Lineaire regressie

Inhoud

1. Inleiding

Less proofs, more models

De titel boven dit stukje is een slogan van we-weten-niet-meer-wie op een internationaal congres over wiskundedidactiek (dat we evenmin precies in de tijd weten te situeren). Wel is het duidelijk dat we in vergelijking met een 20 jaar geleden inderdaad 'less proofs, more models' hebben in ons wiskundeonderwijs. De eerste helft van de slogan interesseert ons nu even niet, de tweede helft des te meer: deze loep gaat over de 'more models', m.a.w. hoe kunnen we op een verantwoorde manier een functie associëren aan meetgegevens en op die manier een fenomeen uit de realiteit min of meer adequaat beschrijven aan de hand van een wiskundig model? In veel gevallen maken leerlingen hiermee al vroeg kennis in de fysicales, via de wonderknop 'Add trendline ...' uit Excel of via de grafische rekenmachine. Onder de noemer 'regressie' kunnen we dit onderwerp uitdiepen in de wiskundeles.

Een loep voor een heterogeen publiek

In de eindtermen vinden we het onderwerp 'lineaire regressie' niet terug, wel in de leerplannen. In heel veel studierichtingen, zowel in ASO als in TSO, zowel in het vrij onderwijs als in het gemeenschapsonderwijs, is het een keuzeonderwerp. Er wordt geadviseerd om er ongeveer 15 lessen voor uit te trekken. In de OVSG-leerplannen is het minder populair en wordt er enkel naar verwezen als mogelijke uitbreiding in het leerplan voor de studierichtingen in het ASO met pool wiskunde.

Je ziet dus dat het onderwerp voor een heel divers publiek geprogrammeerd staat (of beter: kan staan). In deze loep hebben we aandacht voor de basiszaken (tot en met paragraaf 4), maar graven we ook dieper (al een klein beetje in de paragrafen 5 en 6, maar zeker in de paragrafen 7 en 8). Aan jou om uit te maken welke delen je met welke leerlingen kunt en wilt doen.

Regressie voor leerlingen en leerkrachten

We hebben het gevoel dat niet alle leerkrachten die les zullen (moeten) geven over lineaire regressie, zich vanuit hun opleiding voldoende thuis voelen in deze materie. Daarom vind je in deze loep niet enkel een didactische verwerking van het onderwerp. We zorgen ook voor meer achtergrond voor leerkrachten. Het kan dus gerust zijn dat je na het lezen van deze loep vindt dat niet alles spek voor de leerlingenbek is. Dat zal zeker zo zijn als je leerlingen niet zo sterk zijn in wiskunde (en statistiek). We hebben wel geprobeerd om alles zo uit te werken dat het begrepen kan worden door de sterkere leerlingen (en, zoals gezegd, het grootste deel ook door minder sterke leerlingen).

Nadruk op de statistische kant van de zaak

Statistiek is in het secundair onderwijs een onderdeel van het vak wiskunde. In het echte leven is de band tussen beide dikwijls minder sterk en is de verhouding eerder zoals die tussen fysica en wiskunde. Statistiek doet vaak een beroep op wiskunde, maar naast het wiskundige denken zijn er in de statistiek nog heel wat andere dingen belangrijk. We proberen in deze loep het statistische denken te beklemtonen en statistiek minder te zien als louter een onderdeel van de wiskunde. In dat kader kun je begrijpen dat je in deze loep niet de afleiding vindt van de formule voor de coëfficiënten in de regressielijn, dat we niet ingaan op de berekening van de regressielijn en correlatiecoëfficiënt bij een gezamenlijke frequentietabel, ...

Computers en grafische rekenmachines

Lineaire regressie is veel toegankelijker geworden nu grafische rekenmachines en computers vlot beschikbaar zijn. Er komt immers heel wat rekenwerk aan te pas. Dat rekenwerk kan nu in een oogwenk gebeuren door de elektronische rekenhulp. Grafische voorstellingen zijn in de statistiek steeds zeer nuttig en ook op dat terrein bieden computers en grafische rekenmachines de nodige ondersteuning. In deze loep werken we in sommige paragrafen met de grafische rekenmachine TI84. In andere paragrafen, wanneer de sets gegevens groter zijn of wanneer de eisen aan de figuren groter zijn, gebruiken we een rekenblad (zoals Excel).

2. Samenhang

De studie van lineaire regressie start best met enkele inleidende voorbeelden waarin gewerkt wordt met gekoppelde gegevens en grafische voorstellingen (spreidingsdiagrammen) daarvan. De voorbeelden worden gericht op de vraag ‘is er samenhang?’. Dit wordt visueel beoordeeld. In deze voorbeelden moet ook zeer expliciet aandacht gaan naar het feit dat samenhang niet betekent dat er een causaal verband is.

We geven hieronder enkele inleidende oefeningen. In het eerste voorbeeld onderzoeken we of er samenhang is tussen de lengte en de schoenmaat van (mannelijke) personen. Er blijkt, zoals verwacht, een vrij grote samenhang te zijn en de leerlingen kunnen daarbij zelfs een verband formuleren. We werken in de werktekst met gegevens van een militaire keuring. De leerlingen kunnen natuurlijk ook, en misschien zelfs beter, zelf gegevens verzamelen.

Lengte en schoenmaat

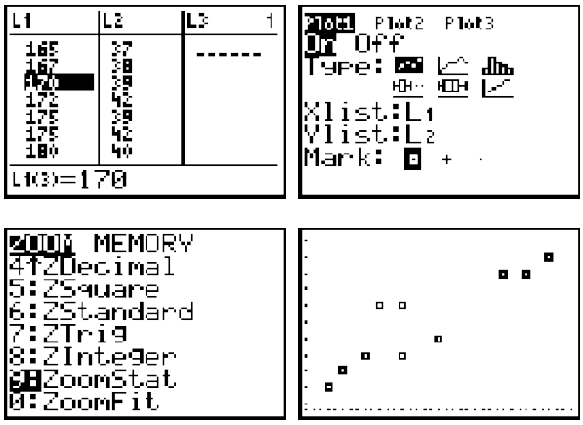
Om te onderzoeken of er een verband bestaat tussen de lengte en de schoenmaat van personen heb je gegevens nodig.

1. Denk je dat er een verband is tussen de lengte en de schoenmaat?

Hieronder vind je de lengte en de schoenmaat van 10 jongens. De gegevens komen van een (oude) militaire keuring.

lengte	165	167	170	172	175	175	180	189	192	195
schoenmaat	37	38	39	42	39	42	40	44	44	45

2. Voer deze gegevens in in je grafisch rekentoestel (bijvoorbeeld in L1 en L2) en laat je rekentoestel een spreidingsdiagram tekenen. (Vergeet niet de gewone functies uit te zetten.)



3. Wat is de richting van de puntenwolk?

   (Als je het eerste punt met het laatste verbindt, vind je als richting $\frac{8}{30}$*.)*

4. Hoe zou je de richting van de puntenwolk in woorden beschrijven? Wat betekent dit dus voor het verband tussen de lengte en de schoenmaat van de kandidaat-soldaten?

   (Als de lengte met 3,75 cm toeneemt, neemt de schoenmaat met 1 toe.)

Hierop laten we best een voorbeeld volgen waarin geen samenhang te ontdekken is.

Schedelomtrek en punten voor wiskunde

In de vorige werktekst vonden we dat in de meeste gevallen een grote lengte overeenkomt met een grote schoenmaat. In deze werktekst willen we onderzoeken of er eventueel een verband zou kunnen zijn tussen intelligentie en de schedelomtrek. Hieronder vind je van een klas (6^{de} jaar, 6 uur wiskunde) zowel de punten van wiskunde van het kerstrapport (2004) als de schedelomtrek van de betrokken leerlingen.

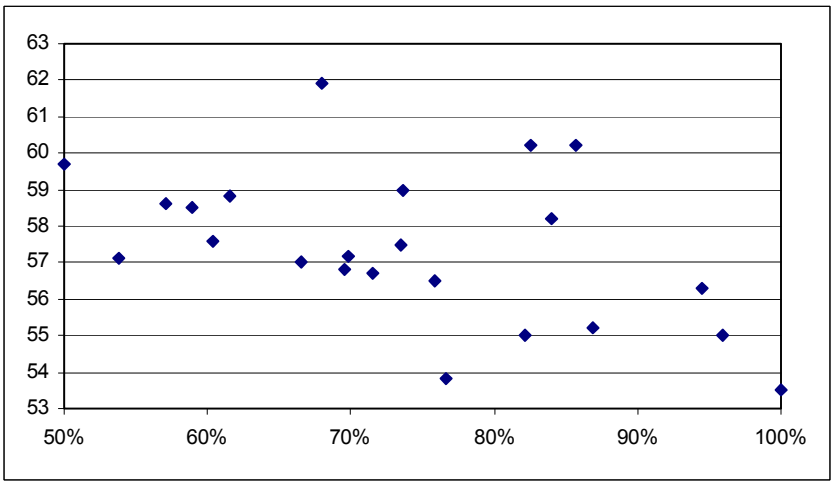
punten (op 100)	schedelomtrek (in cm)	punten (op 100)	schedelomtrek (in cm)
62	58,8	70	56,8
74	57,5	83	60,2
82	55	74	59
86	60,2	50	59,7
54	57,1	70	57,2
77	53,8	60	57,6
95	56,3	100	53,5
67	57	49	58,8
76	56,5	72	56,7
68	61,9	57	58,6
87	55,2	59	58,5
96	55	70	56,8
84	58,2	83	60,2

1. Kan je op basis van deze cijfers iets zeggen over het verband tussen intelligentie en de schedelomtrek?

   (Nee, op basis van deze gegevens kunnen we enkel iets zeggen over een mogelijk verband tussen de punten voor wiskunde en de schedelomtrek. We weten niet of en welk verband er bestaat tussen deze punten en de intelligentie van deze leerlingen.)

We zullen dus enkel nagaan of er misschien een verband is tussen de punten voor wiskunde en de schedelomtrek.

2. Verwacht je een verband?
3. Hieronder zie je een puntenwolk van de gegevens. Kan je op basis hiervan vermoeden dat er een verband is?



   (Nee, op het eerste gezicht zou je kunnen zeggen dat bij hogere punten voor wiskunde eerder een iets kleinere schedelomtrek hoort. Maar er zijn teveel punten die dit tegenspreken.)

In de volgende paragrafen zullen we iets uitvoeriger onderzoeken wanneer we wel en wanneer we niet van samenhang mogen spreken.

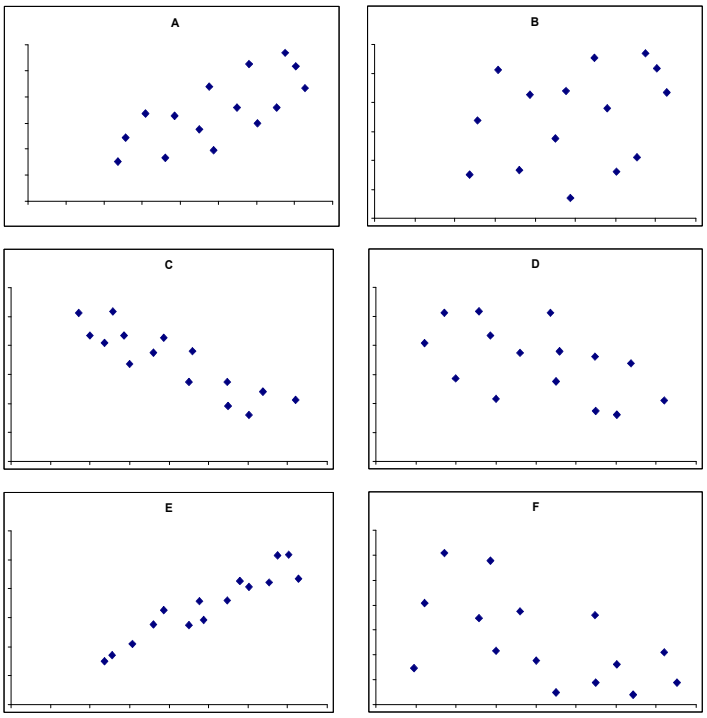
Puntenwolken

Een belangrijk kenmerk bij puntenwolken is de globale vorm. Op basis hiervan kan je eerste besluiten trekken over een mogelijke samenhang. Kijk nog eens terug naar het voorbeeld over de lengte en de schoenmaat. Bij een grotere lengte hoort een grotere schoenmaat. De richting van de puntenwolk is stijgend. Men noemt de samenhang *positief*. Als de richting van de puntenwolk eerder dalend is, dan spreekt men van een *negatieve* samenhang.

Uit de globale vorm kan je soms ook al iets vermoeden over de sterkte van de samenhang. Verderop zal je zien dat je daar voorzichtig mee moet omgaan. Toch is het zinvol om in de inleidende voorbeelden al in te gaan op de sterkte van de samenhang. Later koppelen we daar het begrip correlatiecoëfficiënt aan. In de volgende oefening vergelijken we verschillende diagrammen van gekoppelde gegevens.

Positief/negatief, sterk/zwak

Hieronder zie je zes spreidingsdiagrammen van fictieve scores op twee examens. Alle grafieken hebben dezelfde eenheid op de horizontale as en ook op de verticale as.



Bepaal voor elk van bovenstaande diagrammen de richting (positief/negatief) en de mate van samenhang. Vul daarvoor in elk vakje van onderstaande tabel de bijbehorende letter van het diagram in. (Juist één diagram per vakje.)

	sterk	matig	zwak
positief			
negatief			

Niet-lineaire samenhang

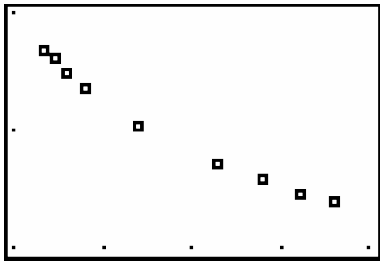
Natuurlijk betekent samenhang tussen twee variabelen niet noodzakelijk dat er een lineair verband is tussen die twee variabelen. In de volgende paragrafen onderzoeken we wanneer het aanvaardbaar is een lineaire benadering te gebruiken voor de relatie tussen twee variabelen. Maar het is belangrijk dat leerlingen niet de indruk krijgen dat alle verbanden lineair benaderd kunnen worden.

