0,981$. In de schermafdruck hieronder is het aantal verwarmingsgraaddagen de verklarende variabele.



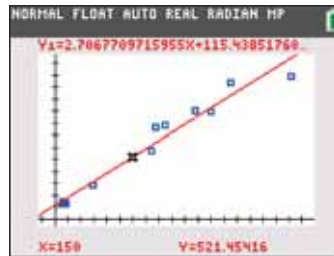
- 2 Stel op basis van de gegevens een lineair model op.

Een vergelijking van de regressierechte is $y = 2,707x + 115,439$.



- 3 In september van het volgende jaar verbruikt de familie 588 kWh. Op een gespecialiseerde website vinden ze 150 HDD voor deze maand. Hoeveel kWh verbruikten ze meer of minder dan wat het model voorspelt?

Aan de hand van het lineair model voorspellen we bij 150 HDD een energieverbruik van 521 kWh. Ze verbruikten dit jaar dus 67 kWh meer.



Opdracht 20 bladzijde 181

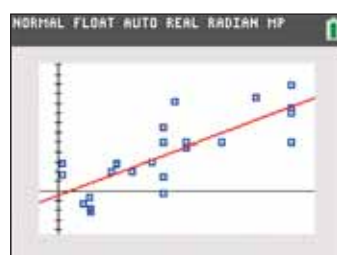


Georges Lemaître (1894–1966) was de grondlegger van de oerknaltheorie. In 1927 opperde hij voor het eerst de hypothese van het uitdijende heelal. Hij baseerde zich hiervoor op de algemene relativiteitstheorie van Einstein. Twee jaar later publiceerde Edwin Hubble een artikel over observaties van de afstand van sterrenstelsels tot de aarde en de snelheid waarmee deze stelsels zich van ons verwijderen. Hij had het verband onderzocht tussen de afstand r van het sterrenstelsel tot de waarnemer, uitgedrukt in megaparsec (Mpc), en de snelheid v (km/s) waarmee de sterrenstelsels zich van de waarnemer weg bewegen. Dit empirisch vastgesteld verband bevestigde de hypothese die Lemaître had geformuleerd.

- 1 De tabel hiernaast bevat de gegevens die Hubble gebruikte in zijn artikel van 1929.

Toon via een puntenwolk aan dat sterrenstelsels op grotere afstand zich sneller verwijderen. Hoe zou je deze correlatie beschrijven?

De puntenwolk verloopt stijgend. De correlatie is positief. Sterrenstelsels op grotere afstand van de aarde verwijderen zich dus sneller.



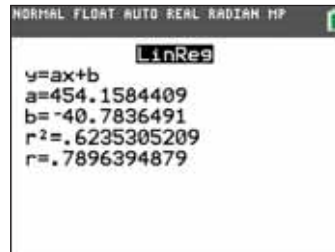
object-naam	afstand r (Mpc)	snelheid v (km/s)
S. Mag.	0,032	170
L. Mag.	0,034	290
6822	0,214	-130
598	0,263	-70
221	0,275	-185
224	0,275	-220
5457	0,45	200
4736	0,5	290
5194	0,5	270
4449	0,63	200
4214	0,8	300
3031	0,9	-30
3627	0,9	650
4826	0,9	150
5236	0,9	500
1068	1,0	920
5055	1,1	450
7331	1,1	500
4258	1,4	500
4151	1,7	960
4382	2,0	500
4472	2,0	850
4486	2,0	800
4649	2,0	1090

- 2 De 'wet van Hubble' stelt dat $v = H_0 \cdot r$, met H_0 een evenredigheidsconstante, die de constante van Hubble wordt genoemd.

Leid uit de vergelijking van de regressielijn een schatting voor H_0 uit die tijd af.

We voeren een lineaire regressie uit met x de afstand tot de aarde. De vergelijking van de regressierechte is dan $y = 454x - 41$. Hierin stelt y de snelheid voor waarmee de

sterrenstelsels zich van ons verwijderen. Hieruit volgt: $H_0 \approx 450 \frac{\text{km}}{\text{s} \cdot \text{Mpc}}$.



