REGRESSIE met de TI-83

Bieke Van Deyck – Studente K.U. Leuven

HOOFDSTUK 1:
INLEIDENDE BEGRIPPEN: CENTRUMMATEN EN SPREIDINGSMATEN.

A. Inleiding.

Statistiek is het verzamelen en bestuderen van numerieke gegevens om vervolgens conclusies te trekken uit deze data.
Zijn we bijvoorbeeld geïnteresseerd in de lengte van alle Belgische 21-jarigen, dan is de ***populatie*** de verzameling van al die lengtes. Dit is een zeer omvangrijke groep die een onderzoeker allicht niet ter zijner beschikking heeft. Daarom beperkt men zich vaak tot het onderzoeken van een ***steekproef***. Dit is een deelverzameling van de populatie, bijvoorbeeld de lengte van 10 21-jarigen uit elke Belgische gemeente. Een goede steekproef moet voldoende elementen bevatten en een goed beeld geven van de volledige populatie. Als we in ons voorbeeld de 10 21-jarigen uit elke gemeente willekeurig kiezen en niet bijvoorbeeld 10 jongeren uit dezelfde familie of 10 jongeren uit een zelfde basketbalclub, zal onze steekproef goed zijn. Men zegt dan dat de steekproef ***representatief*** is.

In de ***beschrijvende statistiek*** zal men de meetresultaten overzichtelijk weergeven in tabellen of grafische voorstellingen (zoals een histogram of sectordiagram) en samenvatten dmv enkele kengetallen.
Hierbij maken we onderscheid tussen twee soorten kengetallen:

de centrumgetallen:
Dit zijn getallen waarrond de waarnemingsgetallen (= de numerieke gegevens) zich concentreren.
We onderscheiden:
- het rekenkundig gemiddelde
- de mediaan

de spreidingsgetallen:
Dit zijn getallen die aangeven hoe ver de waarnemingsgetallen afwijken van de centrumgetallen.
We onderscheiden:
- de spreidingsbreedte
- de interkwartielafstand
- de variantie
- de standaardafwijking

Op basis van een steekproef zal men in de ***verklarende statistiek*** uitspraken doen over de hele populatie.

B. Centrummaten.

(1) Het rekenkundig gemiddelde.

Als iemand zegt: "Dokters verdienen meer dan taxichauffeurs", dan bedoelt hij niet dat ***iedere*** dokter meer verdient dan ***gelijk welke*** taxichauffeur. Wat wel bedoeld wordt, is dat het ***gemiddelde*** inkomen van een dokter hoger ligt dan het ***gemiddelde*** inkomen van een taxichauffeur. Dit gemiddelde noemen we het ***rekenkundig gemiddelde*** (of kortweg gemiddelde).
We bekomen het rekenkundig gemiddelde door de waarden van de waarnemingsgetallen op te tellen en dit te delen door het aantal waarnemingsgetallen.
Of in formule kunnen we dit als volgt noteren:

Het rekenkundig gemiddelde $\bar{x}$ van een reeks numerieke gegevens $x_1, x_2, ..., x_n$ is:

$$\bar{x} = \frac{\text{som van de waarnemingsgetallen}}{\text{aantal waarnemingsgetallen}}$$

$$= \frac{x_1 + x_2 + ... + x_n}{n}$$

$$= \frac{\sum_{i=1}^{n} x_i}{n}$$

$$= \frac{1}{n} \sum_{i=1}^{n} x_i$$

☒ Voorbeeld:

Zes juryleden gaven op de Miss-België verkiezing 1999 de volgende scores op 10 aan de eerste kandidate: 7, 4, 8, 9, 8, 6.
Het rekenkundig gemiddelde is $\bar{x} = \frac{7+4+8+9+8+6}{6} = 7$ en deze score verschijnt op het scorebord voor deze kandidate.

In de formule $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ wordt elk waarnemingsgetal meegeteld. Een getal dat 5 keer waargenomen wordt, of ***frequentie*** 5 heeft, moet 5 keer meegeteld worden. Wanneer er waarnemingsgetallen zijn met een hoge frequentie, bezorgt onze formule voor het rekenkundig gemiddelde ons veel nodeloos rekenwerk. Daarom zoeken we een vereenvoudiging:

Onderstel dat $x_1, x_2, ..., x_p$ alle ***verschillende*** waarnemingsgetallen zijn met respectievelijke frequenties $n_1, n_2, ..., n_p$.

Dan is:

$$\bar{x} = \frac{1}{n}\left(\underbrace{x_1 + x_1 + \ldots + x_1}_{n_1 keer} + \underbrace{x_2 + x_2 + \ldots + x_2}_{n_2 keer} + \ldots + \underbrace{x_p + x_p + \ldots + x_p}_{n_p keer} \right)$$

$$= \frac{1}{n}\left(n_1 x_1 + n_2 x_2 + \ldots + n_p x_p \right)$$

$$= \frac{1}{n} \sum_{i=1}^{p} n_i x_i$$

$$= \sum_{i=1}^{p} \frac{n_i}{n} x_i$$

Men noemt $\frac{n_i}{n}$ de ***relatieve frequentie*** van x_i.

☒ Voorbeeld:

Voor de verkiezing van Miss-België 2000 liet men wat meer juryleden hun zegje doen. Dit keer gaven 24 juryleden hun punten op 10. Dit leverde voor de eerste kandidate de volgende scores op:
2, 6, 3, 3, 7, 8, 10, 8, 6, 5, 4, 4, 2, 8, 7, 7, 6, 4, 5, 7, 7, 5, 7 en 6.
Welke score zal er voor haar op het scorebord komen?
Om dit vlug op te lossen maken we een tabel:

x_i	n_i	$x_i n_i$
2	2	4
3	2	6
4	3	12
5	3	15
6	4	24
7	6	42
8	3	24
10	1	10

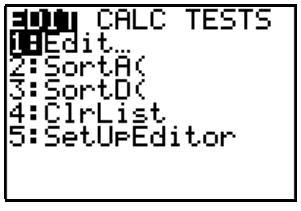
$$\bar{x} = \frac{4+6+12+15+24+42+24+10}{24} = \frac{137}{24} \approx 5.7$$

Dus voor deze kandidate komt de score 5.7 op het scorebord.

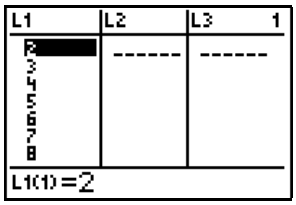
💻 Met de TI-83

We voeren de gegevens in met behulp van lijsten (rijen).

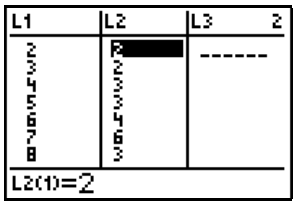
STAT EDIT 1:Edit



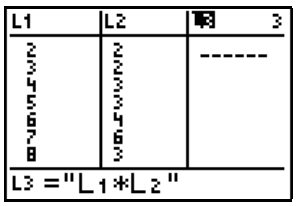
We vullen nu de 9 verschillende scores in bij de lijst L1. Dit doen we door op L1(1) te staan, de eerste score in te vullen (hier 2) en dan op ENTER te drukken. Enz.



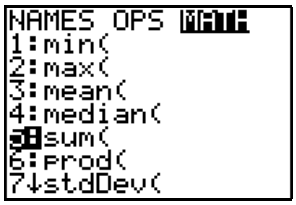
De frequenties vullen we in in de lijst L2.



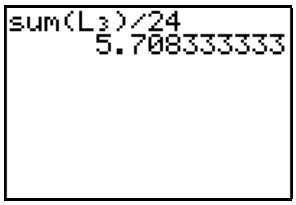
Ga nu op L3 staan, druk op ENTER en vul onderaan in "L1*L2". Hierdoor wordt de derde lijst gedefinieerd als het product van de eerste en de tweede.
Druk dan op ENTER.



Druk 2nd MODE (QUIT). Zo kom je in het basisscherm.
En dan op 2nd STAT (LIST) MATH 5:sum(



Typ dan L3) / 24 en duw op ENTER en je verkrijgt het rekenkundig gemiddelde.



Wanneer we het rekenkundig gemiddelde moeten berekenen van een aantal zeer grote getallen, is het nuttig als we enkele vereenvoudigingen kunnen doorvoeren waardoor we de berekening uit het hoofd kunnen doen. Deze vereenvoudigingen worden gegeven door de volgende twee eigenschappen:

Eigenschap 1:
Als men elk waarnemingsgetal vermindert met een zelfde getal, dan vermindert het rekenkundig gemiddelde ook met dat getal.

$$\frac{\sum_{i=1}^{n}(x_i - a)}{n} = \bar{x} - a \qquad \text{met } a \in \Re$$

Uitgaande van deze eigenschap kunnen we een nieuwe formule opstellen voor $\bar{x}$ en hieruit een praktische werkwijze voor de berekening van $\bar{x}$ afleiden:

$$\bar{x} = \frac{\sum_{i=1}^{n}(x_i - a)}{n} + a$$

Praktische werkwijze:

Bij grote waarnemingsgetallen kan je het rekenkundig gemiddelde als volgt bepalen:
- Verminder elk waarnemingsgetal met een zelfde getal a.
- Bereken het rekenkundig gemiddelde van de nieuw bekomen getallen.
- Vermeerder dit rekenkundig gemiddelde met a.

☒ Voorbeeld:

$x_1 = 401, \quad x_2 = 408, \quad x_3 = 408, \quad x_4 = 416 \quad \text{en} \quad x_5 = 417$

- We definiëren nieuwe getallen: $x_i^* = x_i - 400$
  Hierbij is a = 400.
  $x_1^* = 1, \quad x_2^* = 8, \quad x_3^* = 8, \quad x_4^* = 16 \quad \text{en} \quad x_5^* = 17$
- Hiervan berekenen we het rekenkundig gemiddelde: $\frac{1+8+8+16+17}{5} = \frac{50}{5} = 10$
- Dit vermeerderen we met a = 400.
  $\bar{x} = 10 + 400 = 410$

Eigenschap 2:
Als men elk waarnemingsgetal deelt door een getal, niet gelijk aan 0, dan wordt het rekenkundig gemiddelde ook gedeeld door dat getal.

$$\frac{\sum_{i=1}^{n} \frac{x_i}{b}}{n} = \frac{\bar{x}}{b} \qquad \text{met } b \in \Re_0$$

Uitgaande van ook deze eigenschap kunnen we een nieuwe formule opstellen voor $\bar{x}$ en hieruit een praktische werkwijze voor de berekening van $\bar{x}$ afleiden:

$$\bar{x} = \frac{\sum_{i=1}^{n} \frac{x_i}{b}}{n} . b$$

Praktische werkwijze:

Bij grote waarnemingsgetallen kan je het rekenkundig gemiddelde ook als volgt bepalen:
- Deel elk waarnemingsgetal door een zelfde getal b met $b \neq 0$.
- Bereken het rekenkundig gemiddelde van deze nieuwe reeks getallen.
- Vermenigvuldig dit rekenkundig gemiddelde met b.

☒ Voorbeeld:

$x_1 = 500, \quad x_2 = 800, \quad x_3 = 1100 \quad \text{en} \quad x_4 = 1200$

- We definiëren nieuwe getallen: $x_i^* = \frac{x_i}{100}$

  Hierbij is b = 100.

  $x_1^* = 5, \quad x_2^* = 8, \quad x_3^* = 11 \quad$ en $\quad x_4^* = 12$
- Hiervan berekenen we het rekenkundig gemiddelde: $\frac{5+8+11+12}{4} = \frac{36}{4} = 9$
- Dit vermenigvuldigen we met b = 100.

  $\bar{x} = 9 \,.\, 100 = 900$

Tenslotte bewijzen we nog een eigenschap die later nuttig zal zijn:

Eigenschap 3:
De som van de verschillen tussen de waarnemingsgetallen en het rekenkundig gemiddelde is 0.

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

Gegeven: de waarnemingsgetallen $x_1, x_2, ..., x_n$

Te bewijzen: $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$

Bewijs: Volgens eigenschap 1 geldt: $\frac{\sum_{i=1}^{n}(x_i - a)}{n} = \bar{x} - a$

Kies nu $a = \bar{x}$: $\frac{\sum_{i=1}^{n}(x_i - \bar{x})}{n} = \bar{x} - \bar{x} = 0$

Het verschil tussen een waarnemingsgetal en het rekenkundig gemiddelde, noemt men de ***afwijking*** van het waarnemingsgetal tov het rekenkundig gemiddelde.

(2) De mediaan.

De gemiddelde tijd van de 1447 deelnemers aan de marathonloop in 1976 te Honolulu, was 4 uur en 15 minuten (over de 42.195 km). Misschien verwacht je dat de helft van de deelnemers sneller dan 4 uur en 15 minuten heeft gelopen, maar dat is niet zo. 822 Deelnemers liepen sneller dan 4 uur en 15 minuten, dat is dus bijna 57 %. En 43 % liep dus langzamer dan het gemiddelde.
In dit voorbeeld zien we dat we niet altijd voldoende informatie hebben met het rekenkundig gemiddelde. We kunnen dan nog een andere centrummaat gebruiken: de ***mediaan***. Deze ligt echt in het centrum of het midden van de naar grootte gerangschikte waarnemingsgetallen.

De mediaan is het middelste getal voor een oneven aantal, naar grootte gerangschikte waarnemingsgetallen en het gemiddelde van de twee middelste getallen voor een even aantal naar grootte gerangschikte waarnemingsgetallen.

Voor n oneven: $Med = x_k$ met $k = \frac{n+1}{2}$

Voor n even: $Med = \frac{x_k + x_{k+1}}{2}$ met $k = \frac{n}{2}$

☒ Voorbeeld:

Er zijn op de Miss-België verkiezing 7 juryleden en zij geven de volgende scores aan de laatste kandidate: 3, 6, 1, 8, 5, 7 en 3. Wat is de mediaan?

Hiervoor rangschikken we de scores van klein naar groot: 1 3 3 5 6 7 8
n = 7, dit is oneven

$k = \frac{n+1}{2} = 4$

Het vierde getal in de gesorteerde rij is 5, dus Med = 5

☒ Voorbeeld:

Bij de Mister-Belgium verkiezing zijn er maar 6 juryleden en hun scores voor de laatste kandidaat zijn de volgende: 3, 1, 5, 6, 3 en 7. Wat is hier de mediaan?

We rangschikken opnieuw de scores van klein naar groot: 1 3 3 5 6 7
n = 6, dit is even

$k = \frac{n}{2} = 3$

Med = $\frac{3+5}{2} = 4$

Nochtans is 4 zelf geen waarnemingsgetal.

Ongeveer 50% van de gegevens zijn kleiner dan of gelijk aan de mediaan.

C. Spreidingsmaten.

(1) De spreidingsbreedte.

Een leraar wiskunde doet een zelfde toets in twee verschillende klassen.
De behaalde scores in klas A zijn x_i: 5, 6, 6, 6, 7, 7, 7 en 8.
Terwijl de behaalde scores in klas B zijn y_i: 2, 3, 3, 5, 9, 10, 10 en 10.
We berekenen het rekenkundig gemiddelde in beide klassen:

Voor klas A geldt: $\bar{x} = \frac{5+6+6+6+7+7+7+8}{8} = 6.5$

Voor klas B geldt: $\bar{y} = \frac{2+3+3+5+9+10+10+10}{8} = 6.5$

Het gemiddelde is in beide klassen hetzelfde, maar kan de leraar in beide klassen even tevreden zijn?
Alleen met gemiddeldes weten we niet of er bijvoorbeeld veel onvoldoendes zijn of of er leerlingen het maximum behaalden.
We zien dat de resultaten in klas B veel meer ***verspreid*** liggen dan die in klas A.

De spreiding van de waarnemingsgetallen kan je zeer eenvoudig meten door het verschil te bepalen tussen het grootste en het kleinste waarnemingsgetal. We noemen dit verschil de ***spreidingsbreedte***: de waarnemingsgetallen "variëren" tussen het kleinste en het grootste waarnemingsgetal.

Het verschil tussen het grootste en het kleinste waarnemingsgetal, noemt men de spreidingsbreedte. $R = x_{\max} - x_{\min}$

☒ Voorbeeld:

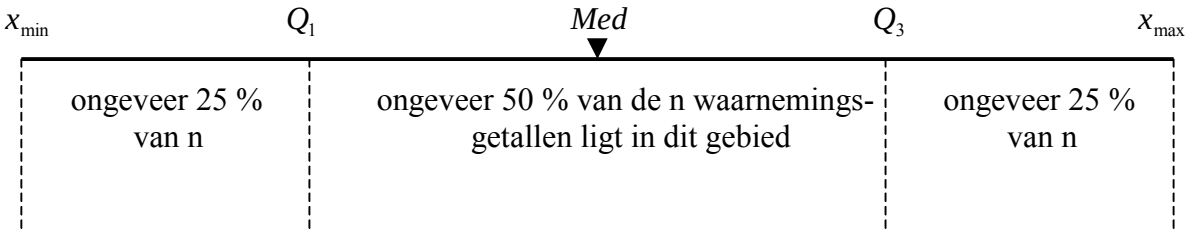
Voor de klassen A en B (hogerop beschreven) geldt:

De spreidingsbreedte van klas A is R = 8-5 = 3.
De spreidingsbreedte van klas B is R = 10-2 = 8.

De grote spreidingsbreedte van klas B wijst erop dat de scores veel meer verspreid zijn dan in klas A.

(2) De interkwartielafstand.

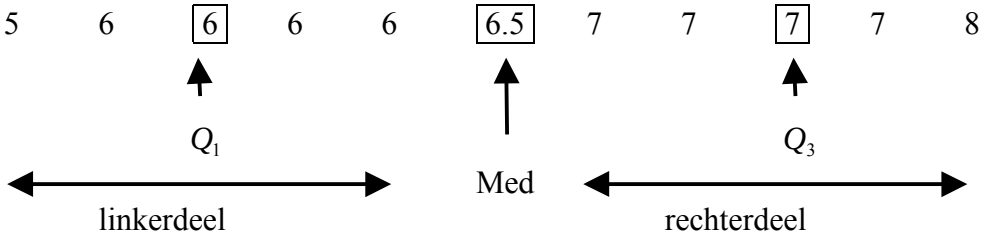
De mediaan verdeelt de geordende gegevens in een linker- en een rechterdeel met in elk deel hetzelfde aantal getallen. De mediaan bepaalt de grens tussen deze twee delen en behoort tot geen van deze delen.
Het eerste ***kwartiel***, Q_1, is de mediaan van het linkerdeel en het derde kwartiel, Q_3, is de mediaan van het rechterdeel.



