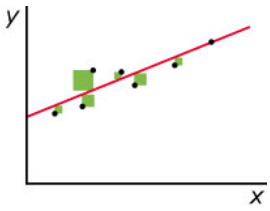
# 6.6 Lineaire regressie

## Inleiding

Naast het toetsen van hypothesen is in de mathematische statistiek het onderzoeken naar statistische verbanden een belangrijke tak van sport: wanneer bestaat er een verband tussen twee statistische variabelen? En kun je dan met zo'n verband tussen twee variabelen ook voorspellingen doen? Met andere woorden kun je een formule vinden die het verband beschrijft? Met de 'kleinste kwadraten methode' kun je bij vrijwel elke puntenwolk wel een (soms kromme) lijn vinden die het verloop beschrijft. Maar hoe zinnig is dat? In dit onderdeel bekijk je alleen lineaire verbanden.



Figuur 1

**Je leert in dit onderwerp**

- bij een puntenwolk bij twee statistische variabelen een regressielijn te tekenen en er een formule van te vinden;
- te bekijken hoe zinvol die formule is en welke voorspellingen je er mee kunt doen.

**Voorkennis**

- de elementaire statistische begrippen (gemiddelde, standaardafwijking, e.d.) gebruiken;
- spreidingsdiagrammen tekenen in de vorm van een puntenwolk en daarbij de correlatiecoëfficiënt berekenen.

## Verkennen

### Opgave V1

Om te onderzoeken of er een verband bestaat tussen lengte en gewicht bij mensen van 15 tot 17 jaar oud heb je gegevens nodig. Op het werkblad **LengteGewicht22h4.xls** vind je de gegevens van een 4HAVO-klas van 22 leerlingen. Bekijk het getekende spreidingsdiagram. De correlatiecoëfficiënt bedraagt ongeveer 0,81, dus er bestaat een lineair statistisch verband tussen lengte en gewicht.

Kun je een formule voor dit verband opstellen? En hoe doe je dit dan?

## Uitleg

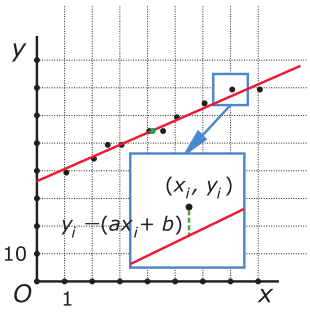
Als er tussen twee variabelen $x$ en $y$ een goede correlatie bestaat, bestaat er een lineair (statistisch) verband tussen. Maar hoe stel je daarbij een formule op? Een regressielijn moet uiteraard door het punt $(\overline{x},\overline{y})$ gaan. De richtingscoëfficiënt (het hellingsgetal) van die lijn kun je op dit moment echter alleen nog maar schatten.

De beroemde wiskundige **Carl Friedrich Gauss** bedacht daarvoor in de negentiende eeuw de 'methode van de kleinste kwadraten'. Stel je voor dat je een regressielijn wilt maken van de vorm $y = a \cdot x + b$. Je gaat dan uit van een regressielijn van $y$ op $x$.
Gauss' methode houdt nu in dat de som van de kwadraten van de verticale afwijkingen van de meetpunten tot deze regressielijn zo klein mogelijk moet zijn. Dat betekent dat

$$\sum_{i=1}^{n} (g_i - (a \cdot l_i + b))^2$$

minimaal moet zijn. Gauss vond dat dit het geval is als

$$a = \frac{\sum_{i=1}^{n} (l_i - \overline{l})(g_i - \overline{g})}{N \cdot \sigma_x^2}$$

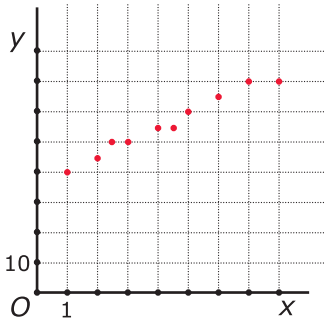


Figuur 2

Deze formule lijkt erg op die van de correlatiecoëfficiënt. In feite is $a = r_{xy} \cdot \frac{\sigma_y}{\sigma_x}$. En hiermee heb je een snelle manier gevonden om het hellingsgetal $a$ te vinden.