Opdracht 21 bladzijde 182

Stel dat we beschikken over n waarnemingen (x_i, y_i) , met $i = 1, 2, \dots, n$. Is $y = ax + b$ een lineair model voor deze gegevens, dan komt met elke x_i een voorspelde y -waarde $\hat{y}_i = ax_i + b$ (1) overeen. De kleinste kwadratenmethode houdt in dat we a en b zoeken zodat de totale kwadratische

voorspellingsfout $\sum_{i=1}^n (y_i - \hat{y}_i)^2$ (2) minimaal is.

- 1 Door in (2) de waarde \hat{y}_i te vervangen door (1), krijg je een uitdrukking in de variabelen a en b . Stel nu even dat a gekend is, dan bevat die uitdrukking enkel nog b als variabele. Je kunt nu onderzoeken voor welke b ze minimaal is met behulp van de afgeleide van die uitdrukking naar b .

Toon aan dat er een minimum bereikt wordt als $b = \bar{y} - a\bar{x}$ (3), met \bar{x} en \bar{y} het gemiddelde van de x - respectievelijk y -waarden.

We berekenen eerst de afgeleide naar b van $\sum_{i=1}^n (y_i - ax_i - b)^2$:

$$\begin{aligned} & \frac{d}{db} \left(\sum_{i=1}^n (y_i - ax_i - b)^2 \right) \\ &= \sum_{i=1}^n \frac{d}{db} ((y_i - ax_i - b)^2) \quad (\text{afgeleide van een som}) \\ &= \sum_{i=1}^n 2(y_i - ax_i - b) \cdot (-1) \quad (\text{afgeleide van een macht, kettingregel}) \\ &= -2 \left(\sum_{i=1}^n y_i - a \sum_{i=1}^n x_i - nb \right) \end{aligned}$$

Vervolgens zoeken we de waarde van b waarvoor deze uitdrukking 0 wordt.

$$-2 \left(\sum_{i=1}^n y_i - a \sum_{i=1}^n x_i - nb \right) = 0$$

$$\Leftrightarrow nb = \sum_{i=1}^n y_i - a \sum_{i=1}^n x_i$$

$$\Leftrightarrow b = \frac{1}{n} \sum_{i=1}^n y_i - a \frac{1}{n} \sum_{i=1}^n x_i$$

$$\Leftrightarrow b = \bar{y} - a\bar{x}$$

- 2 Toon nu met behulp van (1) en (3) aan dat de totale kwadratische voorspellingsfout (2)

geschreven kan worden als $a^2 \cdot \sum_{i=1}^n (x_i - \bar{x})^2 - 2a \cdot \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) + \sum_{i=1}^n (y_i - \bar{y})^2$ (4).

Gebruik je kennis over tweedegraadsfuncties om hieruit de waarde $a = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$ te

halen, waarvoor deze uitdrukking in a minimaal wordt.

$$\begin{aligned} \sum_{i=1}^n (y_i - \hat{y}_i)^2 &= \sum_{i=1}^n (y_i - ax_i - b)^2 = \sum_{i=1}^n (y_i - ax_i - \bar{y} + a\bar{x})^2 = \sum_{i=1}^n ((y_i - \bar{y}) - a(x_i - \bar{x}))^2 \\ &= \sum_{i=1}^n ((y_i - \bar{y})^2 - 2a(x_i - \bar{x})(y_i - \bar{y}) + a^2(x_i - \bar{x})^2) \\ &= a^2 \sum_{i=1}^n (x_i - \bar{x})^2 - 2a \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) + \sum_{i=1}^n (y_i - \bar{y})^2 \end{aligned}$$

We vinden een uitdrukking van de tweede graad in a . Deze bereikt een minimum in de top. (De coëfficiënt bij de tweedegraadsterm is een som van kwadraten en is dus positief, zodat er wel degelijk een minimum optreedt in de top.)

$$\text{De top wordt bereikt als } a = \frac{2 \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{2 \sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

- 3 Toon tot slot aan dat je de formule voor a die je in vorige deelvraag vond, met behulp van de

formule van Pearson voor de correlatiecoëfficiënt kunt herschrijven als $a = r \cdot \frac{s_y}{s_x}$.