Ongeveer 25 % van de gegevens zijn kleiner dan of gelijk aan het eerste kwartiel Q_1. Voor het tweede kwartiel $Q_2 = Med$ en het derde kwartiel Q_3 wordt dit respectievelijk 50 % en 75%. Tussen Q_1 en Q_3 liggen dan ongeveer 50 % van de gegevens.

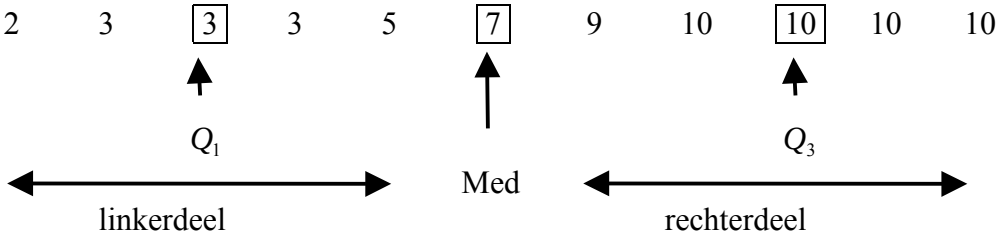
☒ Voorbeeld:

Voor de punten van klas A: 5, 6, 6, 6, 7, 7, 7 en 8 geldt:



$Q_1 = 6$ en $Q_3 = 7$

Voor de punten van klas B: 2, 3, 3, 5, 9, 10, 10 en 10 geldt:



$Q_1 = 3$ en $Q_3 = 10$

De ***interkwartielafstand***, IQR, meet de spreiding van de middelste helft van de geordende gegevens.

Het verschil tussen derde en het eerste kwartiel noemt men de interkwartielafstand. IQR = $Q_3 - Q_1$

☒ Voorbeeld:

Voor klas A is IQR = 7-6 = 1.
Voor klas B is IQR = 10-3 = 7.
Ook hier weer wijst de grotere interkwartielafstand in klas B, op een grotere spreiding van de gegevens.

Eigenschap:
Als de interkwartielafstand klein is, dan betekent dit dat de waarnemingsgetallen goed bij de mediaan aansluiten.

(3) De variantie.

Je kan overwegen om de afwijking van elk waarnemingsgetal ten opzichte van het rekenkundig gemiddelde te gebruiken om een norm van spreiding te bepalen.
Je weet echter dat de som van de afwijkingen ten opzichte van het rekenkundig gemiddelde altijd gelijk is aan 0. (zie eigenschap 3 bij het gemiddelde)
We keren terug naar het voorbeeld van de twee klassen A en B met elk 8 leerlingen. In beide klassen is het rekenkundig gemiddelde 6.5.
Bekijk nu de afwijkingen ten opzichte van het gemiddelde:

Klas A

x_i	$x_i - \bar{x}$
5	-1.5
6	-0.5
6	-0.5
6	-0.5
7	0.5
7	0.5
7	0.5
8	1.5

$$\sum_{i=1}^{8}(x_i - \bar{x}) = 0$$

Klas B

y_i	$y_i - \bar{y}$
2	-4.5
3	-3.5
3	-3.5
5	-1.5
9	2.5
10	3.5
10	3.5
10	3.5

$$\sum_{i=1}^{8}(y_i - \bar{y}) = 0$$

Om de afwijkingen van alle waarnemingsgetallen ten opzichte van het rekenkundig gemiddelde toch te kunnen gebruiken als norm van spreiding, kan men ofwel de absolute waarde ervan nemen, ofwel de afwijkingen kwadrateren. In de statistiek is het echter gebruikelijk de afwijkingen te kwadrateren omdat zo de verder afgelegen waarnemingen een grotere bijdrage leveren tot de spreiding.

Het rekenkundig gemiddelde van de kwadraten van de afwijkingen van de waarnemingsgetallen ten opzichte van hun rekenkundig gemiddelde, noemt men de variantie.

$$Var = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$

We noteren de variantie als Var, s_X^2 of s^2

☒ Voorbeeld:

We berekenen de variantie van de scores van de 8 leerlingen van klas A en van klas B.

Klas A

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
5	-1.5	2.25
6	-0.5	0.25
6	-0.5	0.25
6	-0.5	0.25
7	0.5	0.25
7	0.5	0.25
7	0.5	0.25
8	1.5	2.25

$$\sum_{i=1}^{8}(x_i - \bar{x})^2 = 6$$

Klas B

y_i	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
2	-4.5	20.25
3	-3.5	12.25
3	-3.5	12.25
5	-1.5	2.25
9	2.5	6.25
10	3.5	12.25
10	3.5	12.25
10	3.5	12.25

$$\sum_{i=1}^{8}(y_i - \bar{y})^2 = 90$$

De variantie in klas A is:

$$s_X^2 = Var_A = \frac{6}{8} = 0.75$$

De variantie in klas B is:

$$s_Y^2 = Var_B = \frac{90}{8} = 11.25$$

We merken op dat $Var_A < Var_B$. Zoals verwacht, duidt dit er op dat de spreiding van de scores in klas A kleiner is dan die in klas B.

Let op:

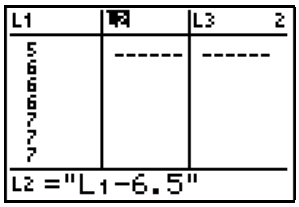
Door het kwadrateren van de afwijkingen, zal de variantie uitgedrukt zijn in een andere eenheid dan de eenheid van de waarnemingsgetallen.

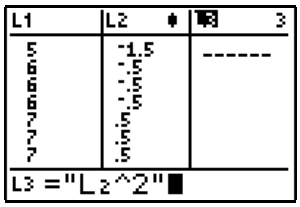
Bijvoorbeeld: Is de eenheid van de waarnemingsgetallen cm, dan zal de variantie cm^2 als eenheid hebben.

Met de TI-83

We voeren, zoals eerder gezien, de scores van de leerlingen van klas A in in L1. Stonden er al gegevens in L1 dan kunnen we die wissen door op L1 te gaan staan en op CLEAR ENTER te duwen.

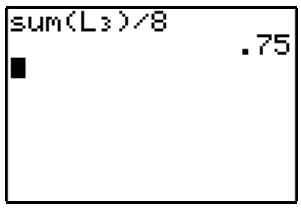
We definiëren L2 als “L1-6.5” (het gemiddelde wordt er afgetrokken) en L3 als “L2^2”.





Druk 2nd MODE (QUIT). Zo kom je in het basisscherm.
En dan op 2nd STAT (LIST) MATH 5:sum(

Typ dan L3) / 8 en duw op ENTER en je verkrijgt de variantie.



(4) De standaardafwijking.

We “verbeteren” het kwadrateren bij het berekenen van de variantie door de positieve vierkantswortel ervan te bepalen. Hierdoor is het bekomen resultaat opnieuw uitgedrukt in dezelfde eenheid als de eenheid van de waarnemingsgetallen. Deze vierkantswortel is een nieuw spreidingsgetal dat men standaardafwijking noemt.

De positieve vierkantswortel van de variantie noemt men de standaardafwijking.
Genoteerd: s_X

$$s_X = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}$$

Voorbeeld:

We berekenen de standaardafwijking van de scores van de 8 leerlingen van klas A en klas B.

Klas A: $s_X = \sqrt{0.75} \approx 0.86$

Klas B: $s_Y = \sqrt{11.25} \approx 3.35$

Met behulp van frequenties kunnen we de formules voor variantie en standaardafwijking als volgt schrijven:

Als $x_1, x_2, ..., x_p$ de verschillende waarnemingsgetallen zijn met respectievelijke frequenties $n_1, n_2, ..., n_p$ dan is:

$$Var = \frac{\sum_{i=1}^{p} n_i (x_i - \bar{x})^2}{n} = \sum_{i=1}^{p} \frac{n_i}{n} (x_i - \bar{x})^2$$

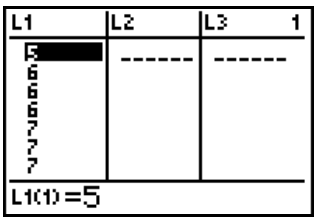
en $$s_X = \sqrt{\frac{\sum_{i=1}^{p} n_i (x_i - \bar{x})^2}{n}} = \sqrt{\sum_{i=1}^{p} \frac{n_i}{n} (x_i - \bar{x})^2}$$

☒ Voorbeeld:

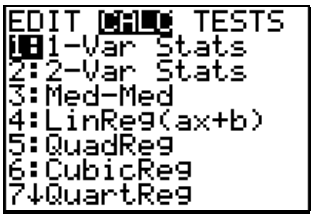
Zo geldt voor klas A met scores 5, 6, 6, 6, 7, 7, 7 en 8 dat

$$\begin{aligned} s_X &= \sqrt{\frac{1.(5-6.5)^2 + 3.(9-6.5)^2 + 3.(7-6.5)^2 + 1.(8-6.5)^2}{8}} \\ &= \sqrt{\frac{(-1.5)^2 + 3.(-0.5)^2 + 3.(0.5)^2 + (1.5)^2}{8}} \\ &= \sqrt{\frac{6}{8}} \\ &= \sqrt{0.75} \approx 0.866 \end{aligned}$$

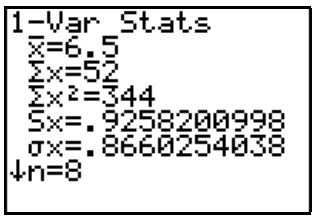
D. Centrum- en spreidingsmaten met de TI-83.

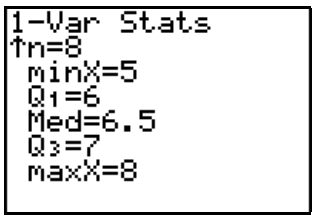


Om het berekenen van de centrum- en spreidingsgetallen te illustreren, definiëren we eerst de lijst L1 als de punten van de leerlingen uit klas A:



Na het indrukken van STAT CALC 1:1-Var Stats en dan de lijst L1, verschijnen er veel kentallen van de lijst L1 op het basisscherm:





Hierbij zijn vooral nuttig:

n = aantal waarnemingsgetallen
σx = standaarafwijking (Let op Sx is iets anders !)
$\bar{x}$ = gemiddelde
minX = x_{min}
Q1 = eerste kwartiel
Med = mediaan
Q3 = derde kwartiel
MaxX = x_{max}

HOOFDSTUK 2:
DE BIVARIATE VERDELING.

A. Probleembeschrijving.

In de beschrijvende statistiek leerden we hoe we een kwantitatieve eigenschap konden onderzoeken binnen een populatie. We nemen een representatieve ***steekproef***, onderzoeken binnen deze deelverzameling die bepaalde eigenschap en bekomen zo de waarnemingsgetallen.
Deze ordenen we in een frequentietabel, stellen ze grafisch voor en trachten ze samen te vatten door centrum- en spreidingsmaten te berekenen.

We hebben ons steeds geconcentreerd op de populatie van één enkele variabele, zoals bijvoorbeeld gewicht, lengte, punten op een examen, punten bij een Miss-België verkiezing,...
Soms is het interessant om de relatie te bestuderen tussen een koppel variabelen, wat de ***bivariate verdeling*** met zich meebrengt.
Bijvoorbeeld, van elke volwassen mens wordt zijn lengte X in cm en gewicht Y in kg gemeten. Het geordende paar (X,Y) heeft een bivariate verdeling.

We stellen ons nu de volgende vraag:
Bestaat er een ***verband*** of ***correlatie*** tussen twee eigenschappen die betrekking hebben op dezelfde populatie?
M.a.w. bestaat er een bepaalde relatie tussen de stochastische veranderlijken X en Y van het koppel (X,Y) met hun bivariate verdeling?

☒ Voorbeelden:

- Is er een verband tussen de leeftijden van huwelijkspartners?
- Is er een verband tussen de lengte van een moeder en haar kind?
- Is er een verband tussen het gewicht en het voetoppervlak (het contactoppervlak met de grond) van Zuid-Amerikaanse slakken?
- Is er een verband tussen het aantal tewerkgestelde mannen en het aantal tewerkgestelde vrouwen van de beroepsbevolking?

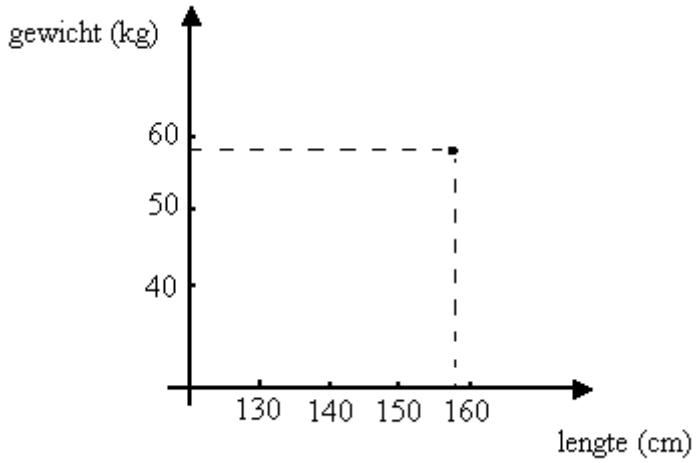
We trachten een eventueel verband te beschrijven door middel van een grafische voorstelling of door middel van een getal. Dit zal in de volgende hoofdstukken uitgewerkt worden op basis van de correlatie- en regressierekening.

B. Het spreidingsdiagram.

De meest eenvoudige methode om twee gemeten variabelen simultaan weer te geven, is een ***spreidingsdiagram***.
Dit gebruikt een horizontale as voor één van de variabelen en een verticale as voor de andere. Er wordt een punt geplaatst voor elk observatie-paar (X_i, Y_i) op de kruising van zijn twee waarden.

☒ Voorbeeld:

Een kind is 1m58 lang en weegt 58 kg.
We kiezen X = lengte (in cm)
Y = gewicht (in kg)
En we plaatsen deze observatie in het spreidingsdiagram als volgt:



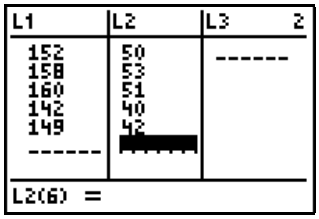
Als je de waarde van één variabele wilt gebruiken om de waarde van een andere variabele te voorspellen, is de conventie om de variabele waarmee je de voorspelling doet (=de ***onafhankelijke variabele***) op de horizontale as te plaatsen en de te voorspellen variabele (=de ***afhankelijke variabele***) op de verticale as.

💻 Hoe kunnen we zelf een spreidingsdiagram tekenen met de TI-83?

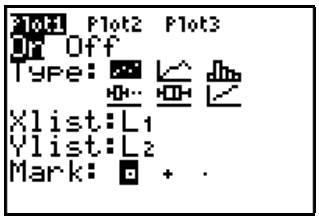
Gegeven 5 kinderen met hun lengte en gewicht:

lengte X	152	158	160	142	149
gewicht Y	50	53	51	40	42

We voeren de gegevens in in de lijsten L1 en L2: L1 zijn de lengtes en L2 de gewichten.

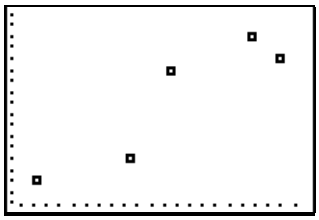


Om een spreidingsdiagram te maken, doen we het volgende:
2nd Y= (STAT PLOT) 1:
en definieer plot 1 zoals hiernaast gegeven:



Hierbij zijn:
- Xlist: de gegevens die je op de horizontale as wilt
- Ylist: de gegevens die je op de verticale as wilt
- Mark: hoe wil je dat je punten er uit zien?

Druk dan op ZOOM 9:Zoomstat en dan verschijnt je spreidingsdiagram.
Deze instructie zorgt er steeds voor dat alle punten op je scherm kunnen. Het bereik wordt dus automatisch aangepast.



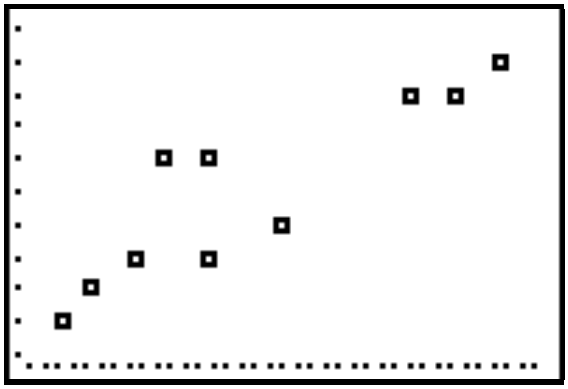
Via TRACE en de pijltjestoetsen kun je nu de coördinaten van elk punt nagaan.

☒ Voorbeeld:

Bij een keuring voor militaire dienst zijn 10 jongens aan een medisch onderzoek onderworpen. Daarbij is van elke jongen de lengte en de schoenmaat gemeten.

Lengte X	165	167	170	172	175	175	180	189	192	195
Schoenmaat Y	37	38	39	42	39	42	40	44	44	45

Dit levert het volgende spreidingsdiagram:



(a) Wat is de richting van de puntenwolk?

(b) Hoe zou je de richting van de puntenwolk in woorden kunnen beschrijven? Met andere woorden hoe kan je de relatie tussen de lengte en de schoenmaat van de kandidaat-soldaten weergeven?

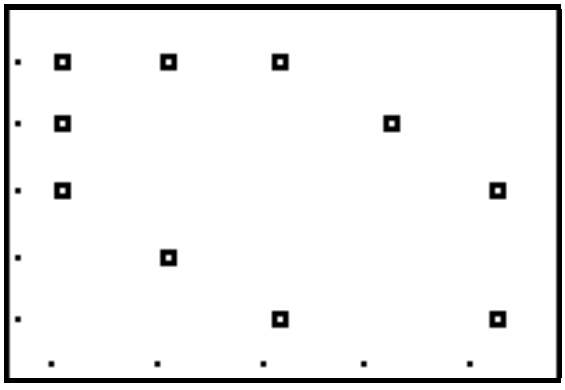
Twee variabelen worden ***positief gecorreleerd*** genoemd als grote waarden van de ene variabele overeenkomen met grote waarden voor de andere variabele.