### Opgave 1

Bekijk dit spreidingsdiagram.



Figuur 3

**a** Maak een tabel van de 10 meetpunten. Voer deze gegevens in je grafische rekenmachine in.

**b** Bereken de coördinaten van het punt $(\overline{x},\overline{y})$.

**c** Als je door deze punten 'op het oog' een regressielijn zou willen tekenen, hoe groot wordt dan de richtingscoëfficiënt ongeveer?

**d** Bereken nu de correlatiecoëfficiënt en de richtingscoëfficiënt van de regressielijn.

**e** Stel een vergelijking op van de regressielijn van $y$ op $x$.

**f** Welke waarde zou $y$ moeten hebben volgens deze regressielijn als $x = 10$?

### Opgave 2

Lees in de **Uitleg** na hoe Gauss de methode van de kleinste kwadraten gebruikte om de richtingscoëfficiënt van de regressielijn te berekenen.

**a** Laat zien (door haakjes uitwerken) dat $p = \sum_{i=1}^{n} (g_i - (a \cdot l_i + b))^2$ een kwadratische functie van $a$ is.

**b** Bereken voor welke waarde van $a$ deze functie minimaal is en leidt zo de formule voor $a$ zelf af.

**c** Leg ook uit hoe je aan de formule komt waarmee je $a$ kunt berekenen vanuit $r_{xy}$.

## Theorie en voorbeelden

### Om te onthouden

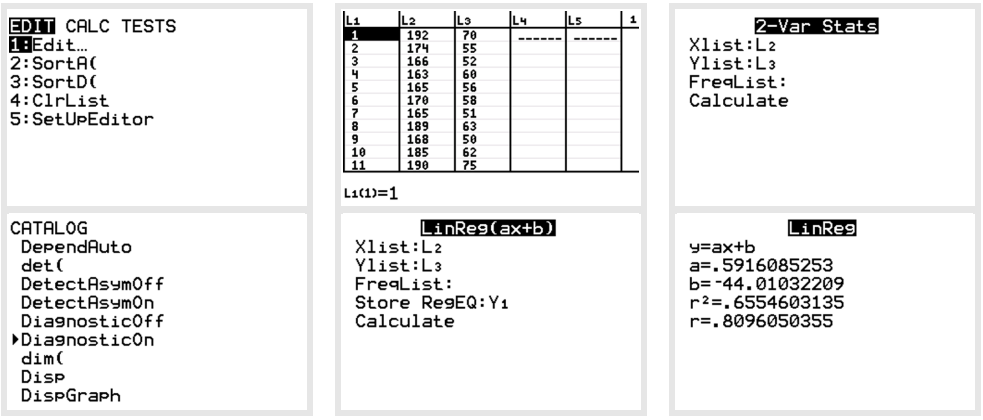
Als de correlatie tussen de variabelen $x$ en $y$ groot genoeg is, kun je een formule van de vorm $y = ax + b$ opstellen die het verband tussen $x$ en $y$ weergeeft. Deze formule heeft als grafiek een rechte lijn, de **regressielijn** of **trendlijn** van $y$ op $x$. Zo'n regressielijn gaat door het punt $(\overline{x},\overline{y})$ en heeft als richtingscoëfficiënt (hellingsgetal):

$$a = r_{xy} \cdot \frac{\sigma_y}{\sigma_x}$$

Deze richtingscoëfficiënt heet wel de **regressiecoëfficiënt** van $y$ op $x$. Met behulp van deze regressiecoëfficiënt en het feit dat de regressielijn door $(\overline{x},\overline{y})$ gaat, kun je de bijbehorende formule opstellen.

## Voorbeeld 1

Op het werkblad **LengteGewicht22h4.xls** vind je de gegevens van een 4HAVO-klas van 22 leerlingen. Je kunt deze gegevens ook in de grafische rekenmachine invoeren en die de regressielijn laten berekenen. In de figuren hieronder zie je hoe dit op de TI84 in zijn werk gaat. Je hebt er het rekenalgoritme LinReg voor nodig. Dat staat voor 'lineaire regressie' en wordt in het volgende onderdeel nader bekeken. In het **Practicum** zie je hoe dit op de diverse rekenmachines gaat.