Fysica: de gaswet van Boyle-Mariotte

In de lessen fysica wordt het verband tussen druk, volume en temperatuur bij gassen onderzocht. Hieronder vind je de resultaten van een leerlingenproef. De leerlingen moesten bij een constante temperatuur het volume van een gas wijzigen en dan de bijbehorende druk meten.

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

1. Voer deze gegevens in in je rekentoestel (of computer) en teken een spreidingsdiagram.



Er is blijkbaar een verband tussen het volume en de bijbehorende druk.

2. Welke eenvoudige samenhang kan je onmiddellijk waarnemen?
   (Als het volume toeneemt, neemt de druk af.)
3. Verbind het eerste en het laatste punt van de puntenwolk. Welke richting heeft deze rechte?
   (dalend met helling $-40{,}12$*)*
4. Vind je dat deze rechte de samenhang goed weergeeft?
   (Nee, alle andere punten liggen onder deze rechte.)

5. In de fysica leert men dat bij constante temperatuur de druk omgekeerd evenredig is met het volume. Welk soort grafiek past bij zulk verband?

   (Een hyperbool)

Samenhang tussen twee variabelen betekent dus niet dat de samenhang kan samengevat worden door een rechte. Heel veel verschillende verbanden zijn mogelijk. In de volgende paragrafen wordt onderzocht wanneer het verband wel samengevat kan worden in een rechte en hoe dit kan gebeuren.

Samenhang ≠ oorzakelijk verband

We moeten de leerlingen er ook op wijzen dat een verband tussen twee variabelen niet hetzelfde is als een 'oorzakelijk verband'. Een bekend voorbeeld hiervan is een vergelijking van het aantal geboortes in een bepaald gebied en het aantal waargenomen ooievaars.

Het volgende voorbeeld is gebaseerd op gegevens over de levensverwachting in verschillende landen van de wereld en het aantal personen per TV-toestel en per arts in dezelfde landen. We beschrijven hoe de leerlingen op basis van die gegevens kunnen redeneren over samenhang en oorzakelijk verband. De aanpak is gebaseerd op [6].

Levensverwachting, TV-toestellen en artsen

In een eerste fase laten we de leerlingen (in groepjes) nadenken over de volgende opdrachten:

1. In België is de levensverwachting op dit moment ongeveer 79 jaar (iets minder voor mannen en iets meer voor vrouwen). Doe eens een zinvolle gok naar de levensverwachting in de volgende landen: Canada, Frankrijk, India, Kenia, Roemenië, Taiwan, VS.
2. Probeer nu ook eens te gokken naar het aantal personen per TV-toestel in die landen.

Omdat dit een wat ongewone variabele is, verplicht dit de leerlingen om even na te denken en te reflecteren. Vervolgens kunnen ze over de volgende vragen nadenken. Uiteraard is het nadenken en bespreken van de vragen belangrijker dan de antwoorden. Toch is het goed de leerlingen te verplichten antwoorden op te schrijven. Ze kunnen verderop ook nog gebruikt worden.

3. Denk je dat er samenhang is tussen de levensverwachting in een land en het aantal personen per TV-toestel.
4. Als er samenhang is, is die dan positief of negatief?

Deze vragen kunnen boeiende discussies opleveren. Ze helpen de leerlingen om bij statistische redeneringen niet enkel naar de data te kijken als getallen maar als getallen in een specifieke context.

In een volgende fase krijgen de leerlingen de volgende gegevens (uit [9]). Je vindt de gegevens ook op onze website.

Land	Levens-verwachting	Personen per TV-toestel	Personen per arts	Vrouwelijke levens-verwachting	Mannelijke levens-verwachting
Argentinië	70.5	4.0	370	74	67
Bangladesh	53.5	315.0	6166	53	54
Brazilië	65.0	4.0	684	68	62
Canada	76.5	1.7	449	80	73
China	70.0	8.0	643	72	68
Colombia	71.0	5.6	1551	74	68
Egypte	60.5	15.0	616	61	60
Ethiopië	51.5	503.0	36660	53	50
Frankrijk	78.0	2.6	403	82	74
Duitsland	76.0	2.6	346	79	73
India	57.5	44.0	2471	58	57
Indonesië	61.0	24.0	7427	63	59
Iran	64.5	23.0	2992	65	64
Italië	78.5	3.8	233	82	75
Japan	79.0	1.8	609	82	76
Kenia	61.0	96.0	7615	63	59

Noord-Korea	70.0	90.0	370	73	67
Zuid-Korea	70.0	4.9	1066	73	67
Mexico	72.0	6.6	600	76	68
Marokko	64.5	21.0	4873	66	63
Myanmar (Birma)	54.5	592.0	3485	56	53
Pakistan	56.5	73.0	2364	57	56
Peru	64.5	14.0	1016	67	62
Filippijnen	64.5	8.8	1062	67	62
Polen	73.0	3.9	480	77	69
Roemenië	72.0	6.0	559	75	69
Rusland	69.0	3.2	259	74	64
Zuid- Afrika	64.0	11.0	1340	67	61
Spanje	78.5	2.6	275	82	75
Soedan	53.0	23.0	12550	54	52
Taiwan	75.0	3.2	965	78	72
Thailand	68.5	11.0	4883	71	66
Turkije	70.0	5.0	1189	72	68
Oekraine	70.5	3.0	226	75	66
Groot-Brittannië	76.0	3.0	611	79	73
Verenigde Staten	75.5	1.3	404	79	72
Venezuela	74.5	5.6	576	78	71
Vietnam	65.0	29.0	3096	67	63

5. Zoek het land met de hoogste levensverwachting en het land met de laagste levensverwachting.
   (Japan, respectievelijk Ethiöpië)
6. En het land met het minste en meeste aantal inwoners per TV-toestel?
   (VS, respectievelijk Myanmar)

Het is nuttig om de verdeling van de aparte gegevens even te bespreken vooraleer dieper in te gaan op de samenhang tussen de gegevens.

Om de samenhang te onderzoeken, maken de leerlingen een spreidingsdiagram van de levensverwachting (verticale as) tegenover het aantal personen per TV-toestel (horizontale as). Nadat ze de negatieve samenhang tussen de variabelen ontdekt hebben, gaan we dieper in op het aspect ‘oorzaak’ aan de hand van de volgende vraag:

7. Als we massa’s TV-toestellen laten aanvoeren naar de landen met een lage levensverwachting, zullen de mensen daar dan langer leven?

De leerlingen zullen dit natuurlijk een belachelijke vraag vinden, maar ze opent wel de ogen. Samenhang wijst in dit geval helemaal niet op een oorzakelijk verband. In een leergesprek kan dieper ingegaan worden op meer plausibele verklaringen voor de samenhang.

Eventueel kan ook de samenhang met de andere gegevens uit de tabel onderzocht worden, bijvoorbeeld het verband tussen het aantal personen per arts en de levensverwachting. Voor die samenhang zijn ook verschillende interpretaties:

- De lage levensverwachting is een gevolg van het gebrek aan artsen.
- De lage levensverwachting en het lage aantal artsen zijn beide een gevolg van een derde factor, bijvoorbeeld de armoede, het gebrek aan scholing, …
- Een laag aantal artsen is een gevolg van een lage levensverwachting.

Op basis van deze gegevens is niet te bepalen welke de juiste is.

3. Lineaire samenhang en regressierechte

In deze paragraaf richten we de blik meer concreet op *lineaire* samenhang. Hier sluit dan de regressielijn bij aan. De regressierechte wordt in deze paragraaf gezien als (de meetkundige evenknie van) een 'vuistregel': een eenvoudige formule die de samenhang tussen de gegevens samenvat. We kunnen deze vuistregel ook gebruiken om 'voorspellingen' te doen in verband met waarden die niet in onze set gegevens zitten.

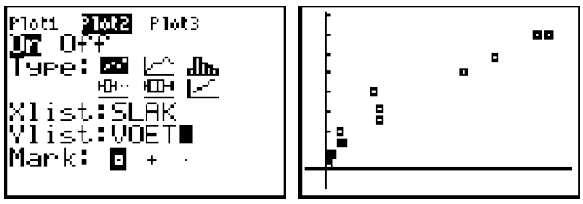
We werken één voorbeeld uit (in de les moet je er natuurlijk verschillende uitwerken ...). De leerlingen krijgen een dataset (uit [7]) waarin een sterke lineaire samenhang te zien is. Ze moeten ook een rechte zoeken die bij de gegevens past, eerst op het gevoel af en daarna met behulp van de rekenmachine (of de computer). We gaan voorlopig nog niet in op de manier waarop deze rechte door de rekenmachine berekend wordt. Dat stellen we uit tot paragraaf 6.

Slakken

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

gewicht (g)	voetopp. (mm^2)	gewicht (g)	voetopp. (mm^2)
0,64	29	0,07	7
0,21	16	0,01	3
0,85	35	0,05	10
0,53	25	0,81	35
0,02	4	0,53	25
0,01	1	0,18	20
0,21	16	0,06	7
0,18	20	0,20	13
0,06	7	0,07	7
0,20	13	0,01	1

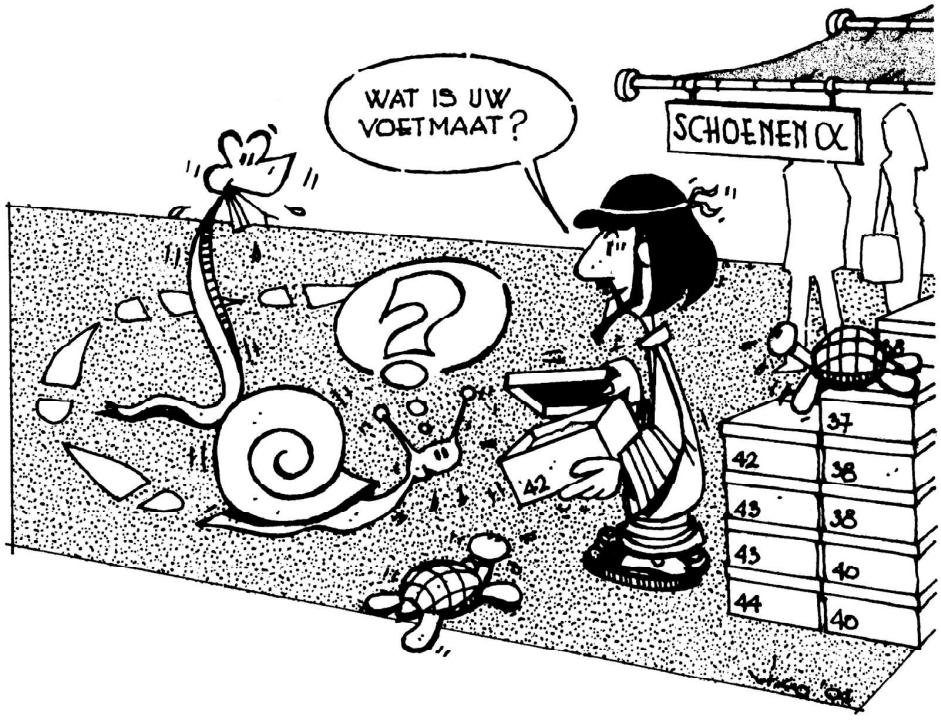
1. Teken met je rekentoestel een spreidingsdiagram van deze gegevens.



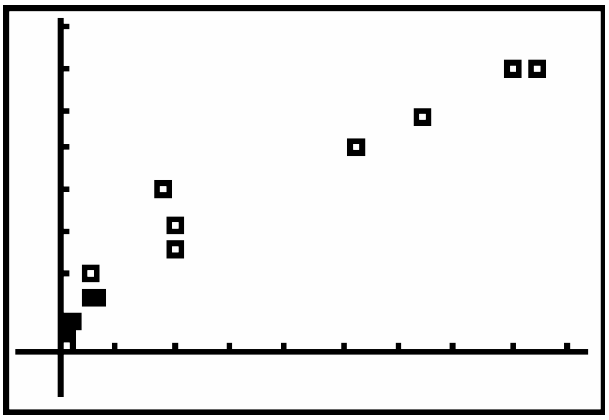
2. De punten op het spreidingsdiagram liggen duidelijk zeer dicht bij een rechte. Welk soort verband is er?

   (Een positief lineair verband)

Als we vermoeden dat er een lineair verband is tussen twee variabelen, kan het interessant zijn de vergelijking te vinden van de rechte (of, wat op hetzelfde neerkomt, de eerstegraadsfunctie) die dit verband beschrijft. De onderzoeker van deze slakken kan deze rechte (beter: haar vergelijking) dan bijvoorbeeld gebruiken als een soort vuistregel. Uit het gewicht van een slak kan hij dan met een paar eenvoudige bewerkingen zeer snel de 'normale' voetoppervlakte bepalen.



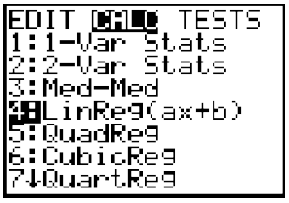
3. Schets op het spreidingsdiagram de rechte die volgens jou het verband het beste weergeeft. Geef ook de vergelijking van de rechte die je getekend hebt. (Op de horizontale as is de eenheid 0,1 gram, op de verticale as 5 mm^2.)