We zien duidelijk dat de lengte en de schoenmaat positief gecorreleerd zijn. Dit zien we aan de richting van de puntenwolk: van links onder naar rechts boven.

☒ Voorbeeld:

Hieronder staat een tabel van het koffie- en theegebruik van 10 huishoudens, in koppen per persoon per dag.

koffiegebruik X	3	6	2	4	5	3	4	2	6	2
theegebruik Y	6	2	4	6	5	3	2	6	4	5



(a) Wat is hier de richting van de puntenwolk?

(b) En hoe zou je in dit geval de richting van de puntenwolk in woorden kunnen beschrijven? Met andere woorden hoe kan je de relatie tussen het koffie- en theegebruik weergeven?

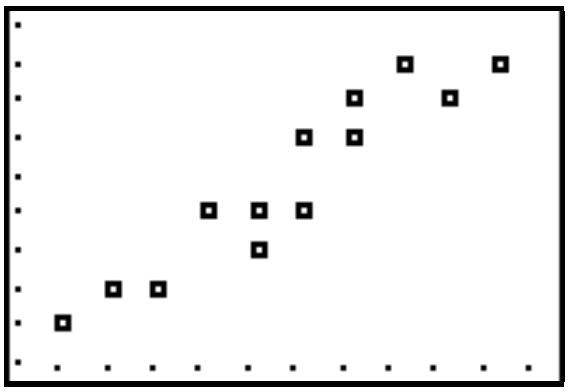
Omgekeerd, worden twee variabelen ***negatief gecorreleerd*** genoemd als grotere waarden van de ene variabele overeenkomen met kleinere waarden van de andere variabele.

Ook hier zien we vrij duidelijk dat het koffiegebruik en het theegebruik in huishoudens negatief gecorreleerd zijn: de puntenwolk gaat van links boven naar rechts onder.

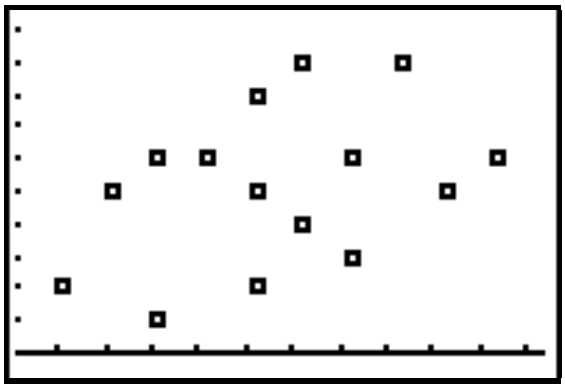
☒ Voorbeeld:

Aan een hondenshow is een wedstrijd verbonden voor de mooiste en meest verzorgde hond. Hiervoor worden drie verschillende juryleden geraadpleegd: een eigenaar van een

hondenschool (jury A), een opleider van blindengeleide honden (jury B) en een hondenliefhebber en –kenner (jury C).
Er zijn 15 honden die aan de wedstrijd deelnemen en zij worden door de drie juryleden beoordeeld.
Wanneer we nu de punten van jury A (op de X-as) en jury B (op de Y-as) vergelijken bekomen we het volgende spreidingsdiagram:



Terwijl wanneer we de punten van jury A (op de X-as) en C (op de Y-as) vergelijken bekomen we dit spreidingsdiagram:



(a) Welk verschil zie je tussen de twee spreidingsdiagrammen?

(b) Hoe kan je dit verschil hier beschrijven in functie van de punten van de drie juryleden?

De ***sterkte van de correlatie*** is afhankelijk van de hoeveelheid punten die de correlatie volgen. Bijvoorbeeld hoe meer punten de positieve correlatie volgen, hoe sterker de positieve correlatie is tussen de twee desbetreffende variabelen en hoe preciezer er een voorspelling kan gedaan worden voor de ene variabele op grond van de andere.

In het voorbeeld was er in het eerste spreidingsdiagram dus een ***zeer sterke positieve correlatie*** terwijl in het tweede spreidingsdiagram is er een ***zwakke positieve correlatie***.

☒ Voorbeeld:

Bekijken we opnieuw het voorbeeld van de kandidaat-soldaten:

Lengte X	165	167	170	172	175	175	180	189	192	195
Schoenmaat Y	37	38	39	42	39	42	40	44	44	45

(a) We hadden gezien dat de lengte en de schoenmaat positief gecorreleerd waren.
Wat wilde dit juist zeggen?

(b) Bekijk de koppels (175,42) en (180,40).
Klopt dit voor deze twee koppels?

Let dus op: het concept van correlatie is slechts een ***statistische tendens***. Er kunnen dus punten zijn die niet aan die correlatie voldoen.

☒ Voorbeeld:

We kunnen in een spreidingsdiagram ook punten met een ***label*** gebruiken om onderscheid te maken tussen verschillende categorieën onder de gegevens.
Bijvoorbeeld zo kunnen we een "+" gebruiken om aan te geven dat de examenscore van een jongen is en een "o" dat het van een meisje is.

💻 **Met de TI-83:**

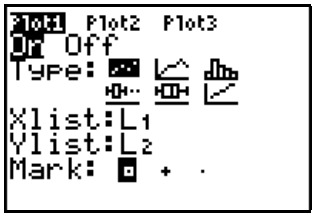
Plaats de volgende examenscores van 5 meisjes-leerlingen in de lijsten L1 en L2 in de TI-83:

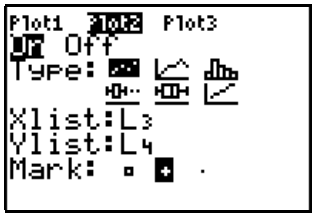
examen 1	9	8	5	9	6
examen 2	10	7	4	8	7

Plaats nu ook de volgende examenscores van 5 jongens-leerlingen in de lijsten L3 en L4 in de TI-83:

examen 1	7	9	2	6	6
examen 2	7	7	8	4	8

Druk op 2nd Y= (STAT PLOT) en definieer plot 1 en plot 2 zoals aangegeven hiernaast:





Wanneer we dan op ZOOM 9:Zoomstat drukken, geeft hij de twee spreidingsdiagrammen over elkaar, met elk hun label.

Maak nu een gelabeld spreidingsdiagram van de examenscores van de 5 meisjes en de 5 jongens.

☒ Voorbeeld:

Een leerkracht geeft een toets maar de punten van de leerlingen zijn zo erbarmelijk slecht, dat de leerkracht de leerlingen een tweede kans wilt geven. Hij geeft een tweede toets over hetzelfde stukje leerstof.
Dit zijn de resultaten op de twee toetsen:

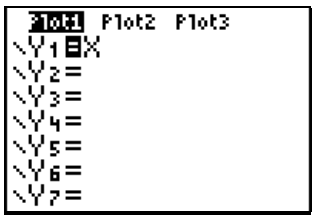
toets 1 X	2	2	3	5	4	6	2	5	8	5
toets 2 Y	9	7	5	3	9	8	2	6	8	6

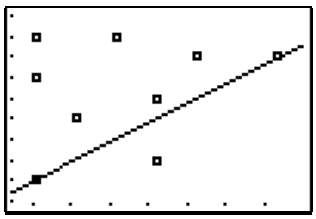
(a) Teken met de TI-83 het spreidingsdiagram.

(b) Teken ook de rechte y = x op het spreidingsdiagram. Deze rechte noemen we de **45°-lijn** (omdat ze een hoek van 45° maakt met de x-as).

Tips voor de TI-83:
Druk op Y= en vul in, bij de juiste plot waarin ook je spreidingsdiagram staat, Y1=X.

Druk dan op ZOOM 9:Zoomstat en je spreidingsdiagram wordt afgedrukt samen met de 45°-lijn.





(c) Wat kun je zeggen over de punten die op het spreidingsdiagram gelegen zijn onder de 45°-lijn? En wat betekent dit hier concreet in het voorbeeld?

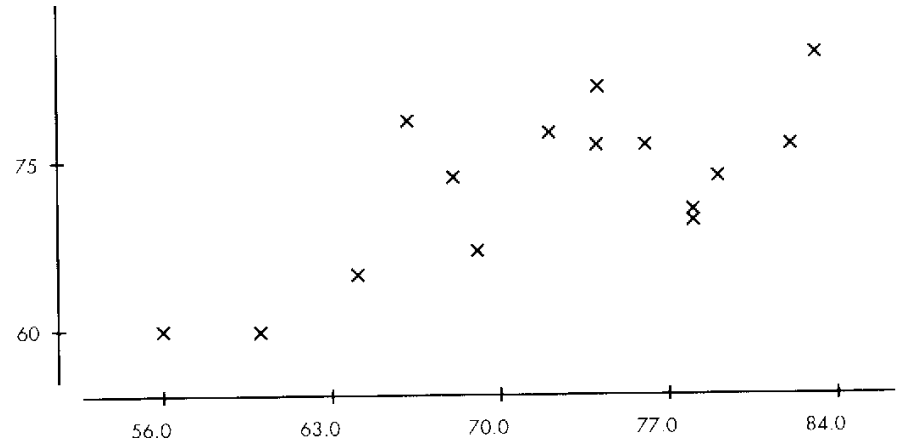
(d) Als de leerkracht wilt weten wie op de tweede toets meer behaalde dan op de eerste, naar wat moet hij dan juist kijken op het spreidingsdiagram?

Dit voorbeeld illustreert dat, wanneer we werken met gepaarde data, de 45°-lijn ons in een spreidingsdiagram nuttige informatie kan geven, over hoe de twee elementen van het paar zich in grootte gedragen tov elkaar.

Om je wat te oefenen in het begrip “correlatie” de volgende oefeningen:

Onderzoeksopdracht 1.
Beschouw het volgende spreidingsdiagram van hypothetische scores op een eerste en een tweede examen voor rijkswacht-commandanten om topfuncties bij de politie te kunnen bekleden:

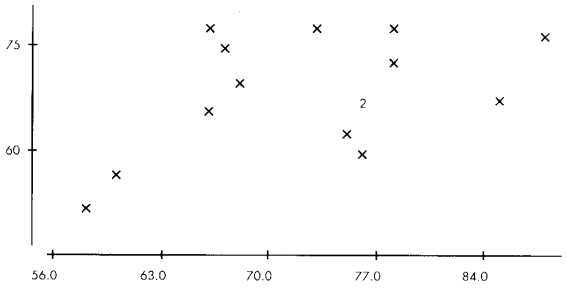
A



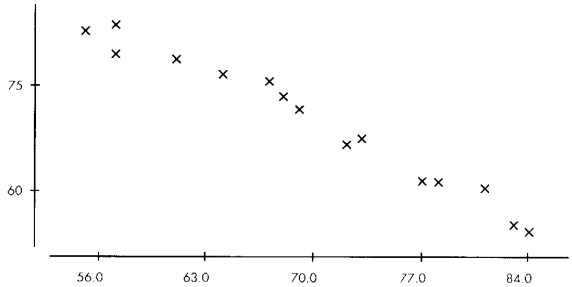
Beschrijf kort wat het spreidingsdiagram toont over de relatie tussen de scores op het eerste examen en die op het tweede examen. Met andere woorden als je de scores weet van het eerste examen, kun je dan vrij goede voorspellingen doen voor het tweede examen? Leg uit.

Onderzoeksopdracht 2.
Hieronder staan 5 andere spreidingsdiagrammen over de hypothetische scores op de twee examens. Jouw taak is om de richting (positief of negatief) en de sterkte (sterk, gematigd of zwak) van de correlatie tussen de scores op het eerste examen en die op het tweede examen te onderzoeken voor elk voorbeeld.

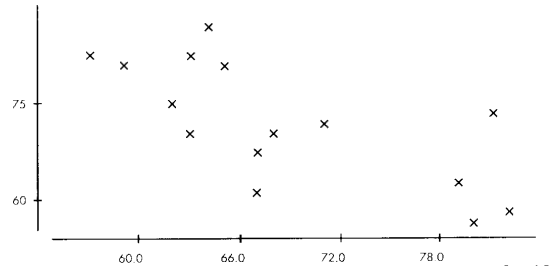
B



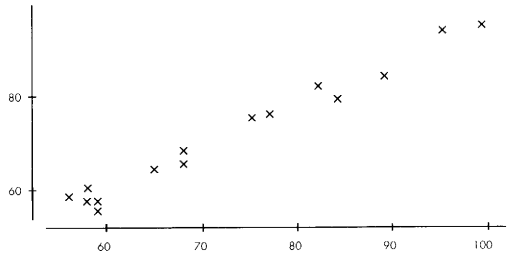
C



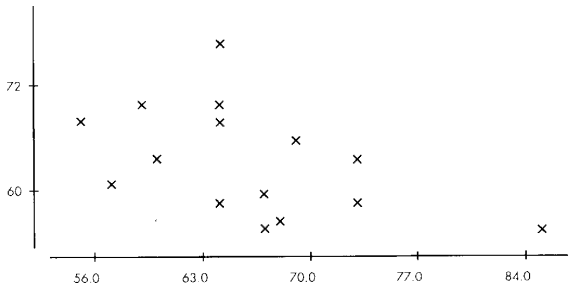
D



E



F



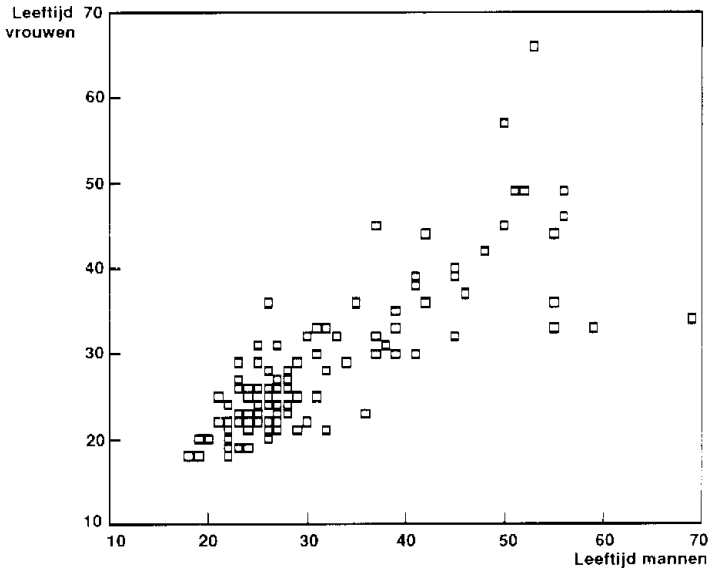
Doe dit door de bijhorende letter (A, ..., F) in te vullen in de tabel hieronder. Elke letter mag slechts één maal gebruikt worden.

	sterk	gematigd	zwak
negatief			
positief			

C. Voorbeelden.

Voorbeeld 1.

We vragen ons af of er een verband bestaat tussen de leeftijd van huwelijkspartners. We besluiten dit te onderzoeken dmv een representatieve steekproef. Daartoe kloppen we aan bij de dienst Burgerlijke Stand van ons gemeentehuis. Hier verschaft men ons volgende gegevens: van 120 willekeurig gekozen huwelijken in het jaar 1992 noteerden we de huwelijksdatum en de geboortedata van de partners. Hieruit hebben we dan de leeftijden van de partners op hun huwelijksdag afgeleid.
Zo bekomen we 120 koppels (X_i, Y_i) met X_i de huwelijksleeftijd van de man en Y_i de leeftijd van zijn vrouw.
We zetten deze punten nu uit en bekomen zo het volgende spreidingsdiagram:



(a) Lijkt er een correlatie te zijn tussen de leeftijd van de twee partners? Zo ja, is deze positief of negatief? En zou je deze als sterk of als zwak karakteriseren? Leg uitgebreid uit.

(b) Zijn er veel koppels die even oud zijn? Hoe kan je dit zien op het spreidingsdiagram?

(c) Zijn er meer mannen die ouder zijn dan hun vrouw of komt het omgekeerde vaker voor? Hoe zie je dit op het spreidingsdiagram?

(d) Vat in eigen woorden samen wat je kan leren over de huwelijksleeftijd van koppels door rekening te houden met de 45° lijn.

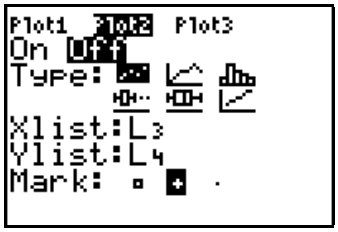
Voorbeeld 2.

In dit voorbeeld willen we nagaan of er een sterke erfelijke afhankelijkheid is tussen de lengte van een moeder en haar kind.
Verzamel daartoe met je klas zo'n 40 paren van moeders met hun kind. Hierbij moeten de kinderen allemaal dezelfde leeftijd hebben om een representatieve steekproef te krijgen. Zorg er ook voor dat er ongeveer evenveel zonen als dochters zijn.
Noteer van elk paar in een tabel, de lengte van de moeder, de lengte van het kind en het geslacht van het kind.

(a) Maak een spreidingsdiagram van je gegevens en gebruik hierbij labels.
Duid goed je assen aan en welk geslacht je voor welk label gebruikt.
Zie je een bepaalde correlatie als je geen rekening houdt met de labels? Kan je besluiten trekken ivm de lengte van moeder en kind?

(b) Maak nu wel onderscheid tussen de labels. Bekom je nu andere correlaties als je telkens maar naar één geslacht kijkt?
Kunnen we van erfelijkheid spreken, denk je?
💻 **Tip voor de TI-83:**

We zagen bij een gelabeld spreidingsdiagram dat dit eigenlijk een over elkaar plaatsen is van twee spreidingsdiagrammen. We kunnen de twee spreidingsdiagrammen dan ook elk apart bekijken door één van de twee op off te zetten.



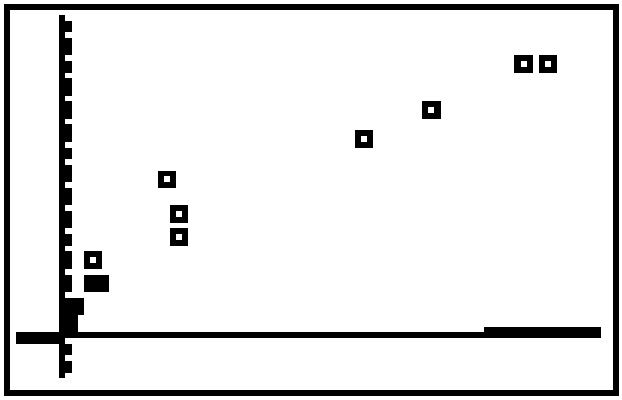
Voorbeeld 3.