Met $r = \frac{1}{n-1} \sum_{i=1}^n \frac{(x_i - \bar{x})(y_i - \bar{y})}{s_x s_y}$ en $s_x^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$ kunnen de teller resp. noemer van a

$$\text{herschreven worden: } a = \frac{r \cdot (n-1) \cdot s_x \cdot s_y}{(n-1) \cdot s_x^2} = r \cdot \frac{s_y}{s_x}.$$

Opdracht 22 bladzijde 182

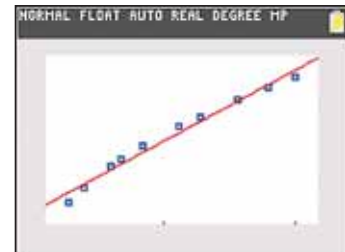
Een model beoordelen aan de hand van een residuendiagram

Stel dat $y = ax + b$ een vergelijking is van de regressierechte bij een reeks gegevens (x_i, y_i) met $i = 1, 2, \dots, n$. Voor elke x_i is $\hat{y}_i = ax + b$ de voorspelde y -waarde en $y_i - \hat{y}_i$ het bijbehorend residu. Een **residuendiagram** is een spreidingsdiagram van de residuen, afgezet tegen de verklarende variabele x .

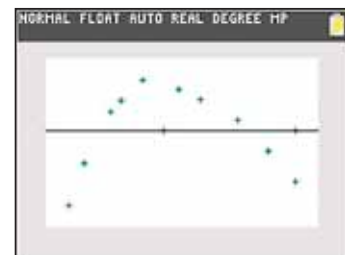
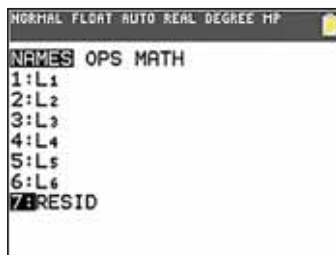
Met behulp van een residuendiagram kunnen we beter beoordelen hoe goed de regressierechte de gegevens modelleert.

Voorbeeld

In het spreidingsdiagram hiernaast is de verklarende variabele de lengte van een slinger (in m) en de responsvariabele de periode van een slingerbeweging (in s). De gegevens kunnen goed door een rechte gemodelleerd worden. De correlatiecoëfficiënt is ongeveer 0,99.

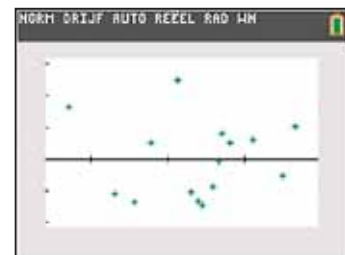


Bij het bepalen van de regressierechte berekent een rekentoestel automatisch ook de residuen en slaat die op in een aparte lijst. Voor de TI-84 heet die lijst RESID. Daarmee kun je snel een residuendiagram laten tekenen.



Uit het residuendiagram hierboven blijkt dat er een duidelijk patroon in de residuen zit. Dat wijst erop dat het lineair model wellicht niet het juiste is voor deze gegevens: voor kleine en grote x -waarden liggen de opgemeten y -waarden telkens onder de voorspelde y -waarden (de residuen zijn systematisch negatief) en voor de centrale x -waarden is het net andersom. Voor deze meetgegevens kunnen we beter naar een niet-lineair model zoeken.