Figuur 4

### Opgave 3

Bekijk **Voorbeeld 1**.

**a** Voer de berekening van de regressielijn bij de gegevens van de 4HAVO-klas zelf uit met behulp van de grafische rekenmachine.

**b** Welke betekenis heeft deze regessielijn als je aanneemt dat de groep leerlingen voldoende representatief is voor alle 15-17 jarigen?

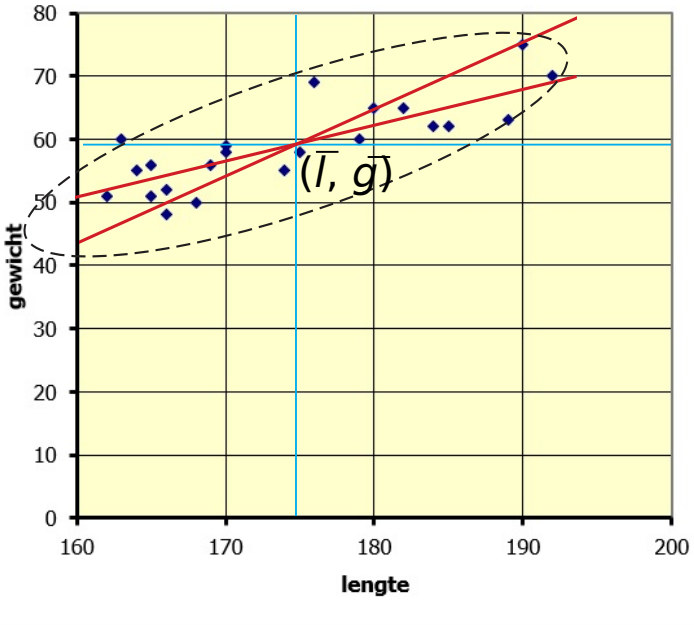
**c** Hoe zwaar zou iemand van 16 jaar moeten zijn als hij 180 cm lang is?

## Voorbeeld 2

**Onderzoek naar verband tussen lengte en gewicht 15-17 jarigen**

**Gegevens:**

| leerling | m/v | lengte (cm) | gewicht (kg) |
|---|---|---|---|
| 1 | m | 192 | 70 |
| 2 | m | 174 | 55 |
| 3 | v | 166 | 52 |
| 4 | v | 163 | 60 |
| 5 | v | 165 | 56 |
| 6 | v | 170 | 58 |
| 7 | v | 165 | 51 |
| 8 | m | 189 | 63 |
| 9 | v | 168 | 50 |
| 10 | m | 185 | 62 |
| 11 | m | 190 | 75 |
| 12 | v | 169 | 56 |
| 13 | m | 176 | 69 |
| 14 | m | 180 | 65 |
| 15 | m | 166 | 48 |
| 16 | v | 170 | 59 |
| 17 | m | 179 | 60 |
| 18 | m | 175 | 58 |
| 19 | v | 162 | 51 |
| 20 | v | 164 | 55 |
| 21 | m | 182 | 65 |
| 22 | m | 184 | 62 |




Figuur 5

Bij het verband tussen de lengte $l$ en het gewicht $g$ bij de groep van 22 leerlingen in het bestand **LengteGewicht22h4.xls** heb je een regressielijn van $g$ op $l$ gemaakt: $g = 0{,}59 \cdot l - 44{,}01$.

Er past echter ook heel goed een regressielijn van $l$ op $g$ bij. Ga na, dat je dan vindt: $l = 1{,}11 \cdot g + 108{,}80$.
Deze tweede regressielijn kun je in dezelfde figuur tekenen als de eerste. Deze twee regressielijnen zijn verschillend!