We bekijken de volgende data van het gewicht en het voetoppervlak van 20 Zuid-Amerikaanse slakken van de soort Biomphalaria Glabrata.

Gewicht (g)	0.64	0.21	0.85	0.53	0.02	0.01	0.21	0.18	0.06	0.20	0.07	0.01	0.05
Voetopp (mm^2)	29	16	35	25	4	1	16	20	7	13	7	3	10

Gewicht (g)	0.81	0.53	0.18	0.06	0.20	0.07	0.01
Voetopp (mm^2)	35	25	20	7	13	7	1

Wanneer we van deze gegevens het spreidingsdiagram tekenen, bekomen we de volgende grafiek:

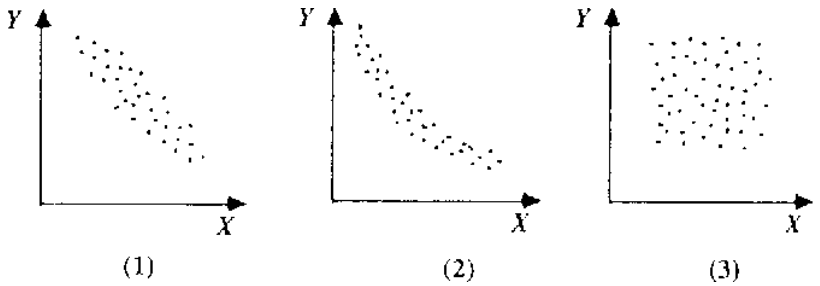


(a) Zie je of er punten samenvallen?

De punten op het spreidingsdiagram liggen duidelijk zeer dicht bij een rechte lijn. We zeggen dat er een ***benaderend lineair verband*** is tussen het gewicht en het voetoppervlak van de slakken.

(b) Welk soort correlatie is er hier?

Natuurlijk is er niet altijd een lineair verband met een positieve correlatie tussen de twee variabelen. Andere mogelijkheden zijn geïllustreerd in de volgende spreidingsdiagrammen:

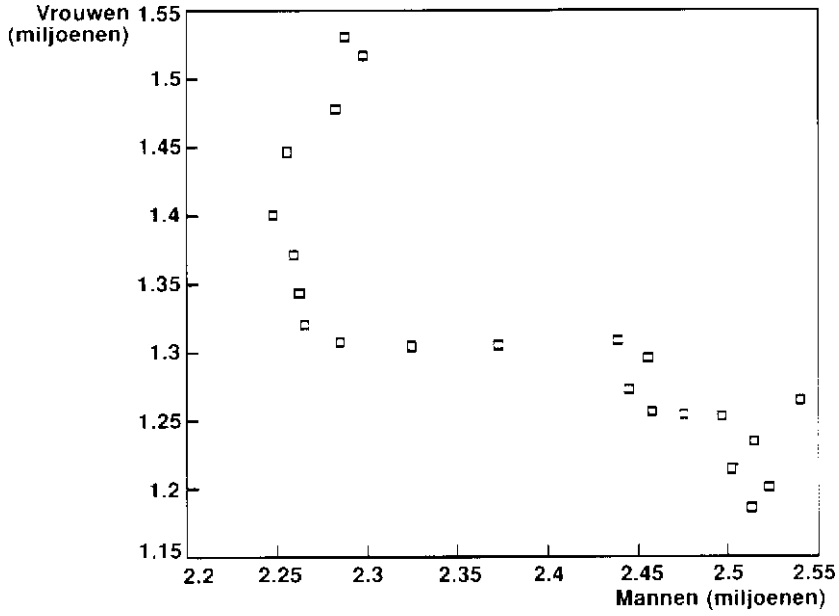


(c) Hoe zou jij, zo nauwkeurig mogelijk, het verband tussen X en Y beschrijven in deze drie gevallen?

Het doel van de komende lessen is te onderzoeken wanneer het aanvaardbaar is een lineaire benadering te gebruiken voor de relatie tussen twee variabelen. En, wanneer dit zo is, de vergelijking te vinden van die best benaderende rechte.
Het volgende voorbeeld toont ons een duidelijk niet-lineair verband.

Voorbeeld 4.

Van het kabinet van het Ministerie van Tewerkstelling en Arbeid kregen we cijfers over de evolutie van de tewerkgestelde beroepsbevolking per geslacht in de periode 1970-1991. Hieruit konden we afleiden dat een groter aantal tewerkgestelde vrouwen gepaard gaat met een kleiner aantal tewerkgestelde mannen.



Kunnen we hier spreken van een lineair verband?

HOOFDSTUK 3:
COVARIANTIE ALS SPREIDINGSMAAT.

A. Opbouw en interpretatie van het begrip covariantie.

De bedoeling van dit hoofdstuk is een eerste maat vinden waarmee we spreiding kunnen uitdrukken bij een bivariate verdeling. We willen over spreiding kunnen spreken aan de hand van een getal en niet meer alleen op basis van het spreidingsdiagram.
In de volgende werktekst is het de bedoeling dat jullie het begrip covariantie als spreidingsmaat zelf opbouwen. Let hierbij op dat je elke stap die genomen wordt, goed begrijpt en stel je hierbij steeds de volgende vraag:

Waarom zou deze stap nodig/nuttig zijn?

Besteed genoeg aandacht aan elke vraag.

Mannelijke krekels sjirpen door hun vleugels tegen elkaar te wrijven. Men wil het verband tussen de sjirpfreqentie en de temperatuur nagaan in drie verschillende landen: België, Zweden en Frankrijk.
Hieronder staan de (fictieve) gegevens voor 12 verschillende testen per land:

België		Zweden		Frankrijk	
temperatuur (°C) X	sjirpfreqentie Y	temperatuur (°C) X	sjirpfreqentie Y	temperatuur (°C) X	sjirpfreqentie Y
17	8.4	13	7.7	24	3.5
19	8.6	14	6.1	28	4.5
20	7.8	17	9.2	31	2.8
22	7.2	18	7.9	33	2.9
25	7.1	18	10.4	34	3.4
26	6.1	18	11.3	35	2.0
28	6.5	21	11.7	35	5.0
29	5.8	22	13.0	38	3.8
31	5.4	22	14.9	41	4.4
33	4.8	23	13.9	43	2.9
33	4.2	26	14.8	48	4.3
35	4.8	26	15.8	49	3.5

(1) Teken op je rekentoestel het spreidingsdiagram voor de drie landen.

💻 **Tips voor de TI-83:**

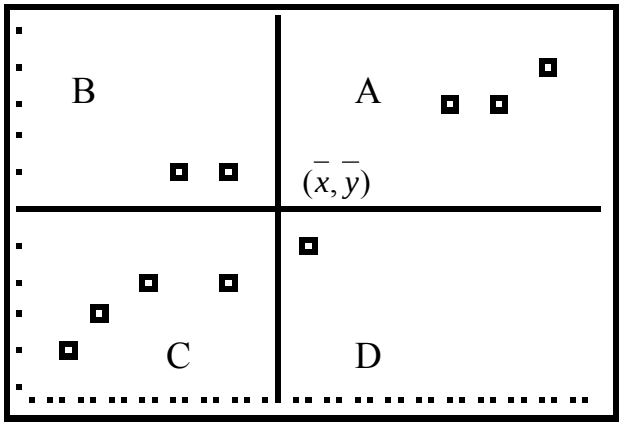
- Gebruik de lijsten L1 en L2 voor België, L3 en L4 voor Zweden en L5 en L6 voor Frankrijk.
  Pas dan ook telkens Xlist en Ylist, wanneer je plot1, plot2 en plot3 definieert, hieraan aan.
- Om van het scherm met het spreidingsdiagram terug te keren naar het basisscherm, druk je op 2nd MODE (QUIT).
- Vergeet niet plot1 en plot3 op off te zetten wanneer je bv plot2 gaat tekenen, anders komen de twee spreidingsdiagrammen gewoon over elkaar.

(2) Bekijk de spreidingsdiagrammen en becommentarieer de correlaties tussen de temperatuur en de sjirpfrequentie voor elk land afzonderlijk.

Het doel van dit werkblad is correlatie nauwkeuriger te kunnen beschrijven. Tot nu toe baseerden we ons alleen op het spreidingsdiagram om te oordelen of er positieve of negatieve correlatie is en of de correlatie sterk is of zwak.
We willen nu een getal gaan ontwikkelen waarmee we de correlatie kunnen beoordelen.

Wanneer we maar één enkele variabele X hebben, geeft de variantie de spreiding weer tov het gemiddelde $\bar{x}$.
Bij een bivariate verdeling, zijn er twee variabelen van belang X en Y. We hebben hier twee gemiddelden $\bar{x}$ en $\bar{y}$ voor handen.
Om nu de spreiding van de punten in de bivariate verdeling te kunnen beschrijven, moeten we rekening houden met beide gemiddelden $\bar{x}$ en $\bar{y}$.
We tekenen dus in het spreidingsdiagram een nieuwe x- en y-as door het punt $(\bar{x}, \bar{y})$.

Het volgende diagram toont opnieuw het spreidingsdiagram van het voorbeeld van de kandidaat-soldaten.



We weten dat de lengte en de schoenmaat van de kandidaat-soldaten positief gecorreleerd zijn.
Vraag 3 verwijst naar dit spreidingsdiagram.

(3) Deze vraag verwijst naar het spreidingsdiagram hierboven.
   (a) In welke twee kwadranten (A, B, C of D) liggen de meeste punten? Waardoor komt dit?

   (b) Moesten de lengte en de schoenmaat negatief gecorreleerd zijn, in welke twee kwadranten zouden dan het grootst aantal punten liggen? Waarom?

   (c) En als de twee variabelen ongecorreleerd zijn, in welke kwadranten liggen dan de meeste punten?

   (d) Als een punt (x,y) in kwadrant A ligt, wat is dan het teken van
      - $x - \bar{x}$

- $y - \bar{y}$
- $(x - \bar{x})(y - \bar{y})$?

Beantwoord nu dezelfde vragen wanneer (x,y) in respectievelijk de kwadranten B, C en D ligt.
Vul alles in, in de volgende overzichtstabel: schrijf een "+" voor positief en een "–" voor negatief.

	A	B	C	D
$x - \bar{x}$				
$y - \bar{y}$				
$(x - \bar{x})(y - \bar{y})$				

Beschouw nu $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$.

(e) Wat zal het teken zijn van $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ als de twee variabelen positief gecorreleerd zijn?
Houd hiervoor rekening met (a) en de tabel in (d).
Leg uit waarom.

(f) Wat zal het teken zijn van $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ als de twee variabelen negatief gecorreleerd zijn?
Houd hiervoor rekening met (b) en de tabel in (d).
Leg ook hier uit waarom je dit denkt.

(g) Wat zal het teken zijn van $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ als de twee variabelen ongecorreleerd zijn?
Houd hiervoor rekening met (c) en de tabel in (d).
Leg ook hier uit waarom.

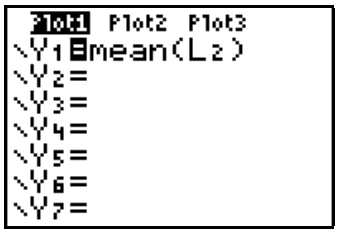
(h) Leg uit (mbv de drie vorige vragen) waarom $\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$ een goede maat is voor correlatie.

(4) Herneem nu het voorbeeld van de sjirpende krekels uit vraag 1.
Vind voor België $\bar{x}$ en $\bar{y}$.

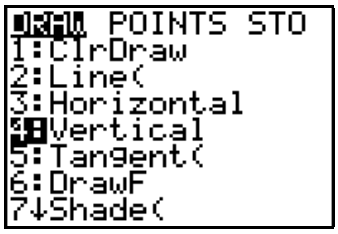
Teken dan op het bijbehorend spreidingsdiagram de twee evenwijdige assen aan de oorspronkelijke assen, door het punt ($\bar{x}$, $\bar{y}$).

Tips voor de TI-83:

Om het gemiddelde te vinden van de elementen van een lijst, gebruik je: 2nd STAT (LIST) MATH 3:mean(
en dan typ je de naam van de lijst in.



Om een horizontale lijn tesamen met je spreidingsdiagram te tekenen, druk je Y=. Je vult dan in bv mean(L2) zoals het hierboven beschreven staat.
Kijk wel na of je in de juiste plot werkt: ben je bezig met de gegevens van België, dus in plot1, moet dit aangeduid zijn bovenaan het scherm in Y=.



Om nu nog een verticale erbij te tekenen, druk je op:
2nd MODE (QUIT) om terug te keren naar het basisscherm, dan op 2nd PRGM (DRAW) DRAW 4:Vertical.
Je typt dan (mean(L1)) zoals aangeleerd hier hoger en drukt op ENTER.
En dan verschijnt het spreidingsdiagram met de twee extra rechten.

(5) Bereken nu voor België $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ en leg de betekenis uit van dit resultaat. Gebruik hierbij je bevindingen in vraag 3(f) en in vraag 4.

(6) Doe nu hetzelfde voor de twee andere landen.

B. Eigenschappen en zwakheden van de covariantie als spreidingsmaat.

Veronderstel dat de steekproef $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ genomen uit de bivariate verdeling (X,Y), een steekproefgemiddelde $(\bar{x}, \bar{y})$ heeft.
Dan definiëren we de covariantie van de variabelen X en Y als:

$$Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Wanneer de covariantie ***dicht bij 0*** ligt, wil dit zeggen dat er weinig of geen correlatie is tussen de desbetreffende variabelen.
Een ***"grote" positieve covariantie*** wijst op een positieve correlatie tussen de twee variabelenen andersom wijst een ***"grote" negatieve covariantie*** op een negatieve correlatie tussen de twee variabelen.

We tonen nu de zwakheid aan van de covariantie als spreidingsmaat:

☒ Voorbeeld:

We hernemen het voorbeeld van de kandidaat-soldaten.
De lengte X in de bivariate verdeling (X,Y) wordt nu uitgedrukt in centimeters.
Wat zal het effect zijn op de covariantie als ze uitgedrukt zal worden in meters?

Oplossing:
Elke lengte uitgedrukt in centimeters, X_i, wordt vervangen in de berekeningen door een lengte uitgedrukt in meters, $\frac{1}{100}X_i$.
We weten dat dan ook het gemiddelde door 100 zal moeten gedeeld worden.
De covariantie verandert van $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ naar $\frac{1}{n}\sum_{i=1}^{n}(\frac{1}{100}x_i - \frac{1}{100}\bar{x})(y_i - \bar{y})$.
De laatste uitdrukking kan geschreven worden als:

$$\frac{1}{n}\frac{1}{100}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

En zo wordt de oorspronkelijke covariantie 100 keer kleiner.

Wat gebeurt er met Cov(X,Y) als alle waarden van X vermenigvuldigd worden met 5 en alle waarden van Y met 4?

Het is duidelijk dat begrippen als "een grote covariantie" heel relatief zijn en afhangen van de grootte van de variabelen en hun eenheden.
We kunnen dus de grootte van de covariantie niet gebruiken om over spreiding te spreken.
Een nieuwe spreidingsmaat dringt zich op.

HOOFDSTUK 4:
DE CORRELATIECOËFFICIËNT ALS BETERE SPREIDINGSMAAT.

A. De correlatiecoëfficiënt met zijn belangrijkste eigenschappen.

In het vorige hoofdstuk zagen we dat de covariantie toch niet zo geschikt bleek te zijn als spreidingsmaat.
Daarom voeren we een nieuw begrip in: de correlatiecoëfficiënt.

De correlatiecoëfficiënt vinden we door de covariantie te delen door het product van de standaardafwijkingen van X en Y.

$$r = \frac{Cov(X,Y)}{s_X s_Y}$$

⌦ Oefening:

Wat gebeurt er met r als alle waarden van X vermenigvuldigd worden met 5 en alle waarden van Y met 4? Is de correlatiecoëfficiënt nog eenheidsgebonden?

⌦ Oefening:

Vind de correlatiecoëfficiënt voor de bivariate verdeling van het voorbeeld van de slakken.

Gewicht (g)	0.64	0.21	0.85	0.53	0.02	0.01	0.21	0.18	0.06	0.20	0.07	0.01	0.05
Voetopp (mm^2)	29	16	35	25	4	1	16	20	7	13	7	3	10

Gewicht (g)	0.81	0.53	0.18	0.06	0.20	0.07	0.01
Voetopp (mm^2)	35	25	20	7	13	7	1

Gebruik hierbij de TI-83 zo nuttig mogelijk om de tussenresultaten te controleren. Met de toets STAT CALC 2:2-Var Stats, krijg je alle nuttige kentallen van zowel X als Y tegelijk.

⌦ Oefening:

Er is een verband tussen het aantal voertuigen in Nederland en het aantal verkeersongevallen per jaar.
In de jaren ’70 waren de aantallen als volgt:

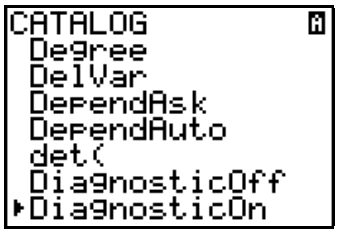
jaar	‘70	‘71	‘72	‘73	‘74	‘75	‘76	‘77	‘78	‘79
Voertuigen (milj) X	2.6	3.1	3.5	3.7	4.1	4.4	4.6	4.9	5.3	5.8
Ongelukken (x 1000) Y	138	163	166	153	177	201	216	208	226	238

Bereken de correlatiecoëfficiënt.

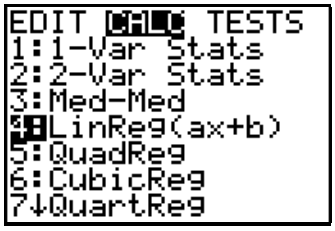
💻 Met de TI-83

We kunnen met de TI-83 ook rechtstreeks de correlatiecoëfficiënt berekenen.