Als er een duidelijk lineair verband is tussen de gegevens, zoals tussen de lengte en het gewicht van 15 baby's (zie opdracht 8), dan vertoont het residuendiagram geen herkenbaar patroon. De figuur hiernaast laat dit zien op het residuendiagram van de regressie van de lengte op het gewicht.



Algemeen

Verstoort een residuendiagram geen duidelijk patroon en liggen de residuen lukraak verspreid rond nul, het gemiddelde van de residuen, dan wijst dit erop dat de meetgegevens niet systematisch van het model afwijken. Het model slaagt er dan goed in het verband tussen de gegevens te beschrijven.

Zit er wel een patroon in de residuen, dan is de kwaliteit van het gebruikte model minder goed en is het aangewezen een beter model te zoeken.

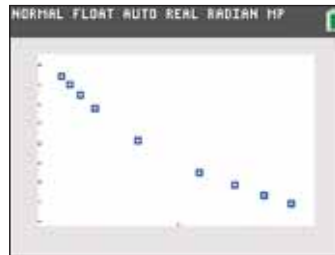
Pas dit nu zelf toe in de volgende situaties.

- 1 In een les fysica wordt bij een constante temperatuur het volume V (in cm^3) van een gas gewijzigd en wordt telkens de bijbehorende druk p (in hPa) opgemeten. Dit levert de onderstaande meetwaarden op.

$V (\text{cm}^3)$	19,1	18,0	17,3	16,7	22,1	26,5	29,0	31,1	33,0
$p (\text{hPa})$	1179	1247	1299	1343	1016	853	785	733	689

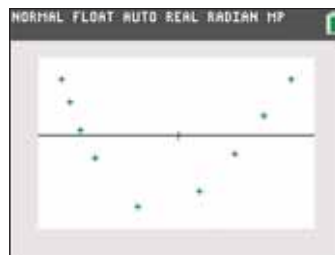
Onderzoek aan de hand van een residuendiagram of het verband tussen p en V goed door een eerstegraadsfunctie gemodelleerd wordt.

De puntenwolk sluit een lineair model niet meteen uit.



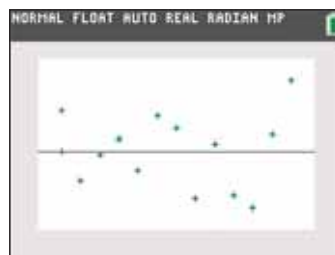
Het residuendiagram geeft een duidelijk patroon weer. Voor kleine en grote x-waarden zijn de residuen systematisch positief en dus liggen de opgemeten y-waarden boven de voorspelde waarden. Voor centrale x-waarden zijn de residuen systematisch negatief. Dat wijst erop dat het lineair model niet het juiste is voor deze gegevens.

Dit klopt met het feit dat druk en volume omgekeerd evenredig zijn (wet van Boyle-Mariotte). Het verband is dus hyperbolisch.



- 2 Onderzoek of het lineair model dat je in opdracht 17 gebruikte goed bij de gegevens past.

Het residuenpatroon vertoont geen duidelijk patroon. De residuen liggen lukraak verspreid rond 0, het gemiddelde van de residuen. Dat wijst erop dat de gegevens niet systematisch van het lineair patroon afwijken.



Opdracht 23 bladzijde 185

We beschikken over de lengtes van 30 vaders en hun volwassen oudste zoon.

lengte vader (cm)	182	176	176	170	175	165	174	177	180	190
lengte zoon (cm)	186	176	181	176	183	169	182	190	189	198

lengte vader (cm)	184	179	166	190	187	176	189	192	167	171
lengte zoon (cm)	191	187	171	202	189	192	201	195	182	184

lengte vader (cm)	172	173	174	172	178	172	178	177	168	170
lengte zoon (cm)	167	177	184	183	189	184	187	189	170	176

De vader van Ruben is 198 cm. Voorspel de lengte van Ruben aan de hand van een lineair model voor het verband tussen beide lengtes.

We zetten de lengte van de vaders op de horizontale as.