Als je van een leerling van 15-17 jaar met een lengte van $l = 180$ cm het gewicht zou moeten voorspellen, vind je volgens de eerste regressielijn ongeveer 62,19 kg, maar volgens de tweede regressielijn hoort bij een gewicht van 62,19 kg een lengte van 177,83 cm!
Dit verschil is het **regressie-effect**.
Dat regressie-effect ontstaat doordat er geen volledige correlatie tussen $g$ en $l$ is, de correlatiecoëfficiënt is 'slechts' ongeveer 0,81 en dat is minder dan 1.

### Opgave 4

In **Voorbeeld 2** zie je dat er twee regressielijnen kunnen worden gemaakt bij elk verband tussen twee variabelen.

**a** Bereken zelf ook de regressielijn van $l$ op $g$.

**b** Bereken bij beide regressielijnen het gewicht dat zou moeten horen bij een 15-17 jarige die precies één standaardafwijking groter is dan de gemiddelde lengte.

**c** Wijkt dit gewicht meer of minder dan één standaardafwijking van het gemiddelde af? Geef voor beide regressielijnen antwoord op deze vraag.

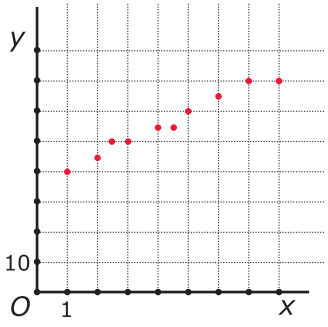
### Opgave 5

Laat zien dat het product van de twee regressiecoëfficiënten precies het kwadraat van de correlatiecoëfficiënt is.

## Verwerken

### Opgave 6

Bekijk dit spreidingsdiagram uit de eerste opgave bij de **Uitleg** nog eens.



Figuur 6

**a** Stel een formule op voor de regressielijn van $x$ op $y$.

**b** Teken zelf het spreidingsdiagram met daarin beide regressielijnen.

**c** Is er sprake van een regressie-effect? Zo ja, laat dit dan met een rekenvoorbeeld zien.

### Opgave 7

Om te onderzoeken of er enig verband bestaat tussen de lengte van een vader en die van zijn zoon zijn de lengtes van 12 vaders en die van hun oudste zoons gemeten op het moment dat die zoons volwassen werden. De gegevens staan in deze tabel.

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| lengte vader $v$ in cm | 173 | 168 | 178 | 170 | 180 | 165 | 185 | 175 | 180 | 178 | 183 | 188 |
| lengte zoon $z$ in cm | 180 | 175 | 180 | 173 | 183 | 175 | 180 | 173 | 188 | 178 | 180 | 185 |

Tabel 1

**a** Was er sprake van een positieve of een negatieve correlatie? Wat betekent dit in de praktijk?

**b** Stel de regressielijn op van $z$ op $v$ bij deze gegevens.

**c** Als een bepaalde vader 1,77 m lang is, hoe lang zou dan zijn oudste zoon moeten zijn?

**d** Wat betekent het optredende regressie-effect voor de bepaling van de lengte van een zoon waarvan de vader bijvoorbeeld 2 m lang is?

## Opgave 8

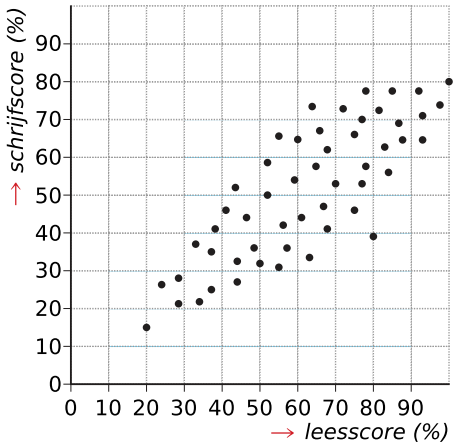
Een basisschool heeft een leestest en schrijftest Nederlands afgenomen bij de leerlingen in groep acht. De resultaten zijn verwerkt in een puntenwolk.

Er lijkt een verband te zijn tussen de schrijfscore $S$ en de leesscore $L$.