Plaats hiertoe je data van de stochastische veranderlijke X in de lijst L1 en die van Y in de lijst L2.
Druk op 2nd 0 (CATALOG) en ga naar DiagnosticOn, druk dan tweemaal op ENTER. Dit moet je maar één keer doen met je toestel, dit dient om extra gegevens te krijgen, waaronder de correlatiecoëfficiënt.



Druk dan op STAT CALC 4: LinReg(ax+b).
En daar verschijnt dan r.
Wat a en b betekenen, zal in het volgende hoofdstuk uitgelegd worden.



⌦ Oefening:

We weten uit de definitie dat $r = \frac{Cov(X,Y)}{s_X s_Y}$.

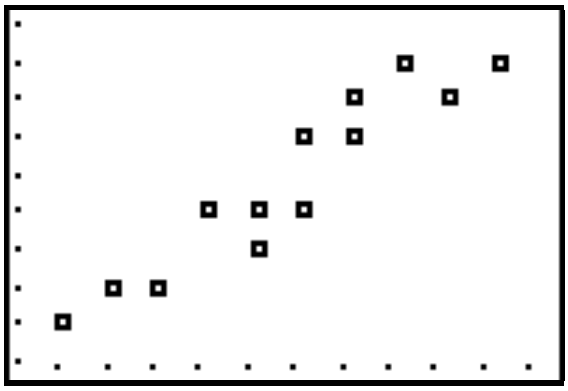
(a) Wat is het teken van Cov(X,Y) als X en Y positief gecorreleerd zijn?

(b) En wat is het teken van Cov(X,Y) als X en Y negatief gecorreleerd zijn?

(c) Wat is dan het teken van r in beide gevallen? Waarom?

☒ Voorbeeld:

We hernemen het voorbeeld van de schoonheidswedstrijd voor honden:
Wanneer we een spreidingsdiagram tekenen van de scores van jury A tov van die van jury B, bekomen we het volgende:

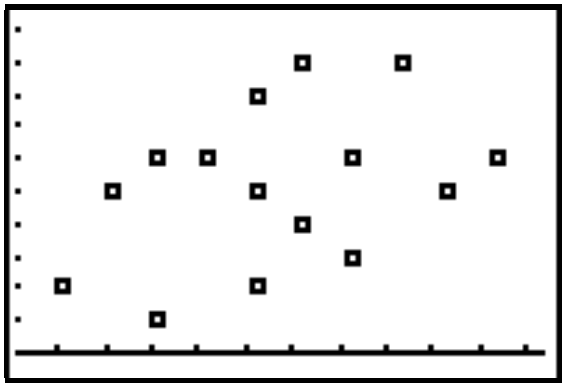


(a) Welk soort correlatie is er hier ook weer?

(b) Welk teken zou dan de correlatiecoëfficiënt moeten hebben?

Wanneer we deze met ons rekentoestel berekenen, bekomen we dat r = 0.95.

Wanneer we nu het spreidingsdiagram tekenen van de scores van jury A ten opzichte van die van jury C, bekomen we:



(c) Welk soort correlatie is er hier?

(d) Welk teken zou dan de correlatiecoëfficiënt moeten hebben?

Wanneer we deze met ons rekentoestel berekenen, bekomen we dat r = 0.39.

(e) Welk zou, denk je, de bovengrens zijn van de correlatiecoëfficiënt als er een positieve correlatie is? Wat is het verband tussen deze bovengrens en de sterkte van de correlatie?

(f) Heb je een vermoeden wat r zal zijn bij zwakke negatieve correlatie en bij sterke negatieve correlatie?

Controleer je vermoeden met het volgende voorbeeld:

☒ Voorbeeld:

We hernemen het voorbeeld van de sjirpende krekels in België. Daar waren de temperatuur en de sjirpfrequentie sterk negatief gecorreleerd.

temperatuur (°C)	17	19	20	22	25	26	28	29	31	33	33	35
sjirpfrequentie	8.4	8.6	7.8	7.2	7.1	6.1	6.5	5.8	5.4	4.8	4.2	4.8

(a) Bereken r en kijk na of je vermoeden van in de vorige vraag klopt?

(b) Welke ondergrens is er, denk je, voor r?

Tot slot, kijken we nog eens wat de correlatiecoëfficiënt gaat zijn, als de twee variabelen ongecorreleerd zijn:

☒ Voorbeeld:

We hernemen het voorbeeld van de sjirpende krekels in Frankrijk. Daar waren de temperatuur en de sjirpfrequentie ongecorreleerd.

temperatuur (°C)	24	28	31	33	34	35	35	38	41	43	48	49
sjirpfrequentie	3.5	4.5	2.8	2.9	3.4	2.0	5.0	3.8	4.4	2.9	4.3	3.5

(a) Bereken r.

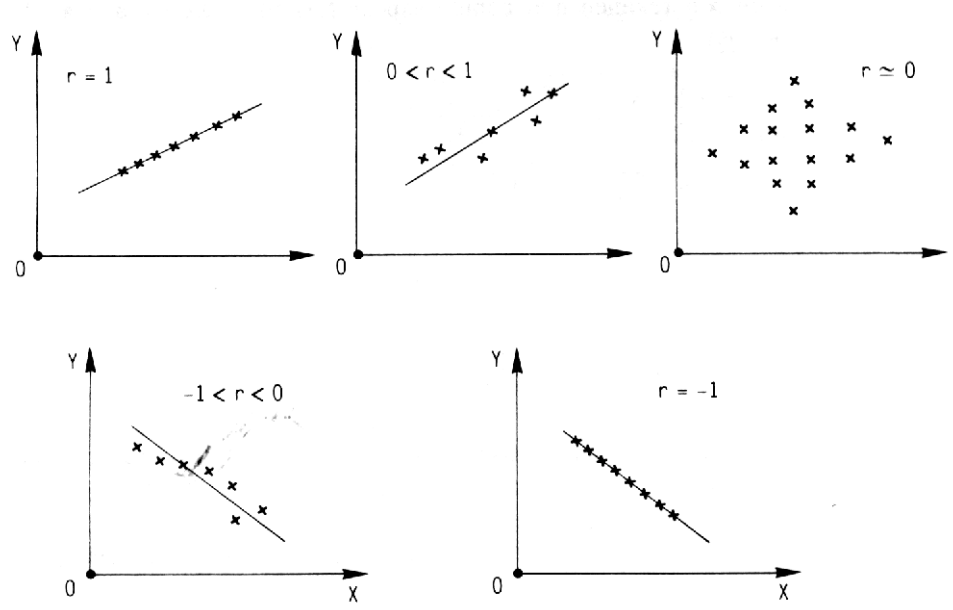
(b) Hoe zal de correlatiecoëfficiënt zijn als twee variabelen ongecorreleerd zijn?

Daar waar de covariantie ons weinig betekenis en verklaring kon geven omwille van het subjectieve idee van “groot” en “klein”, kan de correlatiecoëfficiënt ons meer exacte informatie geven.

- Wanneer r dicht ligt bij 1, kan dit wijzen op een sterke tendens dat grote x_i-waarden overeenkomen met grote y_i-waarden en dat kleine x_i-waarden overeenkomen met kleine y_i-waarden.
  We spreken van een ***sterke positieve lineaire correlatie***.
- Wanneer r dicht ligt bij -1, kan dit wijzen op een sterke tendens dat kleine x_i-waarden overeenkomen met grote y_i-waarden en omgekeerd.
  We spreken van een ***sterke negatieve lineaire correlatie***.
- Wanneer r dicht ligt bij 0, kan dit erop wijzen dat er geen bepaalde tendens is onder de koppels (x_i, y_i).
  We spreken van een ***zwakke lineaire correlatie***.

Wanneer r dicht ligt bij 1, verwachten we dus dat de punten (x_i, y_i) dicht liggen bij een rechte met positieve richtingscoëfficiënt. Terwijl als r dicht ligt bij –1, verwachten we dat de punten (x_i, y_i) dicht liggen bij een rechte met negatieve richtingscoëfficiënt.

Tot slot geven de volgende figuren een overzicht van de mogelijke r-waarden.



B. Enkele oefeningen.

(1) De zee- en luchttemperatuur op een stukje strand in Florida werd, gedurende 10 weken, elke maandagmiddag gemeten.
Dit leverde de volgende gegevens op:

zee X (°C)	19	22	18	19	21	22	18	18	17	16
Lucht Y (°C)	29	34	27	29	33	35	28	27	26	25

(a) Maak een spreidingsdiagram met je rekentoestel.
(b) Bereken r en verklaar je resultaat.

(2) Teken de volgende gegevens in één spreidingsdiagram.
Set 1:

X	1	4	5	7
Y	11	5	3	-1

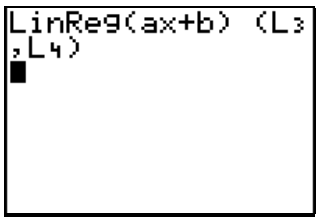
Set 2:

X	0	2	5	6
Y	20	12	0	-4

Duid hierbij de gegevens van set 1 aan met een kruisje en die van set 2 met een vierkantje.
Bereken r voor elk set en becommentarieer het resultaat.

💻 **Tips voor de TI-83.**

Om r te berekenen voor 2 lijsten (indien er meerdere lijsten zijn), druk je STAT CALC 4:LinReg(ax+b) en je typt (L3,L4) bijvoorbeeld voor de lijsten L3 en L4. En dan typ je ENTER.



(3) Men mat de periode T (in seconden) van 7 staanklokken van een verschillende hoogte. Dit gaf de volgende resultaten:

H (cm)	10	20	30	40	50	60	70
T (s)	0.63	0.90	1.10	1.27	1.42	1.56	1.68

(a) Bereken r voor deze gegevens.
(b) Teken het spreidingsdiagram.
(c) Denk je dat de relatie tussen H en T lineair is? Waarom?

(4) Teken een spreidingsdiagram van de volgende vier punten: (1,1), (1,3), (3,1) en (3,3).
Bereken r en verklaar zijn waarde.

C. Enkele bijzonderheden.

(1) Beschouw de volgende hypothetische scores (op 100) op twee examens:

examen 1	49	52	55	58	61	64	67	70	73	76	79	82	85	88	91
examen 2	95	81	69	59	51	45	41	39	41	45	51	59	69	81	95

(a) Teken het spreidingsdiagram met je rekentoestel.
Lijkt er een verband te zijn tussen de twee examenscores?
Zo ja, beschrijf deze relatie.

(b) Bereken de correlatiecoëfficiënt.
Verbaast dit resultaat je? Wat had je verwacht?

Dit voorbeeld illustreert dat de correlatiecoëfficiënt enkel een lineair verband meet tussen twee variabelen. Meer ingewikkelde relaties kunnen met r niet opgemerkt worden.
Dus kan er een verband bestaan tussen twee variabelen, zelfs als de correlatiecoëfficiënt dicht bij 0 ligt. Je moet je dus bewust zijn van deze mogelijkheid en je niet louter baseren op de waarde van r om een besluit te rekken. Onderzoek zeker ook steeds het spreidingsdiagram.

(2) Beschouw de spreidingsdiagrammen van de volgende hypothetische examenscores: teken ze met je rekentoestel.
A:

examen 1	49	52	55	58	61	64	67	70	73	76	79	82	85	88	99
examen 2	54	57	60	62	65	68	70	73	76	78	81	84	87	89	12

B:

examen 1	12	52	55	58	61	64	67	70	73	76	79	82	85	88	91
examen 2	17	62	52	73	69	71	72	80	58	60	67	52	76	69	70

(a) In klas A lijken de meeste observaties een lineair patroon te volgen. Zijn er uitzonderingen?

(b) Terwijl in klas B lijken de meeste observaties eerder willekeurig geplaatst zonder een echt patroon. Zijn er hier uitzonderingen?

(c) Bereken voor beide klassen de correlatiecoëfficiënt. Ben je verrast over één of beide resultaten? Waarom?

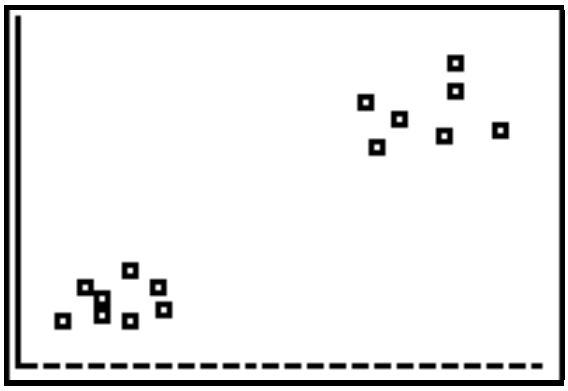
Een punt dat volledig uitspringt uit het patroon in het spreidingsdiagram, noemen we een ***uitschieter***. Dit is meestal te wijten aan een foute meting of een vergissing bij het opschrijven van de gegevens.
Maar soms zijn de gegevens juist en gaat het gewoon om een uitzondering:

- Een leerling met een dikke buis, terwijl de rest van de klas bijna het maximum haalde.
- Een slak met uitzonderlijk kleine voetjes.
- Een hele oude man die trouwt met een vrouw van 20 jaar.
- …

Uitschieters hebben vaak een grote invloed op de correlatiecoëfficiënt waardoor we soms verkeerde besluiten zouden trekken.
Het belang van het spreidingsdiagram te bekijken, is ook hier weer bewezen.
Meestal laat men voor de berekening van r de uitschieters gewoon weg.

(d) Verwijder de uitschieters uit beide klassen en bereken opnieuw de correlatiecoëfficiënt. Becommentarieer hoe deze veranderd zijn en leg uit.
💻 **Tip voor de TI-83.**
Om een element in een rij (lijst) te verwijderen, druk je STAT EDIT, je gaat op het element staan en drukt DEL.
Let op: vergeet het overeenkomstige element niet te verwijderen, als het over een bivariate verdeling gaat.

(3) Beschouw het volgende spreidingsdiagram van een set hypothetische examenscores.



De gegevens zijn:

examen 1	37	44	32	37	35	41	41	45	88	72	75	81	82	71	82
examen 2	39	41	33	34	41	45	33	36	78	74	81	77	94	84	87

(a) Beschrijf wat het spreidingsdiagram jou vertelt over het verband tussen de twee examenresultaten.

(b) Bereken r. Is zijn waarde hoger dan je verwachtte?

Hier zien we dat, zelfs als er geen uitschieters zijn of geen ingewikkelder verband, de correlatiecoëfficiënt toch nog groot kan zijn, hoewel er geen lineaire relatie is tussen de twee variabelen.

(4) De correlatiecoëfficiënt wordt door onderzoekers gebruikt in veel domeinen van sociale wetenschappen tot landbouwwetenschappen. In het algemeen, probeert men aan de hand van de correlatiecoëfficiënt, te bewijzen dat de verandering van één eigenschap leidt tot de verandering van iets anders.
Bijvoorbeeld dat de stijgende werkloosheid een stijging in criminaliteit veroorzaakt.
Eigenlijk kan een resultaat waarbij $r \approx 1$ of $r \approx -1$ op drie verschillende manieren geïnterpreteerd worden:

Als Y stijgt wanneer X stijgt:
- kan de stijging van X de stijging van Y veroorzaakt hebben of omgekeerd, is de stijging van Y de oorzaak van het stijgen van X.
- kunnen beide stijgingen een gemeenschappelijke oorzaak hebben.
- kunnen beide stijgingen totaal niets met elkaar te maken hebben.

Het is dan de taak van de onderzoeker om uit te maken op welke van de drie manieren r moet geïnterpreteerd worden.
Hieruit volgt dus dat op zich, een resultaat $r \approx 1$ of $r \approx -1$, geen informatie geeft over "het veroorzaken".
Om dit te illustreren, bekijken we het volgende voorbeeld:
De volgende tabel geeft informatie over de levensverwachting van de inwoners van 22 landen. Het geeft ook het aantal mensen per televisietoestel in elk land.

land	levensverwachting	Mensen per TV	land	levensverwachting	Mensen per TV
Angola	44	200	Mexico	72	6.6
Australië	76.5	2	Marokko	64.5	21
Cambodja	49.5	177	Pakistan	56.5	73
Canada	76.5	1.7	Rusland	69	3.2
China	70	8	Zuid-Afrika	64	11
Egypte	60.5	15	Sri Lanka	71.5	28

Frankrijk	78	2.6	Oeganda	51	191
Haïti	53.5	234	UK	76	3
Irak	67	18	VS	75.5	1.3
Japan	79	1.8	Vietnam	65	29
Madagaskar	52.5	92	Jemen	50	38

(a) Welk land heeft het minst aantal mensen per televisietoestel? En welk land het meest? Wat betekenen juist deze getallen?

(b) Teken met je rekentoestel het spreidingsdiagram van de levensverwachting versus het aantal mensen per televisietoestel.
Lijkt er een verband te zijn tussen de twee variabelen? Beschrijf dit verband kort.

(c) Bereken r.

(d) Omdat de correlatie zo sterk negatief is, zou men kunnen besluiten dat men, in de landen met een lagere levensverwachting, de mensen langer kan doen leven door veel televisietoestellen naar die landen te sturen.
Becommentarieer deze uitspraak.

(e) Welke van de drie hoger beschreven factoren verklaart hier de stijging van de levensverwachting in functie van de stijging van het aantal mensen per televisietoestel? Leg uit.

Pas nu deze bijzonderheden toe op de volgende oefeningen:

⌦ Oefening:

We hebben gezien dat er drie verschillende manieren bestaan om een sterke correlatie te beoordelen.
(a) In welke categorie zou je het verband tussen lengte en gewicht plaatsen?
(b) In de jaren ’80 was er een stevige toename in het aantal studenten in Sheffield. In dezelfde stad was er toen ook een ferme stijging in autodiefstallen.
In welke categorie zou je dit voorbeeld plaatsen?

⌦ Oefening:

In veel gemeenschappen vindt men een sterke positieve correlatie tussen de smaak van ijs, die in een gegeven maand het meest verkocht wordt en het aantal verdrinkingen door zelfmoord die zich die maand voordoen.
Betekent dit dat ijscrème verdrinking veroorzaakt? Indien niet, kan je dan een alternatieve verklaring geven voor deze sterke correlatie?