Een vergelijking van de regressierechte is $y = 1,030x + 2,366$. Bij een vader van 198 cm voorspelt het model voor de zoon een lengte van ongeveer 206 cm.

Opdracht 24 bladzijde 185

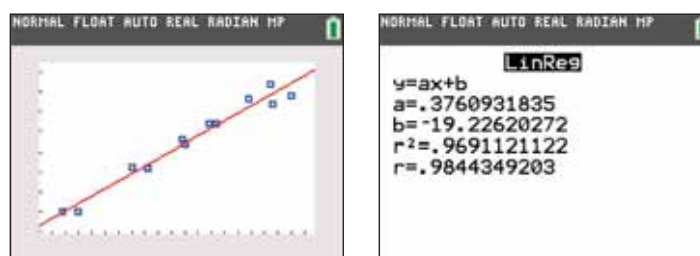
Het menselijk lichaam verbruikt meer zuurstof bij een inspanning dan in rust. Om de spieren van zuurstof te voorzien, moet het hart sneller kloppen. De zuurstofopname wordt genoteerd als VO_2 (van volume O_2) en heeft als eenheid ml/min/kg (milliliter per minuut en per kg lichaamsgewicht); de hartslag noteren we hieronder als HS en we drukken die uit in slagen per minuut (sl./min). Omdat de zuurstofopname moeilijk te meten is en de hartslag relatief eenvoudig, zal men bij sporters een model maken waarmee men uit de hartslag de zuurstofopname bij inspanning kan voorspellen.

Pascal is hardloper en wil laten opmeten hoeveel zuurstof hij opneemt in functie van zijn hartslag. Op basis van twee inspanningstesten beschikt hij over de volgende gegevens.

HS (sl./min)	104	130	149	159	174	182
VO_2 (ml/min/kg)	19,95	31,16	37,94	41,98	48,24	51,89

HS (sl./min)	110	136	150	162	183	190
VO_2 (ml/min/kg)	20	31	37	42	47	49

Stel een model op waarmee hij bij een inspanning vanuit zijn hartslag zijn zuurstofopname kan voorspellen.



We voeren een lineaire regressie uit met x de hartslag en vinden: $y = 0,376x - 19,226$.

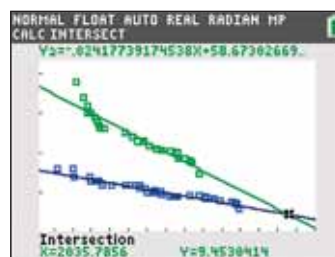
Opdracht 25 bladzijde 186

Het IAAF (International Association of Athletics Federations) werd in 1912 opgericht met als een van de doelen lijsten bij te houden van officiële wereldrecords. In de tabel zie je de Wereldrecordgegevens 100 m sprint mannen en vrouwen.

Wereldrecord 100 m sprint mannen								
jaar	tijd (s)	naam	jaar	tijd (s)	naam	jaar	tijd (s)	naam
1912	10,6	D. Lippincott	1954	10,2	H. Futterer	1983	9,93	C. Smith
1920	10,6	J. Scholz	1956	10,2	B. Morrow	1988	9,92	C. Lewis
1921	10,4	C. Paddock	1956	10,1	W. Williams	1991	9,90	L. Burrell
1929	10,4	E. Tolan	1959	10,1	R. Norton	1991	9,86	C. Lewis
1930	10,3	P. Williams	1960	10,0	A. Hary	1994	9,85	L. Burrell
1932	10,3	E. Tolan	1964	10,0	H. Esteves	1996	9,84	D. Bailey
1933	10,3	R. Metcalfe	1964	10,06	B. Hayes	1999	9,79	M. Greene
1934	10,3	E. Peacock	1967	10,03	J. Hines	2005	9,77	A. Powell
1934	10,3	C. Berger	1968	9,95	J. Hines	2007	9,74	A. Powell
1936	10,2	J. Owens	1972	9,9	E. Hart	2008	9,72	U. Bolt
1941	10,2	H. Davis	1974	9,9	S. Williams	2008	9,69	U. Bolt
1948	10,2	L. Labeach	1975	9,9	S. Leonard	2009	9,58	U. Bolt
1951	10,2	M. Bailey	1976	9,9	S. Williams			