**a** Stel een formule op voor de trendlijn die het verband tussen $S$ en $L$ weergeeft.

**b** Geef met behulp van de formule uit a een schatting van de schrijfscore bij een leesscore van 80%.

**c** Geef met behulp van de formule uit a een schatting van de leesscore bij een schrijfscore van 10%.




Figuur 7

## Opgave 9

In de tabel vind je het aantal inwoners $N$ in een bepaalde stad.

| Jaartal | 1960 | 1970 | 1980 | 1990 | 2000 |
|---|---|---|---|---|---|
| Aantal inwoners $N$ | 23.107 | 25.880 | 28.985 | 32.479 | 36.358 |

Tabel 2

Er wordt aangenomen dat $N$ een exponentiële functie is van $t$, de tijd in jaren met $t = 0$ in 1960.

**a** Maak het spreidingsdiagram van $\log(N)$ afhankelijk van $t$.

**b** Bereken de correlatiecoëfficiënt van $\log(N)$ en $t$.

**c** Voorspel met behulp van de regressielijn van $\log(N)$ op $t$ het aantal inwoners in 2010 en 2020.

**d** Waarom is er vrijwel geen regressie-effect?

## Opgave 10

Om het verband tussen het gewicht $G$ (in pounds) en de braadtijd voor kalkoenen te onderzoeken, werd onder gelijke omstandigheden nagegaan hoeveel minuten $t$ het duurde tot het binnenste van een kalkoen de temperatuur van 85 °C bereikte. Er werden diverse kalkoenen aan dit onderzoek onderworpen. Ze hadden een gemiddeld gewicht van 15,24 pounds met een standaardafwijking van 6,07. Voor de waarden van $t$ vonden de onderzoekers een gemiddelde van 205,4 minuten met een standaardafwijking van 59,1.
De regressielijn van $t$ op $G$ had de vergelijking: $t = 9{,}65G + 58{,}40$.

Hoeveel bedroeg de correlatiecoëfficiënt?

## Opgave 11

In 1947 hielden de wiskundigen Freudenthal en Sittig een statistisch onderzoek ten behoeve van een nieuw maatsysteem voor vrouwenkleding in opdracht van het warenhuis De Bijenkorf. Zij lieten daarbij een grote verscheidenheid aan lichaamsmaten opmeten van 5001 vrouwen. In het bestand **StatFS-Bijenkorf1947.xls** vind je enkele gegevens.
Gebruik de werkbladen ‘mouwlengte en kniehoogte’ en ‘mouwlengte-kniehoogte’.
Op het werkblad ‘mouwlengte-kniehoogte’ zie je een zogenaamde kruistabel waarin de combinaties mouwlengte-kniehoogte zijn weergegeven. De hierbij gevonden correlatiecoëfficiënt is ongeveer 0,6271.

**a** Bereken op het werkblad ‘mouwlengte-kniehoogte’ de standaardafwijkingen van beide statistische variabelen.

**b** Stel vergelijkingen op van de beide bijbehorende regressielijnen met de constanten in twee decimalen nauwkeurig.

**c** Bereken met behulp van deze regressielijnen de gemiddelde kniehoogte van een vrouw met een mouwlengte van 60 cm. Is er sprake van een groot regressie-effect?

# Toepassen

## Opgave 12: Vliegsnelheid en lichaamslengte (vervolg)

Biologen veronderstellen op grond van metingen dat er bij vliegende dieren een verband bestaat tussen de lichaamslengte $L$ (in cm) en de vliegsnelheid $v$ (in cm/s).

| **Vliegsnelheid en lichaamslengte bij verschillende dieren** | | | |
|---|---|---|---|
| Soort | | Lengte $L$ in cm | Vliegsnelheid $v$ in cm/s |
| 1. | *Drosophila melanogaster* (fruitvlieg) | 0,2 | 190 |
| 2. | *Tabanus affinis* (paardenvlieg) | 1,3 | 660 |
| 3. | *Archilochus colubris* (kolibriesoort) | 8,1 | 1120 |
| 4. | *Anax sp.* (waterjuffer) | 8,5 | 1000 |
| 5. | *Eptesicus fuscus* (grote bruine vleermuis) | 11,0 | 690 |
| 6. | *Phylloscopus trochilus* (fitis) | 11,0 | 1200 |
| 7. | *Apus apus* (gierzwaluw) | 17,0 | 2550 |
| 8. | *Cypselurus cyanopterus* (vliegende vis) | 34,0 | 1560 |
| 9. | *Numenius phaeopus* (regenwulp) | 41,0 | 2320 |
| 10. | *Anas acuta* (pijlstaarteend) | 56,0 | 2280 |
| 11. | *Olor columbianus bewicki* (kleine zwaan) | 120,0 | 1880 |
| 12. | *Pelecanus onocrotalus* (witte pelikaan) | 160,0 | 2280 |