D. Individueel project.

De bedoeling bij dit individueel project is, dat je duidelijk laat merken dat je alles tot nu toe goed begrijpt. Het is dan ook een soort van controle voor jezelf: als je, tijdens het maken van deze opdracht, ergens moeilijkheden mee hebt, ga dat onderdeel dan terug bekijken in je nota's. En maak eventueel enkele oefeningen opnieuw of vraag uitleg.
Je krijgt ruim de tijd om je gegevens te verzamelen, een grondige analyse uit te voeren en je verslag te maken. Achteraf is het dan ook de bedoeling dat jullie je project kort komen presenteren voor je klasgenoten.
Zorg dus dat je alles volledig door hebt, zodat je eventuele vragen rustig kan beantwoorden. Steek hier voldoende tijd in zodat je zeker bent van je analyse en je besluiten.

De opdracht zelf nu:

- Kies twee variabelen die gemeten of geteld kunnen worden en waarvan je vermoedt dat er een bepaalde relatie tussen bestaat. De variabelen kunnen zowel eigenschappen van mensen, van dieren als van dingen zijn.
  Denk hier lang genoeg over na, neem niet de eerste de beste. Probeer er twee variabelen uit te kiezen waarvan hun relatie je klasgenoten zal verbazen en zal interesseren.
  Misschien kun je hierover onderling wat brainstormen.
- Verzamel gegevens voor de twee gekozen variabelen en dit bij ten minste 20 verschillende bronnen.
  Vb. 20 verschillende mensen, honden, testen, appelsienen, …
- Toon dat je goed begrijpt wat je twee variabelen juist betekenen, door ze uitgebreid in woorden te beschrijven.
- Leg ook uit hoe je je gegevens hebt verzameld en gemeten. Leg uit hoe je ervoor gezorgd hebt dat je gegevens representatief zijn. (Bv. Neem niet alle 20 appelsienen uit de Delhaize, want zo zou je steekproef onderhevig kunnen zijn aan een externe factor omdat bv. De Delhaize altijd vrij kleine appelsienen verkoopt.)
- Maak een spreidingsdiagram voor je variabelen. Zeg welke variabele je kiest op de X-as en welke op de Y-as en leg je keuze uit.
  Wees nauwkeurig bij je tekening. Je kan controleren met je rekentoestel.
- Geef uitleg over het verband dat er lijkt te zijn tussen de twee gekozen variabelen. Kun je formuleren waarom je dit verband denkt te zien?
  Als je denkt dat er geen relatie is, verklaar dit dan ook. (en kies eventueel opnieuw twee variabelen)
- Toon dat je de relatie volledig begrijpt door ze duidelijk en volledig te beschrijven in woorden: pas hiervoor je statistische besluiten toe op het expliciete voorbeeld, op de twee variabelen dus. Wat betekent de relatie voor deze twee variabelen.
  Kan je nu voorspellingen doen over mensen (honden, testen, appelsienen,…) waarvan je de gegevens niet gemeten hebt?
- Bereken de correlatiecoëfficiënt en geef hier wat uitleg over: wat soort correlatie geeft r aan, klopt deze correlatie met de werkelijkheid, …
- Kan je zeggen dat de ene variabele een stijging of daling in de andere variabele veroorzaakt? Of hoe interpreteer jij anders het verband tussen de twee variabelen? Is er een verborgen factor die beide variabelen beïnvloedt, of is hun verband louter toevallig?

Probeer nu aan de hand van deze vragen een aaneenhangende tekst te schrijven die zo goed en zo duidelijk mogelijk jouw analyse weergeeft.

Houd, terwijl je hieraan werkt, het zelfevaluatieblad dat op de volgende bladzijde staat, zorgvuldig bij en vul telkens de datum in wanneer je een onderdeel volledig beëindigd hebt. Wanneer je werkje af is, vul dan het tweede deel in op het zelfevaluatieblad en gebruik dit blad als “cover” van je werkje.

ZELFEVALUATIEBLAD.

Deel 1.

Gebruik dit blad als de "cover" van je werkje. Schrijf telkens de datum naast een onderdeel wanneer je dit volledig hebt afgewerkt.

…… Ik heb gegevens verzameld om te zien of de twee variabelen gecorreleerd zijn.
De twee variabelen zijn:
……
……

…… Het spreidingsdiagram is netjes getekend en alle eenheden liggen even ver van elkaar.

…… Ik kan exact formuleren wat de twee variabelen betekenen.

…… Ik weet welk verband er is tussen de twee variabelen en heb hiervan een geschreven weergave gemaakt.

…… Ik heb een geschreven uitleg gemaakt over het feit of er een oorzaak is waardoor de ene variabele de andere beïnvloedt of of er een externe factor is.

…… Ik denk dat mijn werk nu volledig is.
Ik vind zelf dat mijn werk (omcirkel)
een grondige studie is
voldoende informatie bevat
nog niet voldoende onderzoek toont.

Deel 2.

Beantwoord de volgende vragen eerlijk nadat je hele werkje af is.

(1) Vind je dat je de leerstof nu, na het maken van dit project, beter begrijpt? Of heb je liever dat je gewoon de theorie moet instuderen?

(2) Vond je het een leuke oefening of verkies je toch gewoon wat meer oefeningen in de les?

(3) Vond je dat je er teveel tijd in moest steken? Of vond je die tijd nuttig besteed voor het volledig begrijpen van de leerstof?

(4) Vond je het moeilijk om alles mooi in woorden uit te drukken?

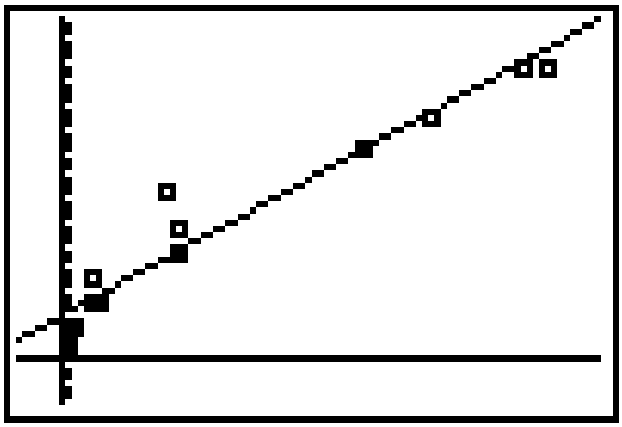
(5) Kon je gemakkelijk twee “sprekende” en “originele” variabelen vinden?

(6) Vind je zo een project voor een herhaling vatbaar op een ander stukje leerstof?

HOOFDSTUK 5:
DE REGRESSIERECHTE.

A. Probleemschets.

In veel voorbeelden, zoals bijvoorbeeld het voorbeeld van de slakken, zagen we dat de punten in het spreidingsdiagram duidelijk zeer dicht bij een rechte lijn liggen. We spraken dan van een ***benaderend lineair verband*** tussen de twee variabelen.
Eens we vermoeden dat er een lineair verband zou kunnen zijn tussen de twee variabelen, moeten we proberen de vergelijking te vinden van de rechte die het best aansluit bij de puntenwolk. Deze rechte noemen we de ***regressierechte***.



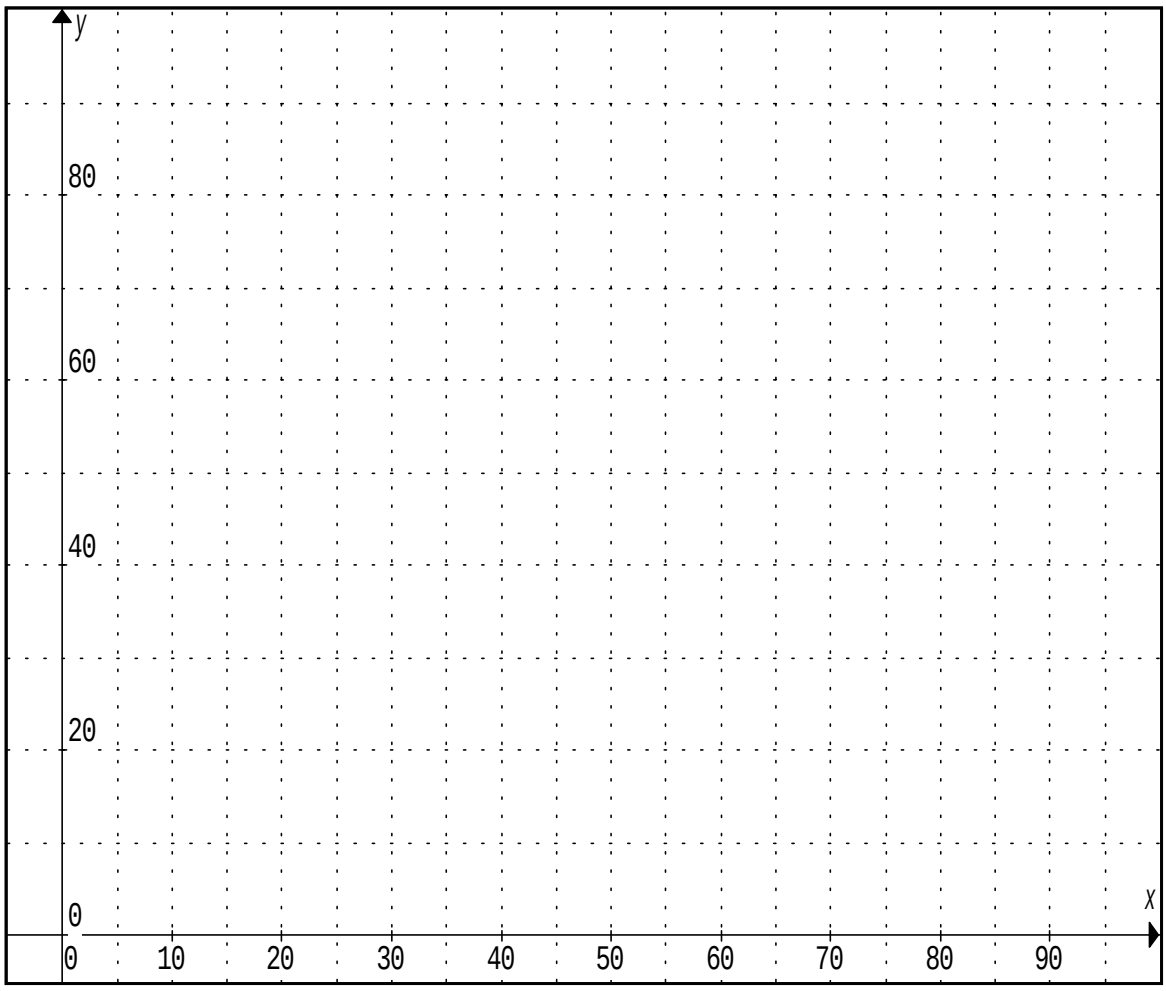
De techniek die gebruikt wordt om wiskundig de vergelijking van de regressierechte te vinden, noemt men dan ***regressie***. De regressierechte wordt o.a. gebruikt om voorspellingen te kunnen doen.

⌦ Oefening:

In Engeland doet men een onderzoek naar de verschillen in het wiskundeonderwijs in de verschillende scholen. Daartoe bekijkt een "moderator" de quotaties van de huiswerken van de leerlingen en past de punten zodanig aan dat de strengheid van verbeteren in elke school dezelfde is.
Natuurlijk kan hij niet elke leerling behandelen en kiest er daarom in elke school 10 leerlingen uit, waarvan de punten heel verscheiden zijn. In een bepaalde school gaf dit de volgende resultaten:

punten van de leraar X	15	24	34	43	47	56	60	70	83	90
punten van de moderator Y	27	36	37	48	47	52	65	66	70	80

(a) Teken een spreidingsdiagram van de gegevens met de hand.



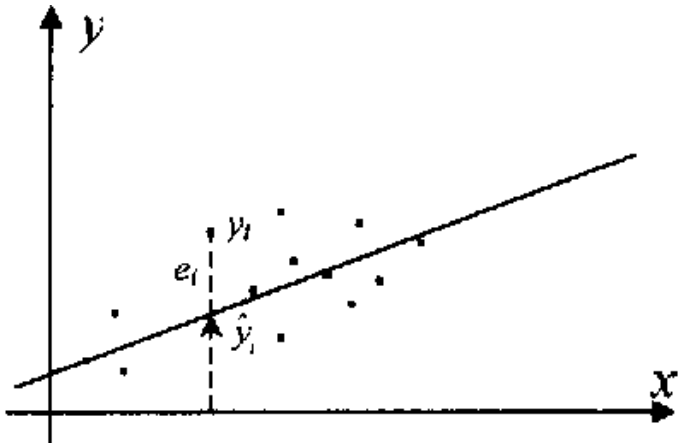
(b) Zoek de correlatiecoëfficiënt en becommentarieer hiermee de correlatie tussen X en Y.

(c) Schat, op het zicht, de rechte die het best aansluit bij de gegevens. Teken deze rechte op het spreidingsdiagram.
Ken je een specifiek punt waar de lijn zal doorgaan? Indien ja, bereken de coördinaten van dat punt en kijk of je geschatte rechte er door loopt.

(d) Gebruik je geschatte rechte om de punten van de moderator te schatten voor een leerling die van zijn eigen leraar 20 punten kreeg.

(e) Hoeveel zouden de punten van de leraar geweest zijn voor een leerling die 56 punten van de moderator kreeg?

Wanneer men gegevens verzamelt van een bivariate verdeling, zijn de X-gegevens meestal de gegevens die onder controle zijn van de persoon die het experiment uitvoert. De Y-waarden daarentegen zullen afhangen van deze X-waarden.
Veronderstel dat je een lineaire relatie $y = ax + b$ vermoedt tussen de onafhankelijke variabele X en de afhankelijke variabele Y. We spreken van "***regressie van Y op X***".
Hierbij bekijken we het verschil tussen de ***geobserveerde*** waarde van Y (y_i) en de ***voorspelde*** waarde van Y ($\hat{y}_i$), die we uit de vergelijking van de rechte halen.



Gegeven is een puntenwolk van n punten: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, die min of meer een lineaire trend vertonen.
Beschouw dan de rechte $y = ax + b$ "door" deze punten waarmee we de grootheid y wensen te voorspellen bij gegeven x.
We definiëren voor elk punt het ***residu*** e_i, met $\hat{y}_i = ax_i + b$, als volgt:

$$\begin{aligned} e_i &= \text{observatie - voorspelling} \\ &= y_i - \hat{y}_i \\ &= y_i - (ax_i + b) \end{aligned}$$

Merk op dat een residu positief is wanneer het punt boven de rechte gelegen is en negatief wanneer het onder de rechte gelegen is.

Om nu de "beste" rechte door de puntenwolk te zoeken, gebruiken we de ***kleinste kwadratenmethode***: hierbij moeten we a en b bepalen zodat $\sum_{i=1}^{n} e_i^2$ minimaal is.

Waarom zou de methode niet werken als we $\sum_{i=1}^{n} e_i$ minimaliseren?

B. Berekening van a en b.

Als we a en b zo kiezen dat $\sum_{i=1}^{n} e_i^2$ minimaal is, komen we tot de volgende formules:

Voor de regressierechte $y = ax + b$ door de punten $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ geldt:

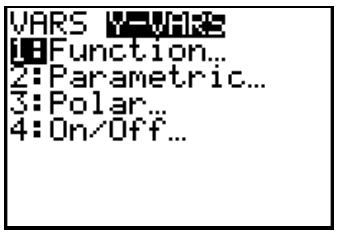
$$a = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{Cov(X,Y)}{VarX}$$

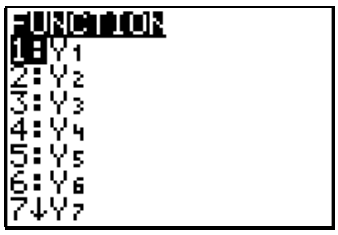
en $b = \bar{y} - a\bar{x}$

💻 Met de TI-83.

Met de TI-83 kunnen we a en b vlug berekenen. We plaatsen de X-gegevens in L1 en de Y-gegevens in L2.

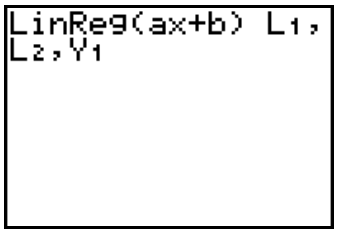
Druk dan STAT CALC 4:LinReg(ax+b).

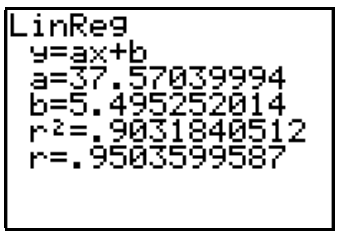




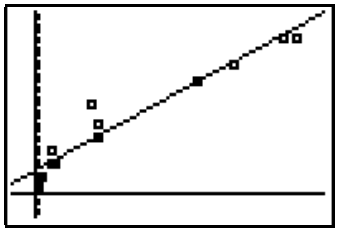
Typ L1,L2,VARS Y-VARS 1:Function 1:Y1.
Door het toevoegen van Y1 wordt de vergelijking van de regressierechte weggeschreven in Y1.

Op ons scherm zou nu dit moeten verschijnen:
Duwen we op ENTER dan verschijnen a en b op ons scherm. In het voorbeeld van de slakken zien we dit scherm:





Definiëren we nu het spreidingsdiagram zoals vroeger aangeleerd en duwen we op ZOOM 9:Zoomstat dan wordt de regressierechte afgebeeld in het spreidingsdiagram.



⌦ Oefening:

We hernemen de oefening van de punten van de leraar versus de punten van de moderator.

punten van de leraar X	15	24	34	43	47	56	60	70	83	90
punten van de moderator Y	27	36	37	48	47	52	65	66	70	80

(a) Als de regressierechte van Y op X de vergelijking $y = ax + b$ heeft, bereken dan a en b.

(b) Teken de regressierechte op het spreidingsdiagram. Gebruik hiervoor je rekentoestel. Controleer nu je resultaten die je in de vorige oefening had bekomen "op zicht".

Werk in vraag (c) en (d) eerst met de vergelijking en controleer je antwoord dan met behulp van het spreidingsdiagram.
Herinner je eraan dat je met de toets TRACE en dan met de pijltjestoetsen, de punten op je scherm krijgt met hun bijbehorende coördinaten.

(c) Gebruik de regressierechte om de punten gegeven door de moderator te voorspellen voor een leerling die van de leraar 20 punten kreeg.
Zat je voorspelling die je in de vorige oefening maakte, er dicht bij?