Wereldrecord 100 m sprint vrouwen								
jaar	tijd (s)	naam	jaar	tijd (s)	naam	jaar	tijd (s)	naam
1922	12,8	M. Lines	1935	11,6	H. Stephens	1972	11,07	R. Stecher
1926	12,4	G. Wittmann	1937	11,6	S. Walasiewicz	1976	11,04	I. Helten
1928	12,2	K. Hitomi	1948	11,5	F. Blankers	1976	11,01	A. Richter
1928	12,0	M. Cook	1952	11,4	M. Jackson	1977	10,88	M. Oelsner
1930	12,0	T. Schuurman	1955	11,3	S. Strickland	1980	10,87	L. Kondratjeva
1932	11,9	T. Schuurman	1958	11,3	V. Krepkina	1983	10,81	M. Gohr
1932	11,9	S. Walasiewicz	1961	11,2	W. Rudolf	1983	10,79	E. Ashford
1933	11,8	S. Walasiewicz	1965	11,1	I. Kirszenstein	1984	10,76	E. Ashford
1934	11,7	S. Walasiewicz	1968	11,08	W. Tyus	1988	10,49	F. Griffith

- 1 Maak een lineair regressiemodel voor de wereldrecords 100 m sprint mannen uitgezet tegen het jaartal waarin het wereldrecord werd gelopen.
- 2 Doe hetzelfde voor de wereldrecords 100 m sprint vrouwen.
- 3 Maak een prognose over wanneer de vrouwen de mannen 'inhalen'.



Indien de huidige trends zich verderzetten, zullen volgens de lineaire modellen in het jaar 2035 de vrouwen (groene grafiek) de mannen (blauwe grafiek) inhalen en zullen beide wereldrecords op ongeveer 9,45 s komen te staan.

Opdracht 26 bladzijde 187

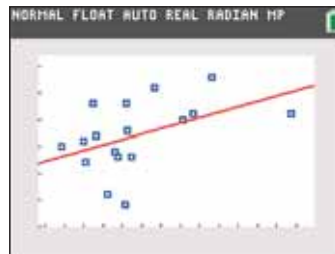
Uit het klimatologisch jaaroverzicht van het KMI van 2000 tot 2016¹⁶ haalden we de volgende gegevens. De eerste lijn bevat het jaartal, op de lijn daaronder staat het aantal uren zonneshijn in Ukkel gedurende dat jaar en op de volgende lijn staat de gemiddelde maximumtemperatuur T uitgedrukt in graden Celsius.

jaar	2000	2001	2002	2003	2004	2005	2006	2007	2008
zon (u)	1392	1455	1480	1987	1537	1563	1559	1472	1449
$T(^{\circ}\text{C})$	14,5	14,2	14,7	15,1	14,3	14,8	15,3	15,3	14,6

jaar	2009	2010	2011	2012	2013	2014	2015	2016
zon (u)	1705	1556	1782	1529	1510	1634	1734	1572
$T(^{\circ}\text{C})$	15,0	13,4	15,8	14,4	13,6	15,6	15,1	14,3

- 1 Ga na dat het verband tussen het jaarlijkse aantal uren zonneshijn en de gemiddelde maximumtemperatuur lineair is.

Het verband is zwak lineair: de puntenwolk ligt sterk verspreid rond een stijgende rechte.



- 2 Bereken de bijbehorende correlatiecoëfficiënt.

Correlatiecoëfficiënt: $r = 0,4621$.