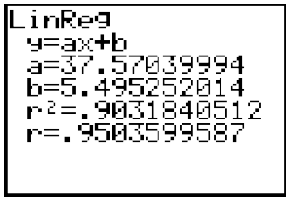
4. Een veldonderzoeker vindt een slak die 0,4 gram weegt. Wat is een 'normale' voetoppervlakte voor zo'n slak?
5. Vergelijk je rechte met die van andere leerlingen. Heeft iedereen dezelfde rechte? En hoe zit het met de voorspelde voetoppervlaktes?

Je rekenmachine kan ook zo'n rechte bepalen (we maken ons voorlopig geen zorgen over de manier waarop de rekenmachine de rechte bepaalt). Ze wordt de 'regressierechte' genoemd. Het gaat als volgt:

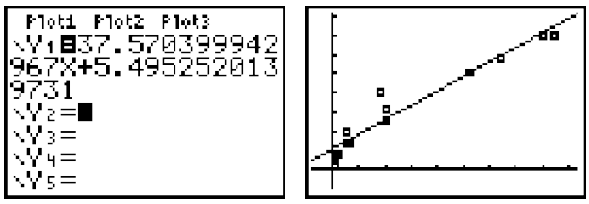
- Bij STAT CALC kies je 4:LinReg (ax+b)



- Vul de opdracht aan met de lijsten waarvan je de regressierechte wil bepalen (L1, L2, … kan je via de toetsen invoeren, de naam van andere lijsten haal je op bij [2nd] LIST NAMES). We geven ook de naam van een functie op zodat de rekenmachine de gevonden vergelijking bewaart. Dat laatste is niet verplicht. In de schermafdruk rechts hieronder zie je de vergelijking van de regressielijn. Het is mogelijk dat de waarden van r^2 en r niet getoond worden. We komen later terug op de betekenis van deze getallen en de manier om ze door de rekenmachine te laten tonen.



- Het voorschrift van de regressierechte zit nu ook in Y1. Als je niets hebt veranderd aan de instellingen wordt met GRAPH de regressierechte in het spreidingsdiagram getekend.



6. Gebruik nu de regressierechte van je rekentoestel om vraag 4 te beantwoorden. Zat je eigen voorspelling dicht bij de resultaten van je rekentoestel?

We kunnen bij elk gewicht van de gegeven data ook de voetoppervlakte voorspellen m.b.v. de regressievergelijking. In onderstaande tabel is dit gebeurd. Dit gaat eenvoudig door op je rekentoestel de lijst Y1(SLAK) te maken en die bv. in L3 op te slaan.

7. Bereken nu bij elk gewicht het verschil tussen de gegeven voetoppervlakte en de voorspelde voetoppervlakte. Dit gaat ook eenvoudig met je rekentoestel door de lijst L3 − SLAK te maken. Men noemt dit verschil het *residu*.

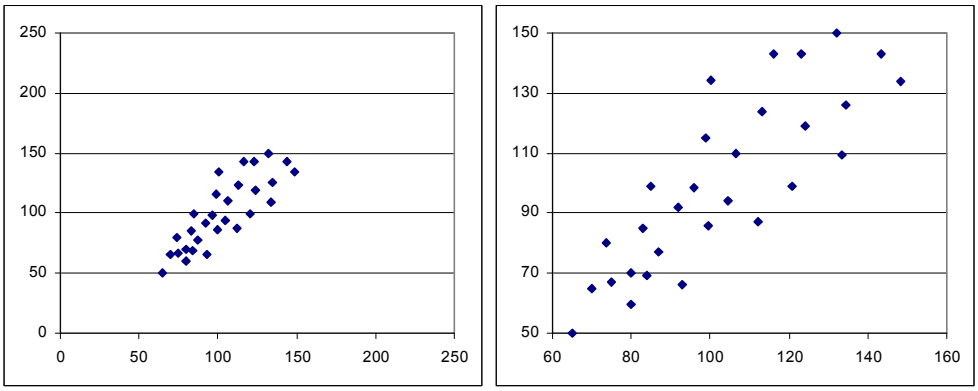
gewicht (g)	voetopp. (mm^2)	voorspelde voetopp.	residu observatie − voorspelling
0,64	29	29,54	
0,21	16	13,385	
0,85	35	37,43	
0,53	25	25,408	
0,02	4	6,2467	
0,01	1	5,871	
0,21	16	13,385	
0,18	20	12,258	
0,06	7	7,7495	
0,20	13	13,009	
0,07	7	8,1252	
0,01	3	5,871	
0,05	10	7,3738	
0,81	35	35,927	
0,53	25	25,408	
0,18	20	12,258	
0,06	7	7,7495	
0,20	13	13,009	
0,07	7	8,1252	
0,01	1	5,871	

8. Wat betekent het dat een residu positief/negatief is?
   (Bij een positief residu ligt het bijbehorende punt boven de regressielijn, bij een negatief residu ligt het punt onder de regressielijn.)

Bij dit eerste voorbeeld hebben de leerlingen de residu's concreet berekend. De rekenmachine berekent bij elke regressie echter ook automatisch de residu's. Ze worden bewaard in de lijst RESID. Je kunt deze lijst oproepen via [2nd] [LIST] [NAMES]. Het zijn de residu's die aangeven in hoeverre het model afwijkt van de realiteit.

4. Correlatiecoëfficiënt

Het inschatten van de mate van lineaire samenhang is met het blote oog niet zo evident. Zo zijn we geneigd de samenhang groter in te schatten in de linkse situatie dan in de rechtse.



Nochtans stellen beide dezelfde gegevensreeks voor. De samenhang *lijkt* alleen maar groter in de linkse afbeelding t.g.v. de keuze van de assen.

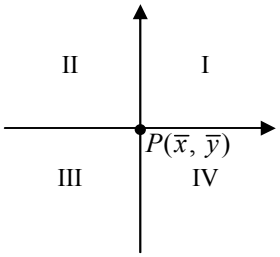
Vandaar de noodzaak om samenhang ook numeriek te bepalen.

In de opbouw hieronder vertrekken we van de begrippen positieve en negatieve correlatie en verfijnen we deze begrippen steeds verder tot we bij het begrip correlatiecoëfficiënt aanbelanden. We inspireerden ons hiervoor op [3].

Positieve en negatieve correlatie

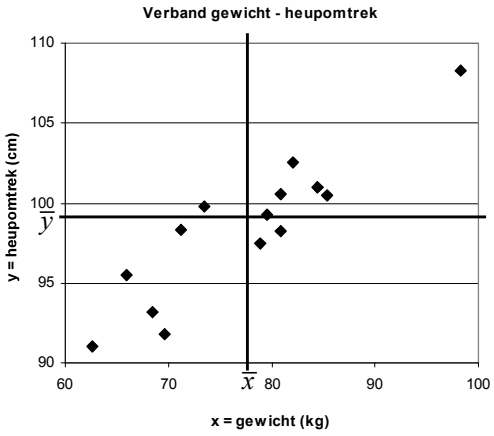
Men zegt dat er een positieve correlatie is, wanneer waarden boven het gemiddelde $\overline{x}$ van de ene variabele de neiging hebben om samen te gaan met waarden boven het gemiddelde $\overline{y}$ van de andere variabele en wanneer bovendien waarden onder het gemiddelde ook veeleer samen optreden.

Tekent men een assenstelsel waarvan de oorsprong door het punt $P(\overline{x}, \overline{y})$ gaat, m.a.w. door het punt dat de gemiddelden van beide grootheden weergeeft, dan zal men bij positieve correlatie de meeste punten aantreffen in het eerste en derde kwadrant.



Er is een negatieve correlatie wanneer waarden boven het gemiddelde van de ene variabele eerder optreden samen met waarden onder het gemiddelde van de andere en omgekeerd. De meeste punten liggen dan in het tweede en vierde kwadrant.

Beschouw nu het spreidingsdiagram hieronder. Het stelt een steekproef van 15 personen voor. We proberen in wat volgt de mate van positieve samenhang tussen beide variabelen cijfermatig uit te drukken.



gewicht (kg)	heupomtrek (cm)
80.9	100.6
68.5	93.2
85.4	100.5
66.0	95.5
69.6	91.8
71.2	98.3
79.5	99.3
73.4	99.8
82.1	102.5
78.9	97.5
98.3	108.3
62.7	91
80.9	98.2
84.4	101
84.4	101
gemiddelde	**gemiddelde**
77.7	98.6

Correlatie door tellen

Een eenvoudige maat voor de correlatie bestaat erin de punten in de kwadranten I en III een waarde +1 te geven en deze in II en IV een waarde –1. De som van de waarden van alle punten is dan een maat voor de correlatie. Indien we het aantal punten in het eerste kwadrant voorstellen door $n(\text{I})$ en die in de andere kwadranten door $n(\text{II})$, $n(\text{III})$ en $n(\text{IV})$, dan vinden we de volgende maat voor de correlatie:

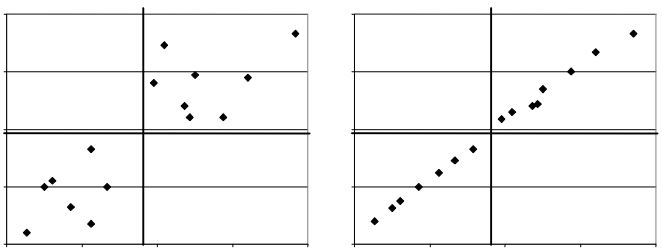
$$\big(n(\text{I}) + n(\text{III})\big) - \big(n(\text{II}) + n(\text{IV})\big)$$

Liggen de meeste punten in I en III, dan is deze waarde positief, in het andere geval is ze negatief.

In het bovenstaande voorbeeld is $\big(n(\text{I}) + n(\text{III})\big) - \big(n(\text{II}) + n(\text{IV})\big)$ gelijk aan + 9, wat een cijfermatige vertaling is van de *positieve* correlatie.

Punten dicht bij de assen minder laten wegen

De bovenstaande formule is echter nogal rudimentair. Zo vindt ze eenzelfde correlatie voor beide situaties hieronder, terwijl het rechtse spreidingsdiagram een sterkere positieve correlatie laat zien.



Punten dicht bij één van de rechten $x = \bar{x}$ of $y = \bar{y}$ (de getekende assen) geven een minder sterke indicatie van lineaire samenhang: een kleine wijziging in de meetwaarden had ze in een ander kwadrant kunnen plaatsen. Daarom is het zinvol deze punten minder zwaar te laten doorwegen. Punten die ver van beide assen liggen, moeten het zwaarst doorwegen. De afstanden (met een teken) van een punt (x_i, y_i) tot de rechten $x = \bar{x}$ en $y = \bar{y}$ zijn $(x_i - \bar{x})$ en $(y_i - \bar{y})$. Het product van beide is positief in I en III, negatief in II en IV, precies wat we nodig hebben. Een betere formule voor de correlatie is dus:

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

Voor het voorbeeld van de gewichten en de heupomtrekken van de 15 personen vinden we nu een correlatie van 520,8.

De invloed van eenheden wegwerken

Een ander nadeel dat hierbij nog overblijft, is dat een verandering van eenheden de berekende waarde beïnvloedt: drukken we de heupomtrek uit in m, dan wordt de waarde 100 keer kleiner dan wanneer ze in cm wordt uitgedrukt.

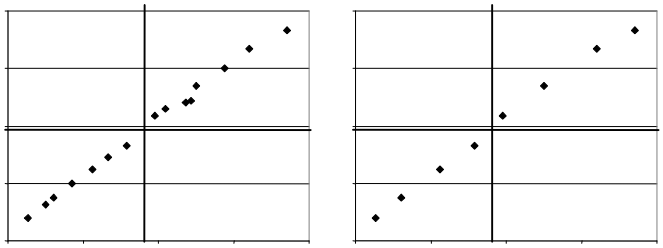
Een 'natuurlijke eenheid' om verschillen tussen een meetwaarde en het gemiddelde uit te drukken is de standaardafwijking. Stellen we de standaardafwijking van de *x*- en *y*-waarden voor door $s(x)$ resp. $s(y)$, met $s^2(x) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ en analoog voor $s(y)$, dan wordt het eenhedenprobleem opgelost m.b.v. de formule:

$$\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s(x)}\right)\cdot\left(\frac{y_i - \bar{y}}{s(y)}\right)$$

In ons voorbeeld is deze waarde 12,7.

De invloed van de steekproefgrootte wegwerken

Laatste probleem: alhoewel de onderstaande gegevens eenzelfde correlatie vertonen, zal de formule hierboven een grotere waarde hebben voor de linkse dataset, enkel en alleen omdat er meer waarden zijn om op te tellen.



Het volstaat het gemiddelde te nemen voor de *n* termen van de formule hierboven om een getal te verkrijgen dat onafhankelijk is van zowel de eenheden als het aantal meetwaarden. Om theoretische redenen delen statistici hier echter door *n* – 1 wanneer de gegevens een steekproef vormen uit de hele populatie. De betekenis blijft echter: een gemiddelde bepalen. De reden voor die *n* – 1 is vergelijkbaar

met die bij de berekening van de standaardafwijking van een steekproef, als schatter voor de standaardafwijking van de onderliggende populatie. Een formule voor correlatie met n in de noemer wordt gebruikt wanneer de gegevens de volledige populatie vormen, net zoals bij standaardafwijking.

De formule die we uiteindelijk verkrijgen, werd opgesteld door Pearson[1] en wordt daarom vaak de Pearson-correlatiecoëfficiënt genoemd. In wat volgt spreken we van correlatiecoëfficiënt zonder meer.

De correlatiecoëfficiënt r van een steekproef is dus gedefinieerd als:

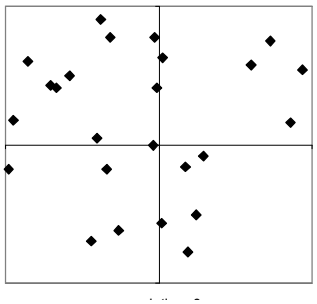
$$r = \frac{1}{n-1} \cdot \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s(x)} \right) \cdot \left(\frac{y_i - \bar{y}}{s(y)} \right).$$

Bemerk dat de waarden $\left(\frac{x_i - \bar{x}}{s(x)} \right)$ en $\left(\frac{y_i - \bar{y}}{s(y)} \right)$ niets anders zijn dan de z-scores of gestandaardiseerde waarden van de overeenkomstige meetresultaten: $r = \frac{1}{n-1} \cdot \sum_{i=1}^{n} z_{x_i} z_{y_i}$.