Tabel 3

**a** Bekijk het spreidingsdiagram voor $\log(L)$ en $\log(v)$ dat je in het vorige onderdeel hebt gemaakt.

**b** Bereken de regressiecoëfficiënt van $\log(v)$ op $\log(L)$.

**c** Er bestaat tussen $L$ en $v$ dus een verband van de vorm $\log(v) = a \cdot \log(L) + b$. Laat zien dat dit betekent dat $v$ een machtsfunctie is van $L$ en stel een formule voor die machtsfunctie op.

# Testen

## Opgave 13

In een Amerikaans laboratorium heeft men proeven genomen waarbij gelet werd op het verband tussen de hoogte van de bewaartemperatuur $F$ in graden Fahrenheit en de werkzaamheid $W$ van een bepaald geneesmiddel. Bij temperaturen van 30°, 50°, 70° en 90° (Fahrenheit) werden drie porties van gelijk gewicht uit eenzelfde productie 20 dagen bewaard. Na deze periode werd op identieke wijze de werkzaamheid van de porties vastgesteld. De werkzaamheid werd uitgedrukt in percentages van de werkzaamheid zoals die was voor het bewaren.

| Bewaartemperatuur $F$ (°F) | 30 | 50 | 70 | 90 |
|---|---|---|---|---|
| Werkzaamheid $W$ (%) | 39, 42, 35 | 32, 26, 33 | 19, 27, 23 | 14, 19, 21 |

Tabel 4

**a** Verwerk deze gegevens in een spreidingsdiagram en bereken de correlatiecoëfficiënt. Is er sprake van een correlatie tussen $W$ en $F$?

**b** Stel de vergelijking op van de regressielijn van $W$ op $F$. Waarom ligt deze regressielijn meer voor de hand dan die van $F$ op $W$?

Het verband tussen de temperatuur in graden Fahrenheit $F$ en die in graden Celsius $C$ wordt zoals bekend gegeven door: $F = 1{,}8C + 32$.

**c** Stel nu een vergelijking op van de regressielijn van $W$ op $C$.

**d** Is de correlatiecoëfficiënt tussen $W$ en $C$ anders dan die tussen $W$ en $F$? Verklaar je antwoord.

Uit andere experimenten is gebleken dat de werkzaamheid bij een vaste bewaartemperatuur exponentieel afhangt van de lengte van de bewaarperiode.

**e** Schat de gemiddelde werkzaamheid van porties die 40 dagen bij een temperatuur van 20 °C zijn bewaard.

## Practicum

Met deze practica leer je hoe je de **de trendlijn** met de grafische rekenmachine tekent en berekent.

- **Trendlijn, correlatie en de TI84**
- **Trendlijn, correlatie en de TInspire**
- **Trendlijn, correlatie en de Casio**
- **Trendlijn, correlatie en de HPprime**
- **Trendlijn, correlatie en de HPprime**

Met het volgende practicum kun je zien hoe je **de trendlijn en de correlatiecoëfficiënt in Excel** berekent. Dat is handig als je een grote set gegevens hebt. Je treft er ook in aan hoe je de **regressielijn**, dat is de meest geschikte lijn door de puntenwolk, kunt tekenen en er door Excel de vergelijking van kunt laten opstellen. In Excel heet die lijn de ‘trendlijn’. In het volgende onderdeel hoor je daar meer over.

- **Correlatie en regressie**

OPMERKING:
Natuurlijk is het veel mooier om een eigen dataset met gegevens van leerlingen in jouw jaargroep te gebruiken. Die moet je dan wel eerst zelf maken: statistisch onderzoek!