- 3 Je kunt de temperatuur ook uitdrukken in graden Fahrenheit. Dat doe je met de omzettingsformule: $^{\circ}\text{F} = ^{\circ}\text{C} \cdot 1,8 + 32$.

Wat is nu de correlatiecoëfficiënt?

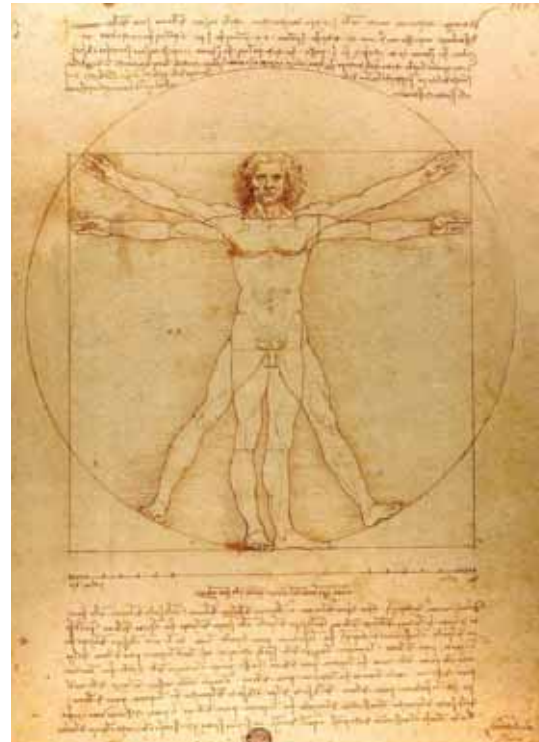
De correlatiecoëfficiënt verandert niet: hij is immers onafhankelijk van de gebruikte eenheid, omdat hij op basis van z-scores van de meetwaarden wordt berekend.

Opdracht 27 bladzijde 187

Leonardo da Vinci gebruikte in zijn beroemde Vitruviusman de afmetingen zoals beschreven door Vitruvius: "... erit eaque mensura ad manus pansas" (de lengte van de uitgestrekte armen is gelijk aan de lichaamslengte). Hij schreef het bij zijn tekening in spiegelschrift.

Leonardo Da Vinci controleerde zijn bevindingen door zelf een aantal mannen op te meten. Recentere metingen toonden echter aan dat bij ongeveer 60 % van de volwassen mannen en bij ongeveer 20 % van de vrouwen de spanwijdte groter is dan de lengte.

Wij willen de lengte van een 17-jarige voorspellen uit de spanwijdte aan de hand van de volgende tabel van 18 jongeren van 17 jaar.



armwijdte (cm)	188	183	183	168	198	171	156	183	173
lengte (cm)	193	188	180	173	196	173	164	177	179
armwijdte (cm)	174	159	166	175	178	160	175	168	200
lengte (cm)	174	163	164	173	173	169	178	166	192

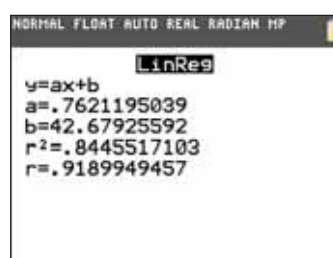
- 1 Maak een spreidingsdiagram. Bespreek vorm, richting en sterkte van de correlatie.

Het spreidingsdiagram vertoont een sterke positieve lineaire samenhang.



- 2 Bereken de correlatiecoëfficiënt en een vergelijking van de regressierechte.

De correlatiecoëfficiënt is 0,919 en een vergelijking van de regressierechte is $y = 0,76x + 42,68$.



- 3 Hoe zou je grafisch kunnen nagaan voor hoeveel proefpersonen de spanwijdte groter is dan de lengte?

Alle punten die onder de eerste bissectrice liggen, hebben een spanwijdte die groter is dan de bijbehorende lichaamslengte. De eerste bissectrice is in het grijs weergegeven in de onderstaande schermafdruk; er zijn 8 personen met grotere spanwijdte.