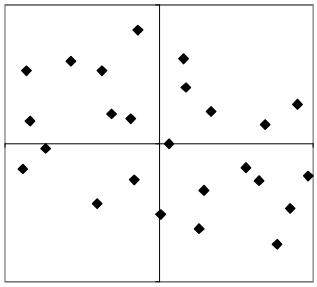
In het voorbeeld van de gewichten en heupomtrekken is de correlatiecoëfficiënt gelijk aan 0,91.

Enkele eigenschappen van de correlatiecoëfficiënt.

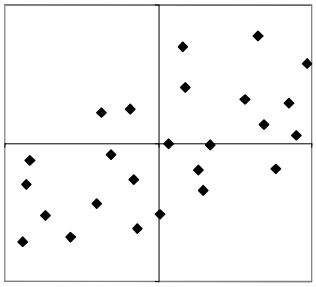
- r is altijd een waarde tussen –1 en 1. Waarden in de buurt van 0 wijzen op een zeer zwak lineair verband. Komt de waarde van r in de buurt van –1 en 1, dan liggen de punten in het spreidingsdiagram dicht bij een rechte. De waarden 1 en –1 worden slechts bereikt wanneer de punten precies op één rechte liggen.



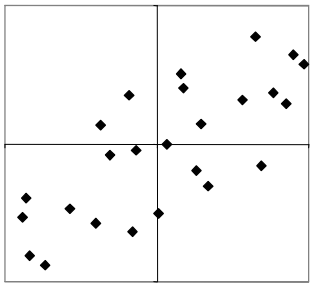
correlatie $r = 0$



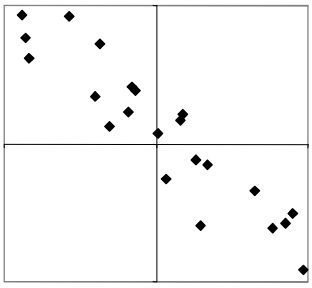
correlatie $r = -0{,}35$



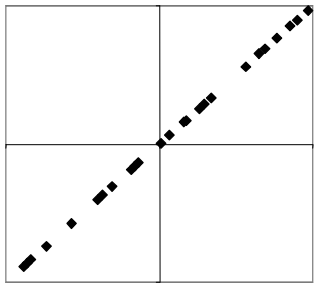
correlatie $r = 0{,}60$



correlatie $r = 0{,}75$



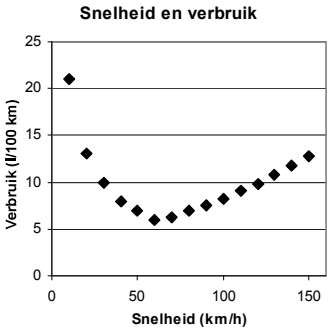
correlatie $r = -0{,}90$



correlatie $r = 1$

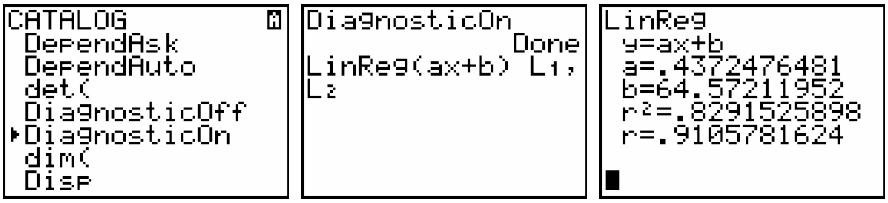
[1] Karl Pearson (1857-1936) gebruikte ze voor het eerst in zijn artikel uit 1896 “Regression, Heredity, and Panmixia,” *Phil. Trans. R. Soc., Ser.* A. 187, 253-318.

- Aangezien de correlatiecoëfficiënt gebaseerd is op gemiddelden en standaardafwijkingen, is hij zeer gevoelig voor uitschieters en scheve verdelingen. Daarom is het belangrijk telkens na te gaan of de *x*- en - en *y*-waarden wel symmetrisch verdeeld zijn.
- De correlatiecoëfficiënt drukt enkel de sterkte van een *lineair* verband uit. De puntenwolk hiernaast geeft het verband tussen de snelheid en het verbruik van een bepaald type auto (data uit [4]). Er is een duidelijk verband tussen beide variabelen. Nochtans is de correlatiecoëfficiënt amper – 0,17. Dit geeft correct aan dat er nauwelijks een lineair verband is.
  Dit voorbeeld maakt ook duidelijk dat het nooit volstaat om enkel de correlatiecoëfficiënt te berekenen om samenhang te onderzoeken: een essentieel startpunt is een spreidingsdiagram.



De meeste computerprogramma's en vele rekentoestellen bevatten een ingebouwde functie om de correlatiecoëfficiënt te berekenen; de TI-83/TI-84 echter niet.

Hij verschijnt wel als 'nevenproduct' wanneer je het toestel een regressierechte laat opstellen en dan nog enkel indien de *Diagnostics* mode is aangezet. Je vindt het nodige commando via de *Catalog*.



Voor de correlatie tussen heupomtrek en gewicht vinden we dus 0,91.

5. Niet-lineaire regressie

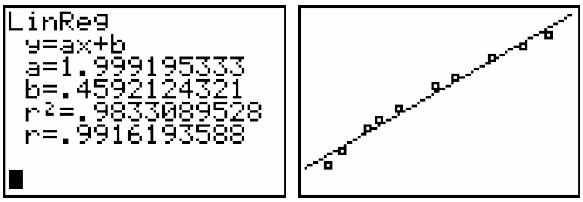
Een signaal dat de regressierechte niet goed is

De onderstaande tabel geeft de resultaten van een proef met een slinger (cijfers uit [2]).

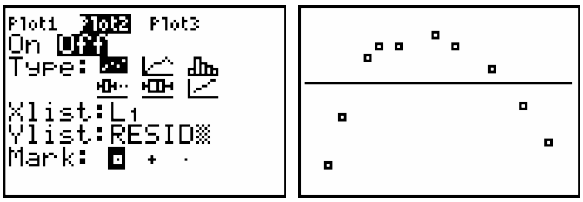
lengte van de slinger (in m)	slingertijd (in s)
0,07	0,53
0,1	0,63
0,15	0,78
0,17	0,83
0,21	0,91
0,28	1,06
0,32	1,13
0,39	1,25
0,45	1,34
0,5	1,41

Het verband tussen de lengte van een slinger en de periode van één slingerbeweging lijkt op het eerste gezicht niet eenduidig bepaald. Wanneer je namelijk start met een grote beginuitwijking zal de slingertijd immers groter zijn dan bij een bescheiden beginuitwijking. Wel, proefondervindelijk kunnen we vaststellen dat deze afhankelijkheid tussen beginuitwijking en slingertijd verdwijnt wanneer we met voldoende kleine uitwijkingen werken. Om een model te krijgen met slechts twee veranderlijken (lengte en slingertijd) moeten we ons bijgevolg beperken tot kleine amplituden.

Het uitvoeren van een lineaire regressie leidt op het eerste gezicht tot een zeer bevredigend resultaat: er is een zeer hoge correlatie en de regressierechte sluit heel goed aan bij de punten van het spreidingsdiagram.



Toch is er een knipperlicht: aan het linker- en rechteruiteinde liggen de punten van het spreidingsdiagram systematisch *onder* de regressierechte en in het midden liggen ze systematisch *boven* de regressierechte. De onderstaande grafiek, die de residu's toont, laat dit nog duidelijker zien:

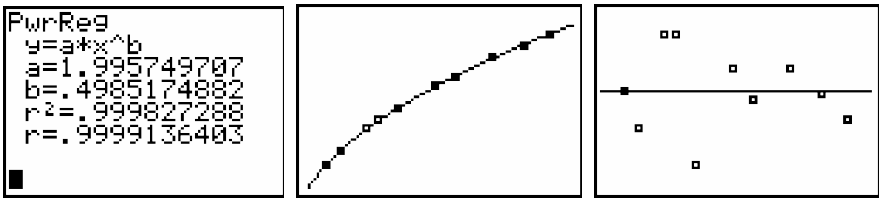


De vaststelling die we deden, wijst er op dat een rechte hier niet de meest aangewezen regressielijn is. Een gebogen lijn zal in dit geval beter bij de punten aansluiten. We moeten de rechte aan de linker- en rechterkant a.h.w. wat naar onder plooien. Dan komt de regressielijn in het midden ook wat naar boven.

Een regressiekromme via niet-lineaire regressie

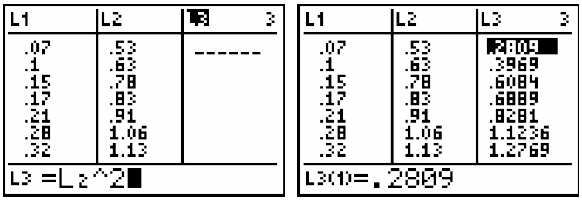
De zoektocht naar een betere regressielijn kan dan beginnen. Met een grafische rekenmachine bij de hand is dat geen probleem meer. Je kunt experimenteren met verschillende regressies: kwadratisch, derdegraads, exponentieel, logaritmisch, ... tot je een bevredigend resultaat vindt.

Maar het kan ook anders. Misschien herinner je je uit de fysicales dat de slingertijd (bij benadering) evenredig is met de vierkantswortel uit de lengte van de slinger. Dat zet je op het spoor van een regressie met een machtsfunctie, die inderdaad een zeer bevredigend resultaat geeft. De berekeningen met de grafische rekenmachine vind je op de volgende bladzijde. De exponent in de berekende vergelijking is bijna gelijk aan 0,5. Het getal r dat getoond wordt, is niet meer de correlatiecoëfficiënt uit de vorige paragraaf. Het is de vierkantswortel uit de determinatiecoëfficiënt, het getal dat in de onderstaande schermafdruk r^2 genoemd wordt. We gaan dieper in op dat getal in paragraaf 7. Bij lineaire regressie is de determinatiecoëfficiënt gelijk aan het kwadraat van de correlatiecoëfficiënt maar in het algemeen is dat niet zo. Het feit dat de determinatiecoëfficiënt zeer dicht bij 1 ligt, geeft aan dat de overeenstemming tussen het model en de meetresultaten zeer goed is. We zien ook dat de getekende kromme bijna perfect door de punten van het spreidingsdiagram gaat. De residu's vertonen geen regelmaat. Op basis daarvan kunnen we geen verdere verbeteringen aan het model voorstellen.

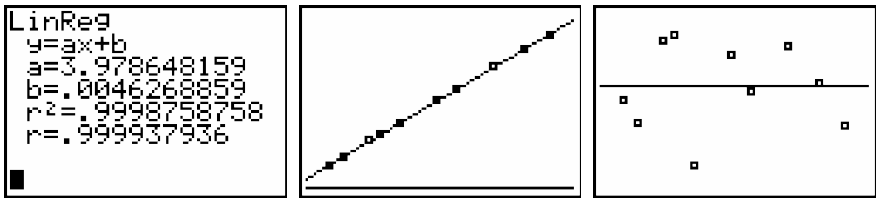


Lineaire regressie na transformatie

Er is nog een andere manier om een goede regressiekromme te vinden. Als we er van uitgaan dat de slingertijd evenredig is met de vierkantswortel uit de slingerlengte, dan is het kwadraat van de slingertijd evenredig met de slingerlengte. Daarom maken we een nieuwe lijst, met daarin de kwadraten van de slingertijden.



Vervolgens voeren we een lineaire regressie uit van L_3 (kwadraten van de slingertijden) op L_1 (slingerlengtes).

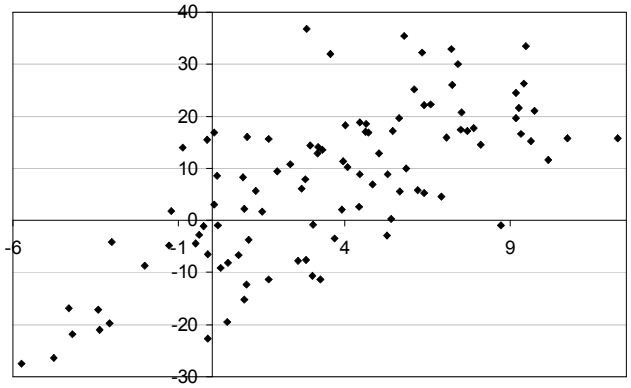


In dit voorbeeld hebben we de niet-lineaire regressie vervangen door een lineaire regressie na het toepassen van een transformatie op *y*-waarden. Dat is een techniek die heel vaak gebruikt wordt. Zo kan een exponentiële regressie vervangen worden door een lineaire wanneer je de logaritme neemt van de *y*-waarden, kan een machtregressie vervangen worden door een lineaire wanneer je zowel de logaritme van de *x*- als van de *y*-waarden neemt ... Merk op dat de regressiekromme die je krijgt bij (bijvoorbeeld) een exponentiële regressie van *y* op *x* niet noodzakelijk overeenkomt met de kromme die je krijgt door de regressierechte van $\ln y$ op *x* omgekeerd te transformeren. We verwijzen naar de loep uit Uitwiskeling 10/3 (over 'maak recht wat krom is') en [5] (over logaritmisch papier) voor verwante zaken, maar dan buiten de context van regressie.