(d) Hoeveel zouden de punten van de leraar zijn als de punten van de moderator 56 waren?
Zat ook hier de voorspelling die je vroeger maakte, er dicht bij?

C. Enkele eigenschappen en bijzonderheden.

(1) Een tomatenkweker gebruikt in elk van de 12 moestuintjes een verschillende hoeveelheid kunstmest.
Dit gaf hem de volgende gegevens:

hoeveelheid kunstmest (g) X	10	12	14	16	18	20	22	24	26	28	30	32
tomatenoogst (kg) Y	2	2	2	3	4	3	4	3	5	6	7	9

(a) Bereken de regressierechte en teken ze in het spreidingsdiagram.

(b) Wat is het koppel $(\bar{x}, \bar{y})$?

(c) Bereken het residu van dit punt.

(d) Wat betekent deze waarde?

(e) Kunnen we dit veralgemenen? Maak je besluit hard.

(2) Beschouw de punten (0,3), (1,4), (2,7), (-1,4) en (-2,7).
(a) Teken deze punten in een spreidingsdiagram.
(b) Bereken de correlatiecoëfficiënt.

(c) Kun je hier besluiten dat er helemaal geen verband is tussen de twee variabelen?

(d) Bereken de regressierechte. Heeft deze hier zin? Kan ze bijvoorbeeld gebruikt worden om voorspellingen te doen?

We bedenken toch even dat met de geziene formules voor a en b, een regressierechte bepaald kan worden uitgaande van om het even welk spreidingsdiagram. Er kan dus – theoretisch gezien – een regressierechte beschouwd worden terwijl er helemaal geen oorzakelijk verband is tussen de twee variabelen. Dit is uiteraard niet zinvol.

Bij het opstellen van de vergelijking van de regressierechte eisen we dat $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ minimaal is.

Aangezien steeds geldt dat $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \geq 0$, is nul de kleinste waarde die $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ kan aannemen. Dit gebeurt als voor elke i geldt dat $y_i = \hat{y}_i$.
Met andere woorden alle punten van het spreidingsdiagram liggen op de regressierechte.
Men zegt dat er een ***perfecte lineaire correlatie*** is tussen X en Y.

Is daarenboven a>0 (of a<0), dan spreken we van een ***perfecte positieve (negatieve) lineaire correlatie***.

In het vorige hoofdstuk zagen we dat ook de correlatiecoëfficiënt ons aanwijzingen geeft over de "goedheid" van de regressierechte.

Eigenschap:

r = 1	asa	de correlatie is perfect positief lineair.
r = -1	asa	de correlatie is perfect negatief lineair.

En hoe dichter r bij –1 of 1 ligt, hoe beter het lineaire model past bij onze punten.

In het voorbeeld waar de punten (0,3), (1,4), (2,7), (-1,4) en (-2,7) perfect op een parabool lagen, was r = 0 en wisten we eigenlijk zo ook al dat het lineaire verband hier niet van toepassing was.

We mogen ons echter ook niet alleen baseren op de waarde van r, dit zal geïllustreerd worden in het volgende hoofdstuk.

D. Oefeningen.

(1) Men vermoedt zeer sterk dat de reactietijd van een persoon in verband staat met zijn hartslagritme. Elf dokters namen elk een verschillende hoeveelheid van een medicijn in, dat het hartslagritme beïnvloedt en testten zo hun vermoeden.
Dit leverde de volgende resultaten op:

hartslagritme (slagen per minuut) X	134	133	132	123	118	110	98	90	84	80	80
reactietijd (ms) Y	438	455	467	505	531	557	541	562	591	603	617

(a) Is er een sterke correlatie tussen X en Y?
(b) Toon de gegevens op een spreidingsdiagram.
(c) Zoek de vergelijking van de regressierechte met de kleinste kwadratenmethode voor regressie van Y op X.
(d) Teken deze rechte op je spreidingsdiagram.
(e) Voorspel de reactietijd van een dokter wiens hartslagritme 95 hartslagen per minuut bedraagt.
(f) Je wordt gevraagd om de reactietijd te voorspellen van een dokter wiens hartslagritme 60 slagen per minuut bedraagt. Geef commentaar bij deze vraag.

(2) In het begin van vorige eeuw onderzocht men in enkele streken van Beieren het verband tussen kindersterfte en flessenvoeding.
Dit gaf de cijfers uit de volgende tabel:

	kindersterfte (aantal sterftes per 1000 levend geborenen) X	aantal kinderen met flessenvoeding (in procent) Y
Niederbeieren	320	70
Oberfranken	170	10
Oberpfalz	300	40
Schwaben	270	60
Unterfranken	190	20
Mittelfranken	250	40

(a) Maak een spreidingsdiagram bij deze gegevens en bereken de regressierechte. Teken deze ook.
(b) In Oberbeieren kreeg 63% van de kinderen flessenvoeding en in Pfalz 15%. Doe aan de hand van de regressierechte een voorspelling voor de kindersterfte.
Ter vergelijking: de werkelijke cijfers waren respectievelijk 290 en 168.

(3) In de tabel hieronder vind je enkele Europese weerstations met hun hoogte boven de zeespiegel en gemiddelde jaartemperatuur.

station	hoogte (m)	temperatuur (°C)
Berlijn	49	9.1
Brocken	1152	2.4
Boedapest	130	10.9

Dobratsch	2140	0.1
Feuerkogel	1592	3.3
Graz	342	9.4
Innsbruck	579	8.4
Klagenfurt	448	8.1
Lugano	276	13.0
Praag	374	7.9
Salzburg	437	8.6
Säntis	2496	-2.3
Sonnblick	6106	-6.4
Wenen	203	9.1
Zugspitze	2962	-5.0

(a) Teken een spreidingsdiagram met daarop de regressierechte.
(b) Is het in het ski-oord Innsbruck relatief warm of relatief koud?
(c) Ukkel ligt op 100 meter boven de zeespiegel. Wat zou op basis van de regressierechte de gemiddelde jaartemperatuur in Ukkel moeten zijn?

(4) De tabel hieronder geeft telkens het gewicht van een boreling en de lengte van zijn moeder.

lengte moeder (cm) X	154	157	160	162	165	168	170	174	177	180	183
gewicht boreling (kg) Y	3.10	3.05	3.10	3.10	3.25	3.25	3.25	3.30	3.35	3.40	3.40

(a) Teken een spreidingsdiagram.
(b) Bereken de correlatiecoëfficiënt.
(c) Construeer de regressierechte.
(d) Denk je dat je hier mag spreken van een lineair verband tussen de lengte van de moeder en het gewicht van de boreling?

(5) Heeft het intelligentiequotiënt (X) een invloed op het schoolresultaat (Y)?
Wat denk je?
Controleer uw vermoeden bij de volgende gegevens van 10 lukraak gekozen zesdejaars, waarbij hun eindprocent in juni gegeven is.

X	92	100	95	118	110	103	105	110	125	122
Y	41	45	54	56	61	62	66	73	75	81

(6) De volgende tabel geeft de gemiddelde maandtemperatuur weer ten opzichte van het bedrag van de elektriciteitskosten die maand in een bepaald gezin.
Merk op dat er gegevens ontbreken voor 3 maanden.

maand	temperatuur (°C)	rekening (fr)	maand	temperatuur (°C)	rekening (fr)
april 91	10.5	1668	juni 92	19	1636
mei 91	16	1706	juli 92	22	1636

juni 91	23	1465	augustus 92	22	1656
juli 91	25	1628	september 92	21	1532
augustus 91	25.5	1540	oktober 92	*	*
september 91	23	1515	november 92	7	1753
oktober 91	15	1438	december 92	4	1776
november 91	9	1574	januari 93	1.5	1850
december 91	6.5	1986	februari 93	*	*
januari 92	1	2220	maart 93	-1	2032
februari 92	0	1912	april 93	9	1906
maart 92	5	1777	mei 93	*	*
april 92	6	1999	juni 93	20	1548
mei 92	14	1579	juli 93	25.5	1899

(a) Maak een spreidingsdiagram.
Kunnen we uit deze grafiek een positieve of een negatieve correlatie besluiten of is er helemaal geen correlatie tussen de temperatuur en de elektriciteitskosten?
En als er een correlatie is, is deze sterk? Wat gebruik je om je antwoord te staven?
(b) Bereken en teken de regressierechte. Is deze goed genoeg denk je, om voorspellingen mee te doen?
(c) Gebruik de vergelijking van deze rechte om het residu te berekenen voor maart 1992.
(d) Gebruik nu de grafiek om te zien welke maanden de grootste en welke de kleinste residu's hebben. Wat wilt dit in dit voorbeeld concreet zeggen?

(7) Onderzoekers aan het "Zee-visserij instituut" houden gegevens bij over het broedgedrag van de Afrikaanse pinguïns op de "Robben Islands". Ze bestuderen ook de ansjovis-bevolking, het geliefkoosde voedsel van de pinguïns.
Ansjovissen zijn te klein om ze stuk voor stuk te kunnen tellen, daarom schat men de hoeveelheid ton ansjovis in het water en dit noemt men de ansjovis-biomassa.
Onderzoekers verzamelden de volgende gegevens over de ansjovis-biomassa en het gemiddeld aantal jongen (die in staat zijn tot vliegen) per broedend paar pinguïns.

jaar	ansjovis-biomassa (milj. ton)	gem. aantal jongen in staat tot vliegen per broedend paar
1989	0.55	0.42
1990	0.47	0.32
1991	1.68	0.59
1992	1.50	0.59
1993	0.75	0.54
1994	0.48	0.46
1995	0.43	0.38

(a) Teken het spreidingsdiagram en beschrijf nauwkeurig het verband tussen de ansjovis-biomassa en het gemiddeld aantal vliegende jongen per broedend paar per jaar.
(b) Wat is de regressierechte? Teken deze op het spreidingsdiagram. Is deze goed benaderend?
(c) Gebruik nu de regressierechte om het gemiddeld aantal vliegende jongen per broedend paar te schatten, als de ansjovis-biomassa 1 miljoen ton is.

(d) Gebruik ook de regressierechte om de biomassa van de ansjovis te schatten, als het gemiddeld aantal vliegende jongen per broedend paar 0.5 bedraagt.
(e) Wat zou het gemiddeld aantal vliegende jongen per broedend paar bedragen als er geen ansjovis zou zijn? Klopt dit met de realiteit, denk je?
(f) Wat zou het gemiddeld aantal vliegende jongen per broedend paar zijn als er zo'n 20 miljoen ton ansjovis in zee zou zijn?
Is dit realistisch?

Afrikaanse pinguïns eten ook sardines. De volgende gegevens geven de sardine-biomassa weer in het water rond de "Robben Islands".

jaar	sardine-biomassa (milj. ton)
1989	0.36
1990	0.27
1991	0.45
1992	0.32
1993	0.47
1994	0.60
1995	0.58

(g) Is er een verband tussen de biomassa's van de beide vissoorten? Onderzoek dit.
(h) En tussen de sardine-biomassa en het gemiddeld aantal vliegende jongen per broedend pinguïn-paar?

(8) In dit hoofdstuk over de regressierechte komen heel veel aspecten samen: het nut van veel eerder aangeleerde begrippen wordt hier getoond.
Probeer dit hoofdstuk heel systematisch samen te vatten, zodat je duidelijk laat zien dat je de essentie begrijpt. Schrijf niet gewoon zinnen over en denk goed na over de volgorde van de begrippen in je samenvatting. Laat zien dat je voor elk begrip goed begrijpt waarom het is ingevoerd.

HOOFDSTUK 6:
MODELVORMING.

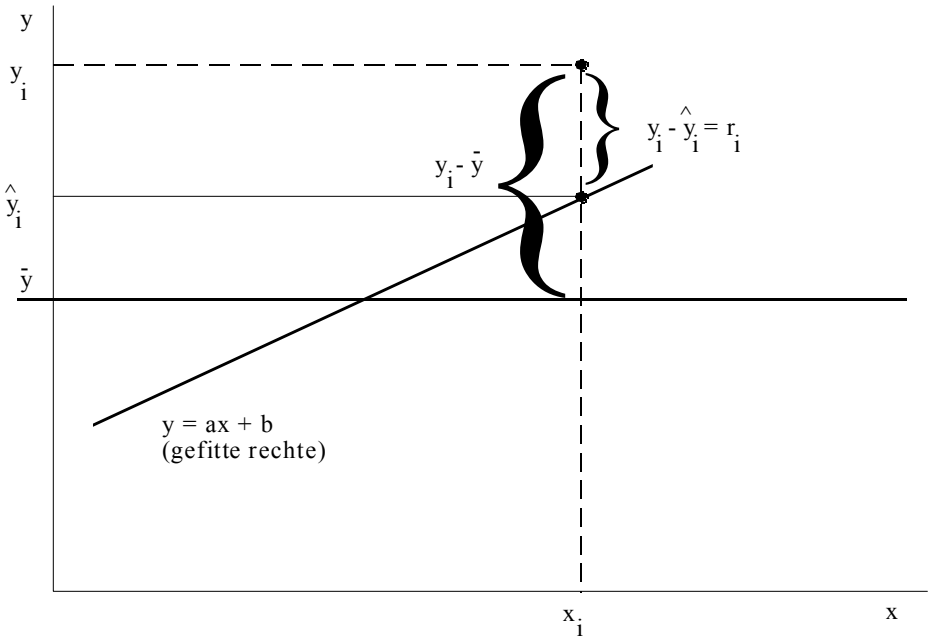
A. De determinatiecoëfficiënt.

De correlatiecoëfficiënt geeft ons eigenlijk alleen nuttige informatie als ons model lineair is. Zo kan er bijvoorbeeld een perfect kwadratisch verband bestaan, hoewel de correlatiecoëfficiënt dicht bij 0 ligt.
We zouden ook andere modellen, dan het lineaire model, moeten kunnen testen.
En hier komt de ***determinatiecoëfficiënt*** op de proppen.
De determinatiecoëfficiënt R^2 is een maat voor de kwaliteit van een regressiemodel dat niet noodzakelijk lineair is.

Als introductievoorbeeld nemen we de data (1,3), (2,1) en (3,5) met als regressierechte $y = x + 1$.
Teken het spreidingsdiagram en de regressierechte op je rekentoestel.
Beschouw nu de onderstaande tabel met $\hat{y}_i = x_i + 1$. Dit is de voorspelde waarde van y met behulp van het lineaire regressiemodel.

x_i	y_i	$(y_i - \bar{y})^2$	$(\hat{y}_i - \bar{y})^2$
1	3	0	1
2	1	4	0
3	5	4	1
	$\sum_{i=1}^{3}$	8	2

Als we enkel de data y_i zouden kennen, is $\bar{y}$ de beste voorspelling voor elke y_i.
Voor de ***totale variatie*** van de gegevens y_i ten opzichte van $\bar{y}$ nemen we dan ook $\sum_{i=1}^{n}(y_i - \bar{y})$ als maat: deze waarde geeft aan in welke mate de puntenwolk verticaal afwijkt van de horizontale rechte $y = \bar{y}$.
Voor de ***regressiemodel-variatie*** van de voorspelde waarden $\hat{y}_i$ ten opzichte van $\bar{y}$ nemen we $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})$ als maat: deze waarde geeft aan in welke mate de gegevens y_i door het regressiemodel verklaard worden. Dit wil zeggen in hoeverre het regressiemodel goed genoeg is voor zoveel mogelijk van de gegevens.



We definiëren nu de determinatiecoëfficiënt als volgt:

De determinatiecoëfficiënt R^2 wordt bepaald door de formule:

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

Ze poogt weer te geven hoe goed het regressiemodel de gegevens beschrijft.

In het vorige voorbeeld is $R^2 = \frac{2}{8} = \frac{1}{4}$.

Met andere woorden het regressiemodel $y = x + 1$ beschrijft de gegevens niet zo goed, want slechts 25% van de gegevens y_i worden verklaard door het regressiemodel.

Voor de correlatiecoëfficiënt r = 0.5 geldt dat $r^2 = R^2$.

Het kan algemeen bewezen worden dat voor een lineair regressiemodel geldt dat $r^2 = R^2$. Daarom noemt men r^2 ook de ***lineaire determinatiecoëfficiënt***.

Uit de definitie volgt dat $0 \le R^2 \le 1$ en dat, indien alle punten het regressiemodel perfect volgen, $R^2 = 1$ aangezien dan $\hat{y}_i = y_i$ voor elke i.

Bijvoorbeeld bij een lineair regressiemodel gebeurt dit als alle punten perfect op een rechte liggen.

Hoe dichter R^2 bij 1 ligt, hoe beter het regressiemodel de gegevens beschrijft.

Aan de hand van de determinatiecoëfficiënt kunnen we dan het beste regressiemodel proberen te vinden: dit kan lineair zijn, kwadratisch of van een hogere orde, exponentieel, logaritmisch,…

De verschillende ingebouwde regressiemodellen in de TI-83 vind je in STAT CALC. De belangrijkste zijn:

4:LinReg(ax+b)	$y = ax + b$
5:QuadReg	$y = ax^2 + bx + c$
6:CubicReg	$y = ax^3 + bx^2 + cx + d$
7:QuartReg	$y = ax^4 + bx^3 + cx^2 + dx + e$
9:LnReg	$y = a + b\ln x$
0:ExpReg	$y = ab^x$
A:PwrReg	$y = ax^b$
C:SinReg	$y = a\sin(bx + c) + d$

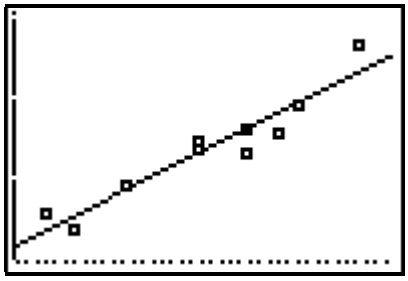
Bij elk regressiemodel kunnen we dan R^2 berekenen en zo het model er uit kiezen met de hoogste determinatiecoëfficiënt, maar dat toch niet te ingewikkeld is.
Bijvoorbeeld:
Stel dat een kwadratisch model een determinatiecoëfficiënt heeft van 0.961 en alleen een model van de vorm $y = ax^4 + bx^3 + cx^2 + dx + e$ doet beter met $R^2 = 0.967$. Dan is dit maar een kleine winst in R^2, maar we krijgen wel een veel ingewikkelder model.
Dus is het kwadratisch model te verkiezen, ondanks dat een ander model een iets betere determinatiecoëfficiënt heeft.