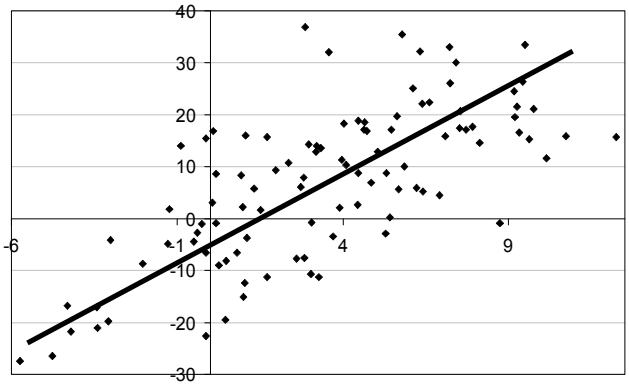
6. De regressierechte nader bekeken

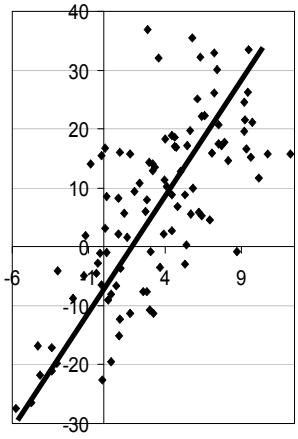
In deze paragraaf leren we dat de regressierechte niet zomaar de rechte is die 'het beste aansluit bij de punten' van het spreidingsdiagram. Het is een heel specifieke rechte die volgens een welomschreven criterium bepaald wordt. We brengen dit (voor jullie) aan via een klein experiment dat we gedaan hebben bij het voorbereiden van deze loep. Het experiment doorbreekt de vanzelfsprekendheid van 'het best aansluiten bij de punten'. Iets dergelijks kan je in de les ook doen om de leerlingen te motiveren voor de definitie van de regressierechte.

De onderstaande figuur toont een puntenwolk.



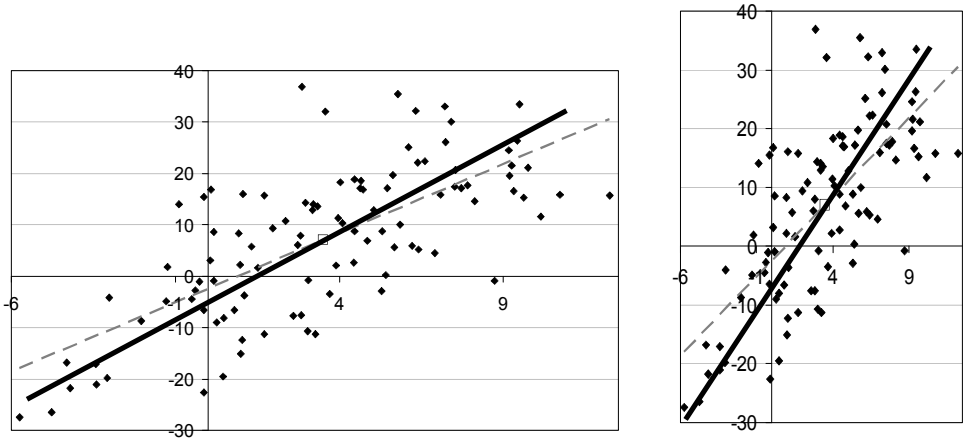
We hebben een (beperkt) aantal proefpersonen gevraagd de lijn te tekenen die het beste aansluit bij deze punten. Hieronder vind je twee typische uitkomsten van ons kleine experiment. In de rechtse grafiek is de puntenwolk dezelfde, maar is de schaalverdeling anders. De proefpersonen wisten niet dat het om dezelfde puntenwolk ging.





Een eerste vaststelling die we kunnen maken, is dat de helling van beide rechten duidelijk verschillend is: in de linkse figuur ongeveer 3 en in de rechtse figuur ongeveer 4. Blijkbaar speelt de vorm waarin de puntenwolk gepresenteerd wordt een rol bij wat we ‘de best aansluitende rechte’ vinden ...

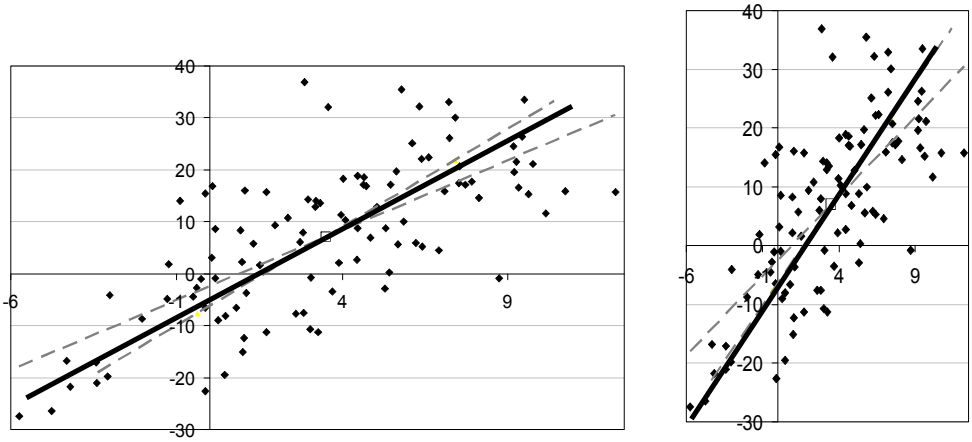
In de volgende twee figuren hebben we de regressielijn toegevoegd (het gaat in beide figuren natuurlijk over dezelfde rechte, al lijkt dat op het eerste gezicht niet zo te zijn omwille van de andere schaalverdeling).



We stellen vast dat de proefpersonen het centrum van de puntenwolk goed weergegeven hebben. Het vierkante blokje dat je ongeveer in het midden ziet, geeft het punt $(\bar{x}, \bar{y})$ aan. Het is bekend dat de regressielijn door dat punt gaat. De rechten die door de proefpersonen getekend werden, gingen systematisch ook (ongeveer) door dat punt.

Met de helling van de rechten is het anders gesteld. De proefpersonen tekenden systematisch een rechte met een grotere helling dan die van de regressielijn!

Hieronder zie je nogmaals dezelfde figuren, aangevuld met weer een nieuwe rechte, die wel eens de *SD-rechte* genoemd wordt (het gaat in beide figuren over dezelfde rechte; we verklaren verderop over welke rechte het precies gaat).



Het lijkt er sterk op dat de proefpersonen bij het 'intuïtief schatten' van de 'best bij de punten aansluitende rechte' *niet* uitkomen bij de regressielijn, maar bij de SD-rechte! In ons experiment was de rechte van de proefpersonen systematisch wat minder steil dan de SD-rechte als het spreidingsdiagram in 'landscape' aangeboden werd en wat steiler als het spreidingsdiagram in 'portrait' aangeboden werd. Je vindt deze SD-rechte bijvoorbeeld in [1], [8]. We verklaren nu wat de SD-rechte is. Ze gaat (net als de regressielijn) door het centrum $(\bar{x}, \bar{y})$ van de puntenwolk.

Bij een puntenwolk met een positieve samenhang heeft ze als helling

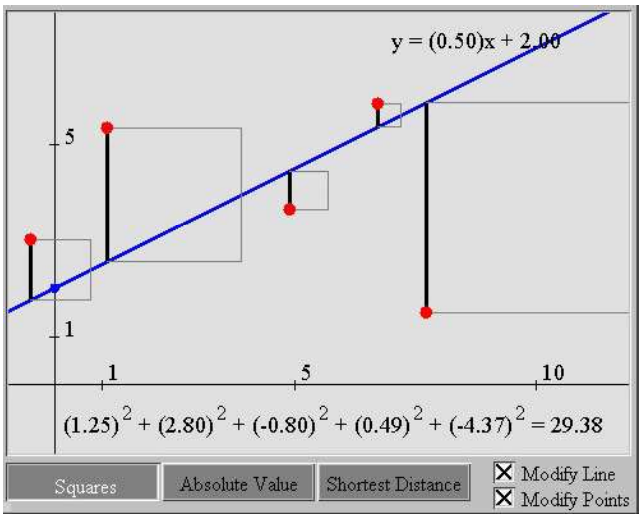
$$\frac{s(y)}{s(x)},$$

d.w.z. één standaardafwijking (van de x-waarden) naar rechts komt overeen met één standaard-afwijking (van de y-waarden) naar boven. Dat verklaart meteen de naam: SD komt van standard deviation, de Engelse benaming voor standaardafwijking. De helling van de regressielijn is kleiner, namelijk

$$\frac{r \cdot s(y)}{s(x)},$$

waarbij r de correlatiecoëfficiënt is. In ons voorbeeld is $r = 0{,}71$. Als de puntenwolk een negatieve samenhang vertoont, moeten we een minteken vóór de breuk plaatsen.

Om het fenomeen dat we hierboven vastgesteld hebben te kunnen verklaren, moeten we eerst wat dieper ingaan op het criterium dat gebruikt wordt om de regressielijn te definiëren. Voor elk punt uit het spreidingsdiagram wordt de verticale afstand tot de rechte bepaald. De regressierechte is dan de rechte waarvoor de som van de kwadraten van al deze verticale afstanden minimaal is. Dat wordt mooi geïllustreerd in een applet die we vonden op de website van NCTM, de vereniging van de Amerikaanse wiskundeleraren (zie [10] en onderstaande figuur). Het kwadraat van de verticaal gemeten afstand wordt gevisualiseerd aan de hand van vierkanten. Om de regressierechte te vinden, moet je met de muis de positie van de rechte veranderen zo dat de gezamenlijke oppervlakte van de vierkanten minimaal wordt. Onderaan wordt de waarde van die totale oppervlakte gegeven. Voor alle duidelijkheid: de rechte in de figuur is (lang) niet de regressierechte!



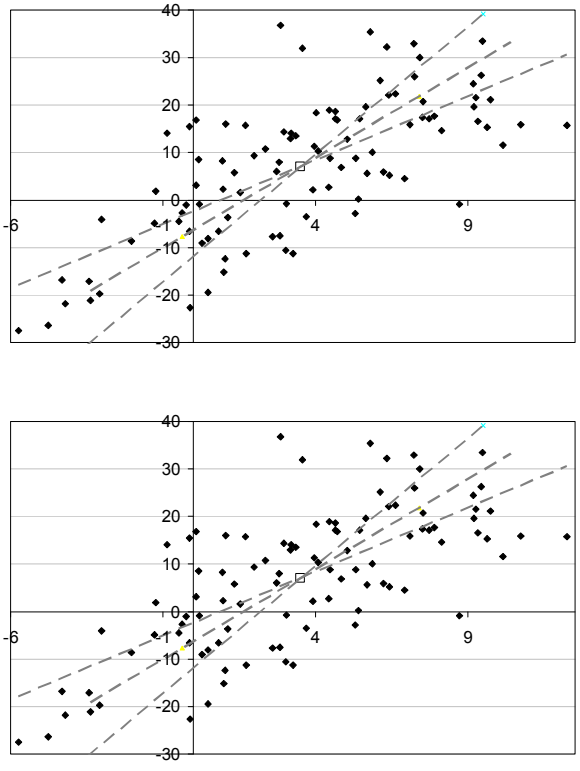
Bij een andere applet ([11], met de titel 'Regression by eye') krijg je geen vierkanten meer te zien (omdat de puntenwolken uit veel meer punten bestaan en daardoor realistischer zijn). De som van de kwadraten van de verticale afstanden wordt wel getoond en die moet je gebruiken om de regressierechte te bepalen. Als je je best gedaan hebt om de goede rechte te vinden (en ook als je je best niet gedaan hebt ...), kan de applet de echte regressielijn voor je tekenen. Beide applets zijn echte aanraders om je leerlingen meer inzicht te geven in de betekenis van een regressielijn.

Het opstellen van de formules voor de coëfficiënten van de regressielijn vereist het bepalen van een extremum van een functie van twee veranderlijken. Het is dus (haast) niet mogelijk om dat in de les te

doen. Als je bereid bent om aan te nemen dat de regressierechte door $(\bar{x}, \bar{y})$ gaat, is het probleem herleid tot het bepalen van een minimum van een functie van één veranderlijke, namelijk de richtingscoëfficiënt van de rechte. Ook dan zouden we het niet onmiddellijk een plaats geven bij het onderdeel statistiek (omdat we meer aandacht willen hebben voor de statistische kant van de zaak dan voor de wiskundige onderbouw), maar eerder in de lessen die gewijd zijn aan extremumproblemen.

Nu we weten hoe de regressierechte gedefinieerd is, begrijpen we waarom de proefpersonen in ons experiment een andere rechte verkozen boven de regressierechte. Bij de regressierechte treedt er immers een – op het eerste gezicht onnatuurlijke – asymmetrie op tussen de *x*- en *y*-waarden. Als we op het gevoel, zonder duidelijk geëxpliciteerd criterium, op zoek gaan naar de 'best passende rechte' gaan we wellicht spontaan meer symmetrisch te werk.

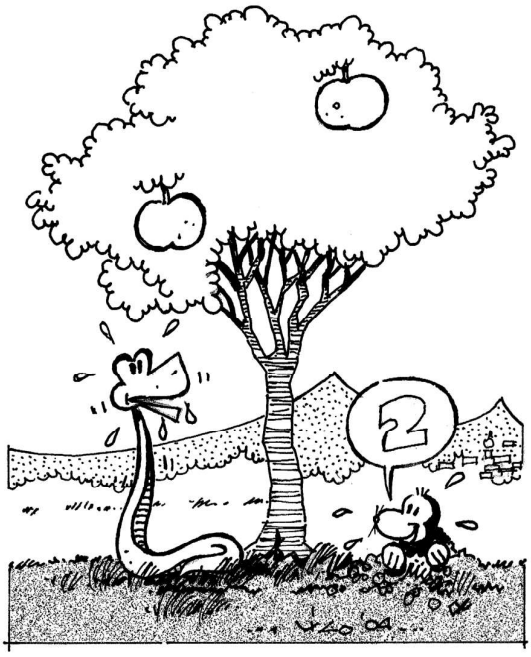
De asymmetrie bij de regressierechte vinden we terug in de benaming: de rechte die we in de vorige figuren getekend hebben, is de regressierechte *van y op x*. De regressierechte *van x op y* is een andere rechte, namelijk de rechte die de som van de kwadraten van de *horizontaal* gemeten afstanden tot de rechte minimaliseert (tenminste als we de assen in het spreidingsdiagram laten zoals ze zijn, wat men meestal niet doet). Dat leidt effectief tot een andere rechte. De onderstaande figuur toont nogmaals het spreidingsdiagram dat we aan onze proefpersonen voorlegden. De rechte die door de proefpersoon getekend werd, hebben we nu niet meer weergegeven. De twee minst steile rechten zijn de regressierechte van *y* op *x* en de SD-rechte. De steilste rechte is nieuw. Het is de regressierechte van *x* op *y*. De helling van deze regressierechte is $\frac{s(y)}{r \cdot s(x)}$. De regressielijn van *x* op *y* is altijd steiler dan de regressielijn van *y* op *x*. De verhouding van hun richtingen is gelijk aan de determinatiecoëfficiënt r^2 (zie paragraaf 7).



Uit de formules voor de hellingen van de drie rechten wordt duidelijk dat ze dichter bij elkaar liggen als de correlatiecoëfficiënt (in absolute waarde) groter wordt. In het bijzonder: als de correlatie dicht bij 1 of -1 ligt, dan stemt de regressierechte ongeveer overeen met de rechte die op het gevoel 'het best bij de punten aansluit'. In paragraaf 3 was het dus geen probleem dat we de leerlingen vroegen om de best aansluitende rechte op het gevoel te tekenen.

Misschien vraag je je af *waarom* er met de verticaal of horizontaal gemeten afstand gewerkt wordt en niet met de gewone loodrechte afstand tot de rechte. Daar is een goede reden voor. Bij het berekenen van de loodrechte afstand worden de x- en y-coördinaten van de punten met elkaar gecombineerd tot één getal. Meestal wordt echter in de horizontale en verticale richting met andere eenheden gewerkt (bijvoorbeeld: leeftijden in jaar op de horizontale as en lengtes in cm op de verticale as). De loodrechte afstand heeft dan helemaal geen betekenis.

De asymmetrie tussen x en y blijft niet beperkt tot de berekeningen. Bij regressie hebben de twee veranderlijken een verschillende rol, net zoals bij functies: y is de afhankelijke variabele en x is de onafhankelijke variabele. Meestal gebruikt men wel andere namen: y is de *te verklaren variabele* en x is de *verklarende variabele*. Als we de regressierechte gebruiken om voorspellingen te maken, mogen we dat alleen in de goede richting doen. De regressierechte van y op x wordt gebruikt om op basis van een x-waarde een voorspelling te maken voor een y-waarde, maar niet omgekeerd (het is in het kader van deze loep niet mogelijk om helemaal uit te leggen waarom dat zo is). Dat zijn we natuurlijk niet gewoon vanuit de functieleer. Daar kunnen we bij een gegeven voorschrift onbeperkt beelden en inverse beelden berekenen. Bemerk ook dat er geen asymmetrie tussen x en y is wanneer we ons beperken tot het bestuderen van de samenhang m.b.v. de correlatiecoëfficiënt.



7. Determinatiecoëfficiënt

In deze paragraaf ontmoeten we een mooie interpretatie van de correlatiecoëfficiënt die verband houdt met de regressierechte. We leggen alles uit aan de hand van een voorbeeld.

Determinatiecoëfficiënt

Veronderstel dat we van 12 jongens de lengte (in cm) gemeten hebben:

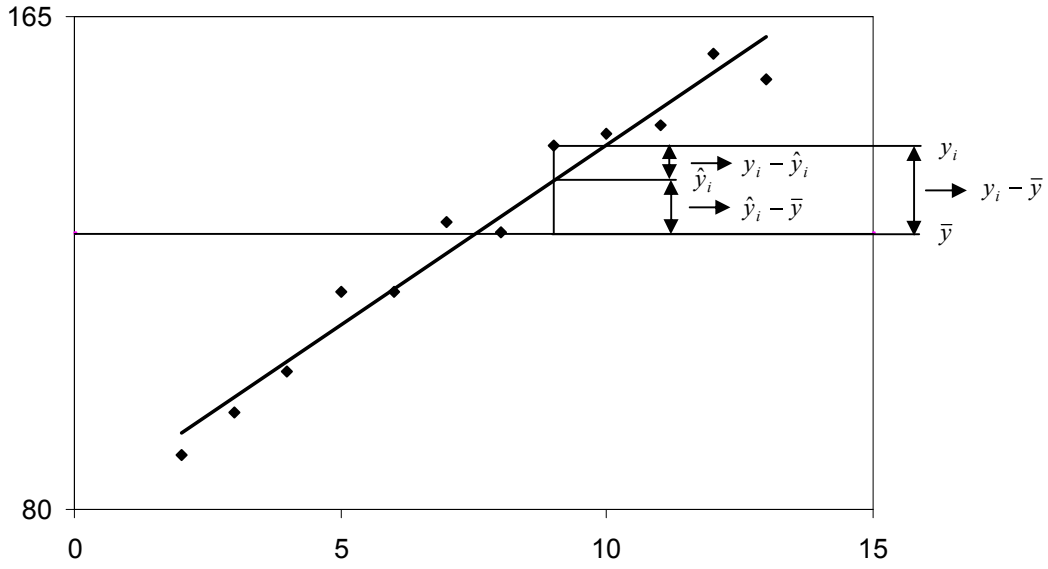
142,8 96,8 144,8 117,5 158,6 154,1 89,3 146,1 117,5 103,6 129,4 127,7

We zien op het zicht dat er nogal wat spreiding is in de lengtes. In de tabel hieronder vinden we nu ook de leeftijd van de jongens.

leeftijd (in jaar)	lengte (in cm)
2	89,3
3	96,8
4	103,6
5	117,5
6	117,5
7	129,4
8	127,7
9	142,8
10	144,8
11	146,1
12	158,6
13	154,1

Het is bij nader inzien nogal wiedes dat de lengtes sterk gespreid zijn. Dat heeft te maken met de leeftijd. De leeftijden van de jongens zijn verschillend, en bij een verschillende leeftijd hoort een verschillende lengte. Met andere woorden: de spreiding van de lengtes is (minstens voor een deel) terug te brengen tot de spreiding in de leeftijden. Men zegt: een deel van de spreiding van de lengtes *wordt verklaard door* de spreiding van de leeftijden. Hoewel de term ‘verklaren’ impliciet verwijst naar een oorzakelijk verband (wat hier wel degelijk het geval is), wordt hij ook gebruikt wanneer er geen oorzakelijk verband is.

We zullen nu onderzoeken *hoeveel* van de spreiding van de lengtes verklaard wordt door het verschil in leeftijd. We zullen daarbij gebruik maken van de variantie om de spreiding te meten. De variantie is gebaseerd op de afwijkingen van de lengtes ten opzichte van de gemiddelde lengte. We bekijken deze afwijkingen eerst in detail m.b.v. het onderstaande spreidingsdiagram van de lengte op de leeftijd.



We hebben één van de punten (x_i, y_i) gekozen. Het verschil tussen de specifieke lengte y_i en de gemiddelde lengte $\bar{y}$ is aangegeven. De schuine rechte is de regressierechte van de lengte op de leeftijd. Deze regressierechte bepaalt voor x_i een bijbehorende *y*-waarde $\hat{y}_i$ (a.h.w. de lengte die theoretisch gezien bij de leeftijd x_i hoort). We kunnen het verschil $y_i - \bar{y}$ nu als volgt schrijven als een som van twee termen:

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i) = (\hat{y}_i - \bar{y}) + e_i .$$

Beide termen zijn ook in de figuur aangeduid. De eerste term drukt uit dat het niet meer dan normaal is dat de lengte van deze jongen afwijkt van het gemiddelde: hij is immers ouder dan gemiddeld. Dit gedeelte van $y_i - \bar{y}$ wordt dus verklaard door de samenhang van lengte en leeftijd. Het andere gedeelte van $y_i - \bar{y}$, het residu, kan niet aan de leeftijd toegeschreven worden (maar bijvoorbeeld aan genetische factoren, voeding ...).

Men kan aantonen dat men de variantie op dezelfde manier kan opsplitsen in twee delen. Dat is niet vanzelfsprekend omdat we bij de variantie werken met de kwadraten van de afwijkingen! Concreet:

$$s^2(y) = s^2(\hat{y}) + s^2(e) .$$

In het linkerlid staat de *totale variantie*. De eerste term in het rechterlid is de variantie die we zouden krijgen als de lengtes zich perfect zouden gedragen zoals de regressierechte bepaalt. Dat is het gedeelte van de variantie van de lengtes dat terug te brengen is tot de spreiding in de leeftijden. Deze term noemt men de *variantie die door de regressie verklaard wordt*. De tweede term in het rechterlid is de variantie van de residu's. Dat is het gedeelte van de variantie dat niet teruggebracht kan worden tot de verschillen in leeftijd (en waarvoor er een aantal andere oorzaken bestaan).

De *determinatiecoëfficiënt* (ook het *verklarend vermogen* genoemd) is de verhouding

$$d = \frac{s^2(\hat{y})}{s^2(y)}$$

en wordt vaak onder de vorm van een percentage gegeven (de *d* die we voorlopig gebruiken om de determinatiecoëfficiënt aan te geven, is niet het algemeen gebruikelijke symbool). In ons voorbeeld is de determinatiecoëfficiënt gelijk aan 0,9636. Dat betekent dat 96,36% van de variantie in de lengtes verklaard wordt door het verschil in leeftijd.

Bemerk dat het criterium dat gebruikt wordt om de regressierechte te bepalen, inhoudt dat we $s^2(e)$ minimaliseren. Met andere woorden: de regressierechte is de rechte waarvoor het verklarend vermogen *d* zo groot mogelijk is.

Determinatiecoëfficiënt en correlatiecoëfficiënt

De definitie van de determinatiecoëfficiënt geeft de betekenis van dit begrip weer: hoeveel procent van de variantie van de *y*-waarden wordt verklaard door de regressie ervan op een andere variabele *x* ? Voor het berekenen van de determinatiecoëfficiënt kunnen we echter beter op een andere manier te werk gaan. De vergelijking van de regressielijn in het voorbeeld is $\hat{y} = 6{,}2140x + 80{,}7$. Op basis van deze vergelijking vinden we een verband tussen de variantie van de waarden die $\hat{y}$ aanneemt en de variantie van de *x*-waarden. We moeten de *x*-waarden eerst vermenigvuldigen met het getal 6,2140. De variantie wordt met het kwadraat van deze factor vermenigvuldigd. Daarna moeten we nog 80,7 optellen. Dat heeft geen verdere invloed op de variantie. We vinden dus (in het algemeen): $s^2(\hat{y}) = a^2 s^2(x)$, met *a* de helling van de regressierechte. De determinatiecoëfficiënt wordt dan

$$d = \frac{a^2 \cdot s^2(x)}{s^2(y)}.$$

Als we deze uitdrukking combineren met de formule $a = r \cdot \frac{s(y)}{s(x)}$ voor de helling van de regressierechte zien we in dat de determinatiecoëfficiënt niets anders is dan het kwadraat van de correlatiecoëfficiënt:

$$d = r^2.$$

De determinatiecoëfficiënt bij niet-lineaire regressie

We zijn hierboven vertrokken van lineaire regressie. Bij niet-lineaire regressie is de opsplitsing van de variantie in twee termen niet langer geldig. De determinatiecoëfficiënt (verklarend vermogen) wordt dan gedefinieerd als

$$d = 1 - \frac{s^2(e)}{s^2(y)}$$

(wat in het geval van lineaire regressie neerkomt op de oorspronkelijke definitie). Bij niet-lineaire regressie is de determinatiecoëfficiënt ook niet langer gelijk aan het kwadraat van de correlatie-coëfficiënt. Soms wordt de determinatiecoëfficiënt toch nog met het symbool r^2 aangegeven (zoals op de rekenmachine), soms wordt het symbool R^2 gebruikt.

8. Regressielijn en verklarende statistiek

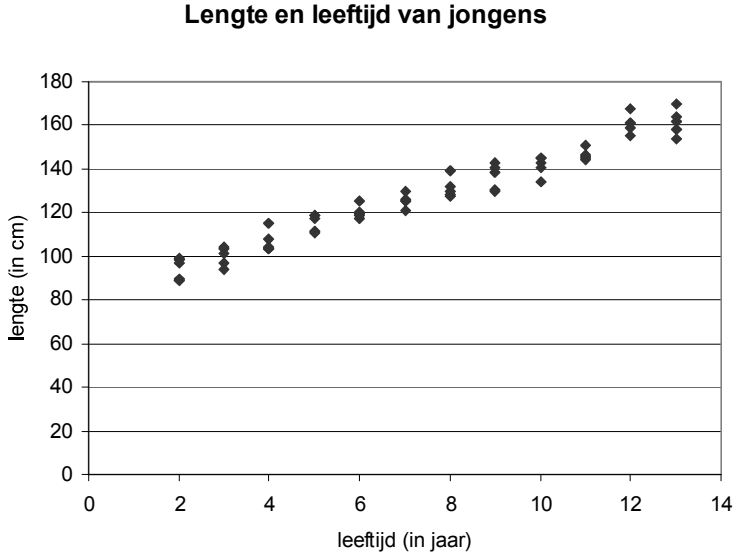
De verklarende statistiek houdt zich bezig met situaties waarin we gegevens hebben over een steekproef, terwijl we conclusies willen over de gehele populatie. In de voorgaande paragrafen hebben we niet stilgestaan bij het verschil tussen steekproef en populatie. De regressielijn is toen behandeld geworden als een soort 'vuistregel' die de samenhang tussen de twee variabelen weergeeft. Nu bekijken we de regressielijn wel vanuit het perspectief van de verklarende statistiek. Op basis van een dergelijke analyse kunnen we dan (met een zekere betrouwbaarheid en foutenmarge) uitspraken doen over de coëfficiënten, voorspellingen doen voor andere waarden van x ...