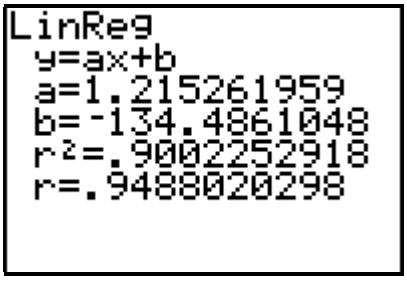
B. De residuplot.

Als voorbeeld beschouwen we de lengte X in cm en het gewicht Y in kg van 10 lukraak gekozen studenten:

lengte X	163	185	180	175	168	175	191	180	160	183
gewicht Y	60	90	78	81	71	79	104	84	64	83

We brengen deze gegevens op de TI-83 in de lijsten L1 en L2 en tekenen de puntenwolk. Zo te zien is een lineair model een goede benadering.



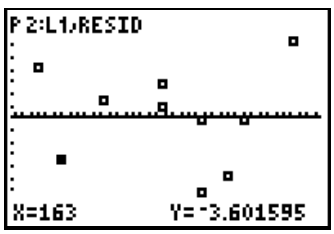


Bij lineaire regressie is de som van de residu's steeds nul:

$$\begin{aligned}\sum_{i=1}^{n} e_i &= \sum_{i=1}^{n} (y_i - \hat{y}_i) \\ &= \sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i - nb \\ &= n\bar{y} - an\bar{x} - nb \\ &= n\bar{y} - an\bar{x} - n(\bar{y} - a\bar{x}) \\ &= n\bar{y} - an\bar{x} - n\bar{y} + na\bar{x} \\ &= 0\end{aligned}$$

Voor een goed model moet de residuplot, dit is een grafiek van de residu's uitgezet tegen de gegevens x_i, een lukrake verdeling tonen ten opzichte van de x-as. Dus er mag geen bepaald patroon zichtbaar zijn in de residuplot: de punten moeten lukraak verdeeld zijn en schommelen rondom de x-as. Dus liggen er best ongeveer evenveel punten boven als onder de x-as: dwz dat er ongeveer evenveel residu's positief als negatief zijn en dus dat het regressiemodel de oorspronkelijke gegevens goed benadert.

Wanneer we dit gaan controleren bij ons voorbeeld, bekomen we:

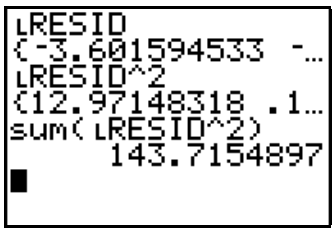


De determinatiecoëfficiënt R^2 is 0.9. Er wordt 90% van de variatie van de gegevens y_i ten opzichte van $\bar{y}$ verklaard door het lineaire regressiemodel.

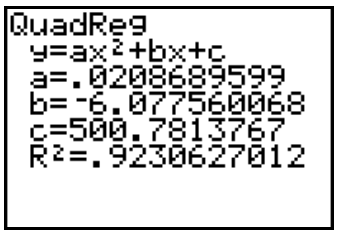
Reken na dat de som van de kwadraten der residu's $\sum_{i=1}^{n} e_i^2$, gelijk is aan 143.7.

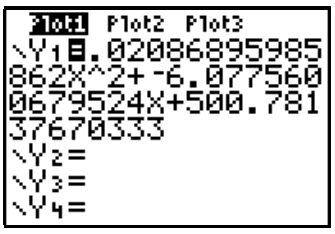
Tip voor de TI-83:

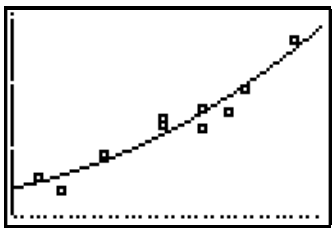
Met 2nd STAT (LIST) NAMES 1:RESID krijg je automatisch de lijst van de residu's.
Voor het berekenen van de som van de kwadraten der residu's, gebruik je 2nd STAT (LIST) MATH 5:sum(



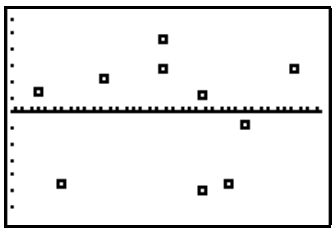
We onderzoeken of het kwadratisch model beter is.

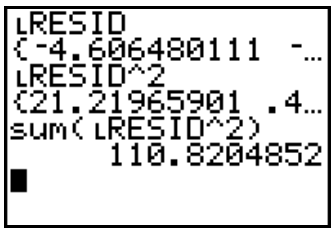






We zien dat er hier bij het kwadratisch model, een iets betere benadering is als we ons baseren op het spreidingsdiagram.
We maken nog een residuplot om te kijken of de residu's wel willekeurig verdeeld zijn.





Tot slot berekenen we de determinatiecoëfficiënt: R^2=0.923.
Uit dit alles kunnen we dus besluiten dat het kwadratisch model iets beter is dan het lineaire model.

Onderzoek van de andere modellen die de TI-83 berekent, geeft:

model	R^2	r^2	r
LinReg		0.900	0.949
QuadReg	0.923		
CubicReg	0.938		
QuartReg	0.963		
LnReg		0.892	0.945
ExpReg		0.815	0.957
PwrReg		0.911	0.955

In dit voorbeeld lijkt het model QuartReg het beste als we ons baseren op de determinatiecoëfficiënt. Het valt dan af te wegen of we dit model of het model QuadReg, dat veel eenvoudiger is, gaan gebruiken. Dit hangt o.a. af van de vereiste nauwkeurigheid voor ons probleem.

C. Enkele oefeningen.

(1) Onderstaande tabel geeft de snelheid en het verbruik van een wagen. Hoe verandert het verbruik Y van die wagen in functie van de snelheid X?

snelheid (km/h)	10	20	30	40	50	60	70	80
verbruik (l/100 km)	21	12	10	8	7	5.9	6.3	6.95

snelheid (km/h)	90	100	110	120	130	140	150
verbruik (l/100 km)	7.57	8.27	9.03	8.87	10.79	11.77	12.83

(a) Teken de puntenwolk in een spreidingsdiagram en bepaal de beste rechte door die punten.
(b) Zou jij deze rechte gebruiken om het verbruik bij een bepaalde snelheid te voorspellen?
Observeer hiervoor ook de determinatiecoëfficiënt en de residuplot.
Staaf je antwoord voldoende.
(c) Suggereer een beter model.

(2) Om je te oefenen in het zoeken van een geschikt model, is deze oefening heel nuttig.
Teken de volgende gegevens in een spreidingsdiagram en vind een geschikt model.

X	0	1	2	3	4	5	6	7	8	9	10
Y	0.2	3.6	7.5	11.5	15	17	20.4	22.7	25.9	27.6	30.2

Schrijf stap per stap op wat je probeert, wat je denkt en wat je besluit.

(3) De volgende tabel geeft de levens verwachting X (in jaren) versus de draagtijd van jongen Y (in dagen) voor een aantal dieren.
De gegevens zijn gepubliceerd in “The 1993 World Almanac and Book of Facts”.

dier	draagtijd	levens-verwachting	dier	draagtijd	levens-verwachting
Ezel	365	12	Guinees varken	68	4
Baviaan	187	20	Nijlpaard	238	25
Zwarte beer	219	18	Paard	330	20
Grizzlybeer	225	25	Kangoeroe	42	7
Ijsbeer	240	20	Luipaard	98	12
Bever	122	5	Leeuw	100	15
Buffel	278	15	Aap	164	15
Kameel	406	12	Eland	240	12
Kat	63	12	Muis	21	3
Chimpansee	231	20	Buidelrat	15	1
Aardeekhoorn	31	6	Varken	112	10
Koe	284	15	Poema	90	12
Hert	201	8	Konijn	31	5

Hond	61	12	Neushoorn	450	15
Olifant	645	40	Zeeleeuw	350	12
Wapiti	250	15	Schaap	154	12
Vos	52	7	Eekhoorn	44	10
Giraf	425	10	Tijger	105	16
Geit	151	8	Wolf	63	5
Gorilla	257	20	Zebra	365	15

(a) Teken het spreidingsdiagram, bereken de regressierechte en teken deze rechte op het spreidingsdiagram.
(b) Interpreteer de richtingscoëfficiënt van de regressierechte. Met andere woorden, leg nauwkeurig uit wat deze waarde zegt over de relatie tussen de levensverwachting en de draagtijd.
(c) Hoeveel % van de gegevens wordt verklaard door het model? Wat heb je hiervoor berekend?
(d) Maak een residuplot. Is er een verband tussen de residu's en de levensverwachting?
(e) Welk dier is duidelijk zowel in levensverwachting als in draagtijd een uitschieter? Bereken zijn residu-waarde. Lijkt dit dier het grootste residu (in absolute waarde) te hebben?

In de context van regressierechte, zijn uitschieters observaties met een groot (in absolute waarde) residu. Met andere woorden uitschieters liggen ver van de regressierechte en volgen het patroon niet.
De olifant hier is wel een uitschieter in levensverwachting en draagtijd maar is geen echte uitschieter in de regressiecontext.

(f) Welk dier heeft het grootste (in absolute waarde) residu? Is zijn draagtijd langer of korter dan verwacht voor een dier met zo een levensverwachting?
(g) Elimineer nu de gegevens van de giraf.
Wat is nu de regressierechte? En wat is nu r^2? Wat betekent dit?
(h) Teken nu de twee regressierechten op het spreidingsdiagram. Verschillen deze twee rechten erg?
(i) Voeg de gegevens van de giraf er terug bij, maar verwijder nu de gegevens van de olifant.
Wat is nu de regressierechte en r^2?
(j) Teken nu de laatste regressierechte samen met de regressierechte uit (a) op het spreidingsdiagram.
Is er nu meer verschil of minder als in (h)?

Bij kleinste-kwadratenregressie, is een ***invloedrijke observatie*** een observatie die een grote invloed heeft op de regressierechte.

(k) Is de olifant een invloedrijke observatie? En de giraf?
(l) Verander nu de draagtijd van de olifant van 645 dagen naar 45 dagen.
Bereken nu de regressierechte en r^2. Bespreek hun waarden.

D. Computerzitting.

Klik op het icoontje **Teach Me** op je scherm.
Klik op **Textbook** en daarna op **Bivariate data**.

Opdracht 1.

Klik op **Pearson's correlation**.

Lees nog eens de formule van de correlatiecoëfficiënt om op te frissen.
We weten ook dat de correlatiecoëfficiënt ligt tussen –1 en 1.

Klik op **interactive example**.

Het diagram toont een aantal punten die niet erg gecorreleerd zijn. Gebruik de scrollbar aan de linkerkant om de correlatiecoëfficiënt te veranderen.
Verander de punten zodanig totdat je denkt dat je een zeer sterke positieve correlatie (maar geen perfecte) hebt verkregen.
Hoe groot denk je dat r nu op dit moment bij jouw punten is?

……..

Klik dan op **next** en controleer de waarde van r links boven.

Hoeveel bedraagt de werkelijke correlatiecoëfficiënt van jouw punten?

……..

Klik op **reset** en probeer nu de punten zodanig te veranderen totdat je denkt dat r = 0.

Waar zorg je dan precies voor?

……..
……..

Klik opnieuw op **next** en controleer de waarde van r.

Klik op de **rode pijl** ⇦ om terug te keren naar het overzicht.

Opdracht 2.

Klik op **introduction onder de titel regression**.

Het spreidingsdiagram toont het aantal passerende auto's door een tunnel in de Alpen per uur, versus de benzine-concentratie in de tunnel-lucht.
In dit voorbeeld gebruikt men het lineaire model.
Lijkt dit jou een goed model als je je enkel baseert op het spreidingsdiagram? Of weet je een beter model? Leg uit.

……..
……..
……..
……..

Klik op **interactive example**.

We hebben gezien dat een uitschieter heel veel invloed kan hebben op de regressierechte. Dit zal in dit voorbeeld geïllustreerd worden.
Zie hier punten die bijna perfect op een rechte liggen. Het lineaire model lijkt voor deze data uiterst geschikt.

Klik nu op **Toggle**.

Je ziet hoe een uitschieter de ligging van de regressierechte enorm kan doen wijzigen.
Als de uitschieter in dit geval, 5 cm meer naar links zou liggen, hoe zou de regressierechte er dan uitzien? Teken deze rechte en breng ook de rode rechte vanop je scherm op je tekening om te kunnen vergelijken.

Klik nu op **exit**.

Klik op **leverage effect**.

Hier zie je hetzelfde spreidingsdiagram als daarjuist.

Klik in de figuur.

Hier zie je een aantal data-punten die zeer dicht tegen een rechte liggen. Begin met 10 punten.

Klik op **next** om de regressierechte te zien.
Klik nu op een willekeurig punt en leg dat punt (door te slepen met de muis) op de plaats met coördinaten (6,2).

Je hebt nu dezelfde situatie gecreëerd als in de vorige opgave waar je moest proberen de regressierechte te tekenen. Controleer nu je antwoord van de vorige oefening: sleep hiervoor je punt naar de plaats met coördinaten (4,2).
Hoe zal de regressierechte veranderen?

……...

Waardoor komt dit, denk je?

……...
……...
……...

Klik op **next**.

Hoe verder de uitschieter ligt van de andere punten, hoe meer de regressierechte zal veranderen.

Klik op **next** om een grotere omgeving te creëren.

Om het gevaar van een uitschieter nog eens extra te beklemtonen, probeer je het punt zodanig te verslepen totdat de regressierechte bijna loodrecht op de lijn met de 9 andere punten staat.
Welke coördinaten heeft je punt nu?

……

Je ziet dus dat een uitschieter enorm veel kwaad kan verrichten, als je hem niet opmerkt tijdens je onderzoek.

Klik nu op **reset**.

Verander nu het aantal punten van 10 naar 100 en neem er opnieuw één punt uit, ongeveer op dezelfde plaats als je punt van daarjuist. Sleep dit nu verder weg, maak je omgeving weer groter en zet het opnieuw op de plaats die je daarnet gekozen had.
Staat nu de regressierechte terug bijna loodrecht op de lijn met de 99 andere punten?

……

Bestaat er zo een plaats waar je de uitschieter moet leggen en waar dit wel zo is? Zo ja, welke coördinaten heeft die plaats?

……

Wat kan je hierover besluiten over het verband tussen het aantal data-punten en uitschieters?

✎……...
……...
……...

Eventueel kun je alles nog eens herhalen met 50 punten als je nog niet zeker bent over je antwoord of om je besluit te staven.

🖱 Klik op **exit**.

🖱 Klik op de **rode pijl** ⇦ om terug te keren naar het overzicht.

Opdracht 3.

Klik op **straight line onder de titel regression**.

De tekening toont nog eens wat x_i, y_i en $\hat{y}_i$ juist is. Hier gebruiken ze wel $\hat{y}_i = a + bx_i$, de rol van a en b is gewoon omgekeerd. Het verschil tussen $\hat{y}_i$ en y_i noemen we het residu.

Klik op **interactive example**.

Gegeven zijn een aantal data-punten en een rechte. De bedoeling is nu dat jij, op zicht, de best benaderende rechte tekent: de regressierechte.

Klik op **next**.

De rechte heeft twee parameters: k controleert de richtingscoëfficiënt van de rechte en d geeft de waarde op de y-as waar de rechte de y-as snijdt.
Op de rechtse grafiek kan je, door het puntje te slepen, de ligging van de rechte wijzigen. Door horizontaal te bewegen, verander je de richtingscoëfficiënt. Door verticaal te bewegen verander je d. In de linkse grafiek zie je de rechte dan bewegen.

Klik op **next**.

Nu zie je de residu's getekend in rode lijnen. Je weet dat je de regressierechte bekomt door de som van de kwadraten van de residu's te minimaliseren. Verander nu de rechte zodanig tot je denkt dat je de best benaderende rechte hebt.
Onderaan staat de som van de gekwadrateerde residu's.
Wat is die som bij jou?

……

Klik op **next**.

In de linkse figuur zie je nu wanneer je rechte goed was.
Welke kleur heeft het gebied waar jouw gekozen punt (het grootste punt) in ligt?

……

Ligt het grootste puntje in het knalgroene gebied, dan heb je goed geschat. Het turkoois gebied heeft een al iets groter som van gekwadrateerde residu's en als je punt in het blauwe gebied ligt dan heb je slecht geschat.
Het kleine puntje geeft de regressierecht weer (=best benaderende rechte, het optimale geval). Viel jouw punt hiermee samen, dan had je dus perfect geschat.
Leg nu eens jouw punt op het kleinste puntje.
Wat is de som van de gekwadrateerde residu's voor de regressierechte?

……

Klik op **exit**.
Klik op de **rode pijl** ⇦ om terug te keren naar het overzicht.

Opdracht 4.

Klik op **analysis of residuals onder de titel regression**.

Het analyseren van de residu's is zeker belangrijk voor de verdere studie van ons model. Dit kunnen we doen aan de hand van de residuplot. We weten dat voor een goed model de residuplot aan bepaalde voorwaarden moet voldoen:

- De punten in de residuplot moeten willekeurig zijn en mogen geen bepaald patroon vertonen.
- Er moeten ongeveer evenveel punten liggen boven y = 0 als er onder.

Klik op **interactive example**.

In deze oefening krijg je 4 verschillende gegevens-sets te zien, waarop men lineaire regressie wil toepassen. Slechts één van deze is geschikt voor lineaire regressie. Je krijgt telkens het spreidingsdiagram en de residuplot te zien en op grond hiervan moet jij kunnen zeggen of hier lineaire regressie kan toegepast worden.

Klik op **next**.

Het eerste geval is het ideale en het enige waarbij lineaire regressie mag toegepast worden: de relatie tussen X en Y is lineair, de punten in de residuplot zijn willekeurig verspreid en er is geen bepaald patroon.

Klik op **next**.

Aan welke factoren kan je hier zien dat een lineair model niet goed is?

……...
……...
……...

Beantwoord deze vraag ook voor de volgende gevallen.

……...
……...
……...

……...
……...
……...

……...
……...
……...

Klik op **exit**