De gegevens

We werken met (fictieve) gegevens over de lengte (in cm) en leeftijd (in jaar) van 60 jongens (zie de onderstaande tabel). Verderop maken we ook gebruik van een grotere dataset, met gegevens over 1200 jongens. Je vindt de gegevens ook op onze website.

2	89,3	4	103,6	6	117,5	8	127,7	10	144,8	12	158,6
2	98,8	4	103,3	6	125,1	8	128,2	10	143,0	12	167,5
2	98,4	4	115,5	6	120,4	8	131,6	10	140,6	12	161,3
2	97,2	4	108,0	6	119,2	8	129,4	10	145,0	12	154,9
2	88,6	4	104,1	6	118,7	8	138,9	10	134,2	12	161,2
3	96,8	5	117,5	7	129,4	9	142,8	11	146,1	13	154,1
3	104,2	5	110,8	7	126,1	9	140,6	11	146,6	13	161,5
3	93,7	5	111,2	7	125,6	9	130,8	11	145,1	13	169,9
3	103,2	5	118,5	7	125,1	9	129,5	11	151,2	13	164,0
3	101,0	5	119,1	7	121,0	9	138,7	11	144,4	13	158,5

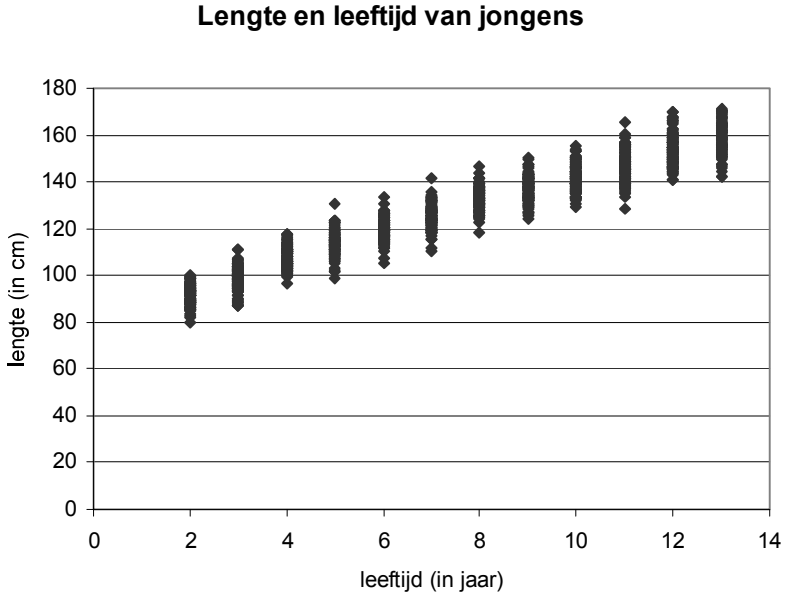
Spreidingsdiagrammen



In het bovenstaande spreidingsdiagram is de lengte van de jongens uitgezet tegen hun leeftijd. Op de x-as is de leeftijd (in jaar) uitgezet en op de y-as staat de lengte (in cm). We gaan immers de lengte schatten op basis van de leeftijd en niet omgekeerd.

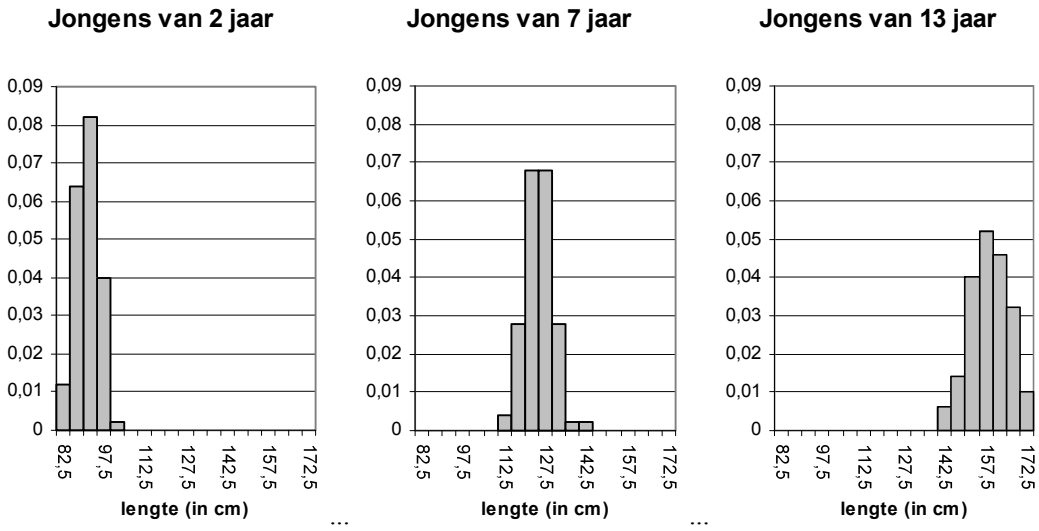
We zien dat er bij éénzelfde x-waarde (leeftijd) verschillende y-waarden (lengtes) horen. We hebben voor elke leeftijd vijf (willekeurig gekozen) jongens gemeten (in het spreidingsdiagram zijn echter niet altijd vijf duidelijk onderscheiden punten te zien). Deze vijf jongens hadden niet dezelfde lengte. Het is niet moeilijk om ons voor te stellen hoe dit komt: enerzijds wordt de lengte van een jongen natuurlijk wel beïnvloed door de leeftijd, maar ook heel wat andere factoren spelen een rol (lengte van de ouders, ziektes die de jongen eventueel meegemaakt heeft, voeding ...).

De jongens uit het spreidingsdiagram vormen een steekproef uit de populatie van alle (Belgische) jongens. In feite horen bij één waarde van x niet 5, maar een enorm groot aantal y-waarden. In het onderstaande spreidingsdiagram hebben we met een grotere steekproef gewerkt: voor elke leeftijd 100 jongens.

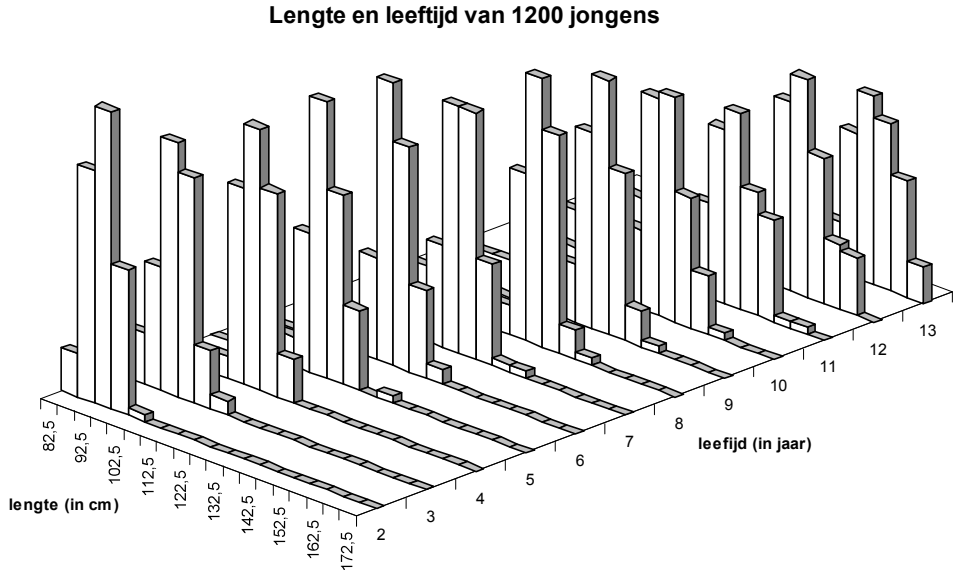


Histogrammen

We bekijken de verdeling van de lengtes bij de verschillende leeftijden nu meer in detail. De figuren hieronder tonen histogrammen. De eerste figuur toont het histogram voor de lengte van de 100 jongens van 2 jaar. Zo kunnen we voor elke leeftijd een histogram maken (we hebben gewerkt met klassen van 5 cm breed en we hebben op de verticale as de frequentiedichtheid uitgezet). Voor de leeftijd van 7 jaar en van 13 jaar hebben we de histogrammen ook gemaakt.



Hieronder zie je alle histogrammen van de lengtes bij vaste leeftijd in één figuur.

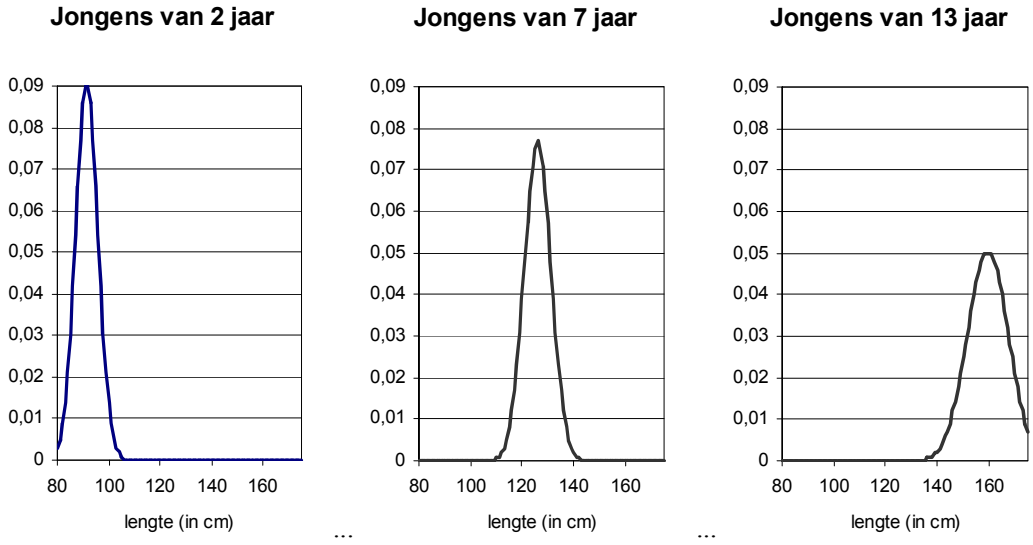


Ook in deze figuren zien we

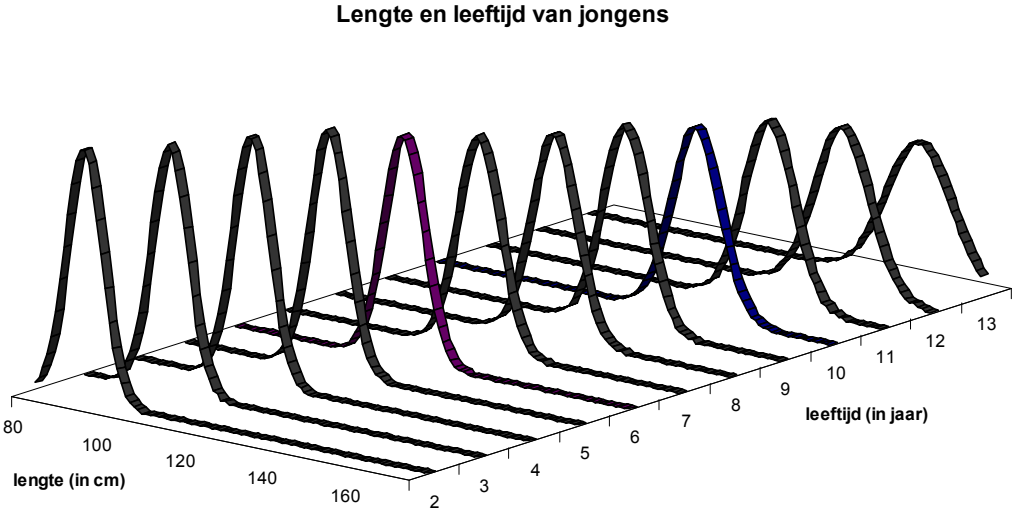
- enerzijds de invloed van de leeftijd (het histogram schuift naar rechts), en
- anderzijds de invloed van de andere factoren (er is niet één vaste lengte per leeftijd, maar variabiliteit in de lengtes die bij één leeftijd horen).

De hele populatie

Als we niet met een (kleine of grote) steekproef werken maar met de hele populatie, zullen we idealiserende veronderstellingen maken over de verdeling van de lengtes. Meer bepaald zullen we er van uitgaan dat bij elke leeftijd de lengte van de jongens van die leeftijd normaal verdeeld is. De onderstaande figuur toont de normale dichtheden die horen bij de jongens van 2, 7 en 13 jaar oud.



Deze normale dichtheden hebben een verschillende verwachtingswaarde (soms zegt men ‘gemiddelde’ i.p.v. ‘verwachtingswaarde', maar dat is eigenlijk verkeerd). Dat geeft weer dat de lengte afhangt van de leeftijd. De andere factoren die de lengte van de jongens beïnvloeden, zorgen er voor dat de lengtes variëren rond deze verwachtingswaarde (bemerk, terzijde, dat de variabiliteit groter is bij oudere jongens). Hieronder zie je de normale dichtheden bij elke leeftijd in één figuur.



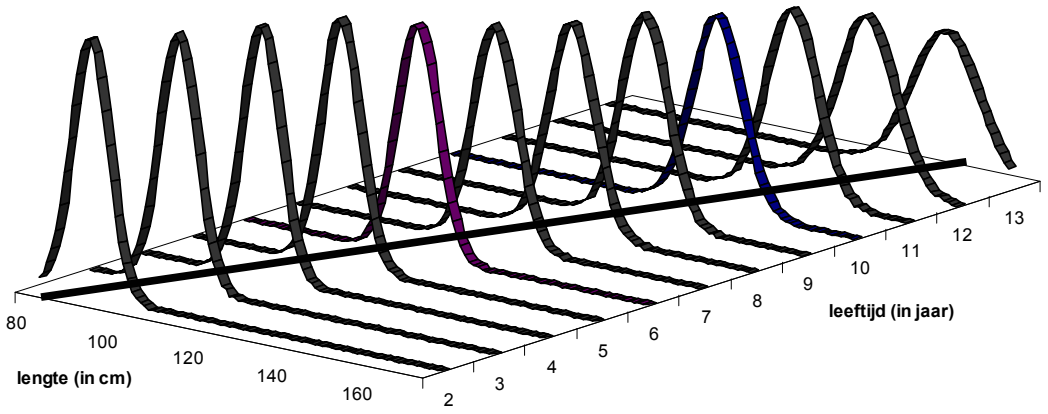
Het verband met regressie

Als we een lineaire regressie uitvoeren, dan gaan we er impliciet van uit dat de verwachtingswaarden van die normale verdelingen afhangen van de leeftijd volgens een eerstegraadsfunctie (in toepassingen spreekt men meestal over een lineaire functie i.p.v. een eerstegraadsfunctie). Met andere woorden: we gaan er van uit dat er getallen α en β bestaan zo dat

- de lengtes van de 2-jarige jongens normaal verdeeld zijn met verwachtingswaarde $\alpha \cdot 2 + \beta$,
- de lengtes van de 3-jarige jongens normaal verdeeld zijn met verwachtingswaarde $\alpha \cdot 3 + \beta$,
- ...
- de lengtes van de 13-jarige jongens normaal verdeeld zijn met verwachtingswaarde $\alpha \cdot 13 + \beta$.

In de figuur hieronder is de grafiek van deze eerstegraadsfunctie getekend.

Lengte en leeftijd van jongens



Laten we er even van uitgaan dat $\alpha = 6$ en $\beta = 83$ (hoewel dat wellicht niet helemaal de juiste waarden zijn). De verwachtingswaarde voor de lengte van de 7-jarige jongens is dan 125 cm. Als we één bepaalde jongen nemen van 7 jaar, dan is de kans zeer klein dat zijn lengte gelijk zal zijn aan 125 cm. Er zijn immers nog andere factoren dan de leeftijd alleen die de lengte bepalen. De eerste 7-jarige jongen uit onze kleine steekproef bijvoorbeeld, meet 129,4 cm. De afwijking (in cm) t.o.v. de verwachtingswaarde zullen we met ε noteren, voorzien van een onderindex die het rangnummer aangeeft van het gegeven. De eerste 7-jarige jongen is de 26-ste jongen uit de steekproef. Dus is $\varepsilon_{26} = 4{,}4$. Voor de laatste 7-jarige jongen uit onze kleine steekproef (met een lengte van 121,0 cm) vinden we $\varepsilon_{30} = -4{,}0$. Deze afwijkingen, die dus de invloed weergeven van de andere factoren die de lengte bepalen, zijn normaal verdeeld, net als de lengtes van de 7-jarige jongens, maar nu met verwachtingswaarde 0.

Voor elk individueel meetresultaat kunnen we dan schrijven:

$$y_i = \alpha \cdot x_i + \beta + \varepsilon_i,$$

waarbij ε_i de afwijking is die de lengte van de jongen vertoont t.o.v. de *y*-waarde die via de eerstegraadsfunctie berekend wordt. Dit is de formalisering van de gedachte die we in deze paragraaf

aangebracht hebben. Het eerste deel in deze formule (de eerste twee termen) drukt uit dat de *y*-waarden voor een deel afhangen van de *x*-waarden (meer bepaald volgens een eerstegraadsfunctie). Het tweede deel staat dan voor de andere factoren die de *y*-waarden beïnvloeden (en we veronderstellen dus dat deze invloeden elkaar globaal genomen uitmiddelen).

In de vorige paragrafen hebben we telkens gewerkt met gegevens waar met één *x*-waarde nooit meer dan één *y*-waarde overeenkwam. Als we dit bekijken in het licht van wat we in deze paragraaf geleerd hebben, zien we dat we voor elke *x*-waarde gewerkt hebben met een steekproef van 1 element. Zoals we in deze paragraaf uitgelegd hebben, kan de waarde van *y* opgesplitst worden in twee delen: een eerste deel, dat bepaald wordt door de *x*-waarde, en een tweede deel, dat door andere factoren bepaald wordt. Als de gegevens bijvoorbeeld metingen voorstellen, dan hebben we per *x*-waarde één meting gedaan. Als we de meting over zouden doen (of als iemand anders de meting zou uitvoeren), zou de invloed van de andere factoren er voor zorgen dat we een lichtjes afwijkende waarde zouden vinden.

De waarde van α en β schatten

Als we een lineaire regressie uitvoeren, dan is het onze bedoeling om de waarde van α en β te weten te komen voor de populatie. Omdat we in de praktijk nooit beschikken over de gegevens van de volledige populatie, zal het onmogelijk zijn om het gestelde doel te bereiken. We zullen ons daarom tevreden moeten stellen met een steekproef en op basis daarvan zullen we via het kleinste-kwadraten-criterium een schatting voor de waarde van α en β maken.

Als we voor de regressie gebruik maken van de steekproef van 60 jongens, vinden we de vergelijking $\hat{y} = 6{,}1203x + 82{,}4579$. We hebben dus 6,1203 als schatting voor α en 82,4579 als schatting voor β. Voor een 14-jarige geeft deze formule 168,1 cm. Dat is een schatting voor de verwachtingswaarde van de lengte van een 14-jarige jongen.

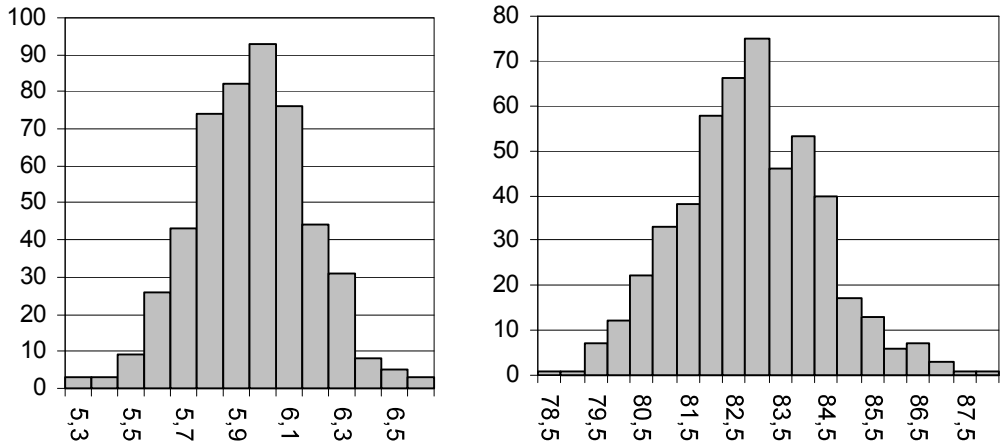
Als we een andere steekproef van 12 keer 5 jongens zouden nemen, dan zouden we een andere waarde als schatting voor α en β vinden. Je vindt de resultaten van 5 steekproeven die we genomen hebben in de onderstaande tabel.

steekproef	schatting voor α	schatting voor β	schatting voor de verwachtingswaarde van de lengte van een 14-jarige jongen
1	6,1203	82,4579	168,1
2	6,0975	81,1756	166,5
3	5,9390	82,5485	165,7
4	5,7506	83,9836	164,5
5	6,1541	81,5463	167,7

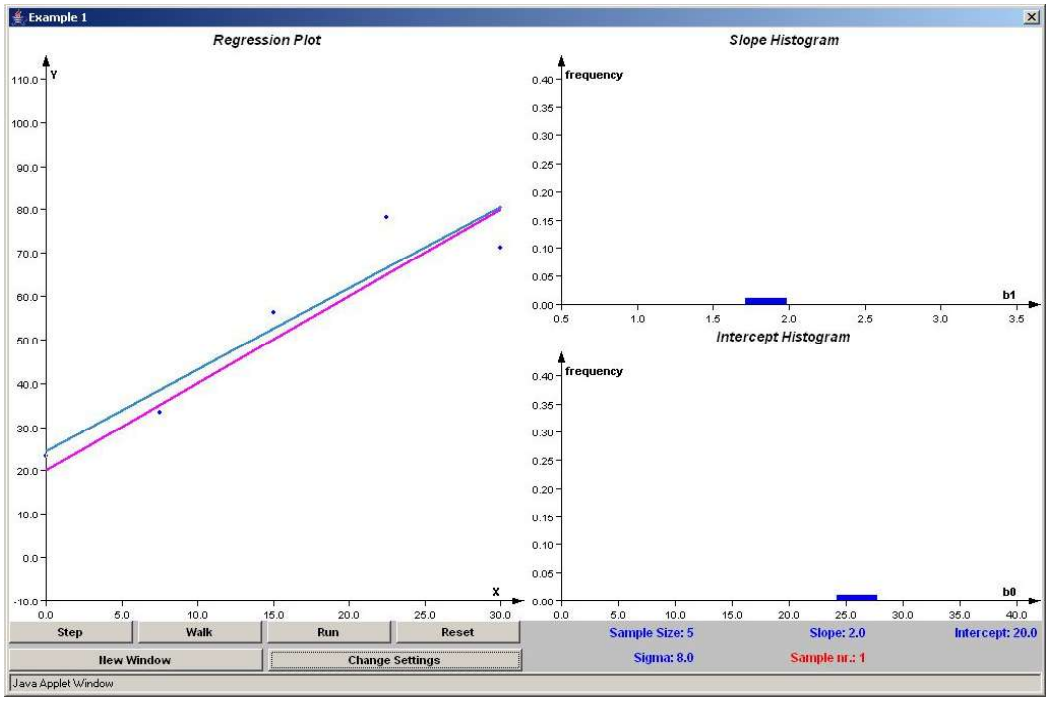
Elke steekproef levert dus een andere (schatting voor de) regressielijn en andere voorspellingen op!

De onderstaande grafieken tonen een grafische verwerking van de schattingen voor α en β voor 500 steekproeven.

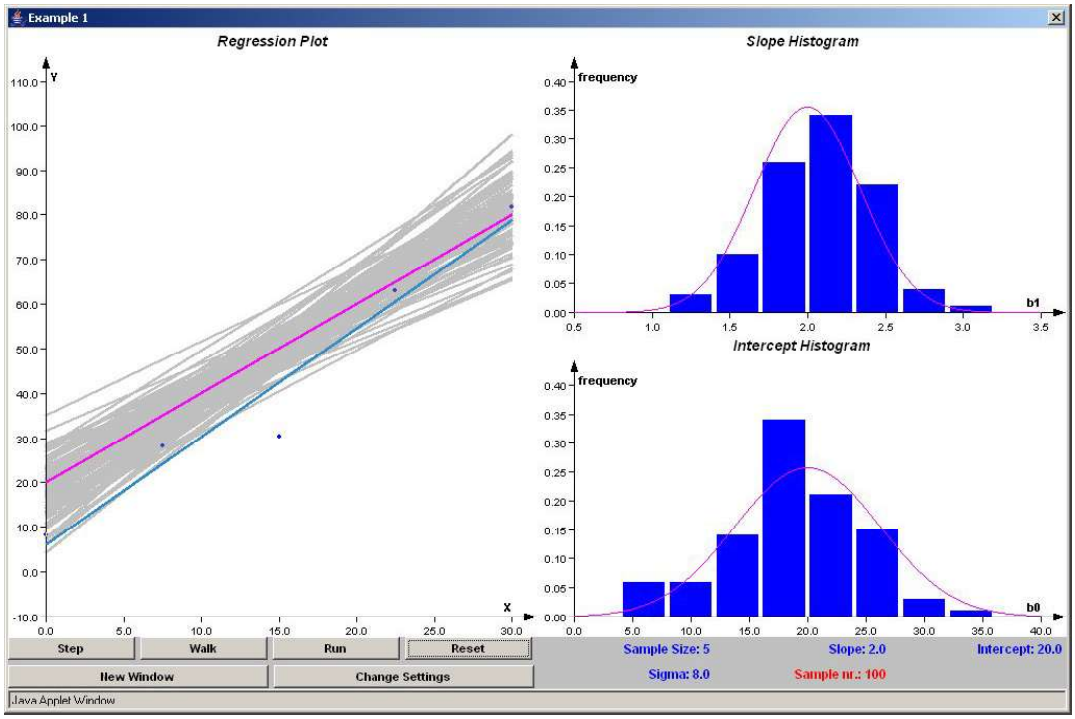
schattingen voor α **schattingen voorβ**



Op de website www.kuleuven.ac.be/ucs/java (volg Regressions en vervolgens Histograms of slope and intercept) vind je een applet die heel mooi illustreert wat we hierboven uitgelegd hebben.



De donkerste lijn (roze op de website) in het linkerdeel is de grafiek van de eerstegraadsfunctie die de verwachtingswaarden van de y-waarden in de gehele populatie uitdrukt in functie van de x-waarden. Als je op 'Step' drukt, wordt een steekproef gegenereerd (standaard met 5 elementen) en wordt de bijbehorende regressielijn getekend (de lichtere lijn in het linkerdeel, blauw op de website). Druk je op 'Walk', dan gebeurt dit een aantal keer opnieuw (standaard 100 keer). De applet maakt de positie van de 100 regressielijnen op een mooie manier zichtbaar en houdt in een histogram de 100 schattingen voor α en β bij. Tot slot wordt een normale dichtheid getekend op de beide histogrammen (zie ook verder). Met 'Run' krijg je hetzelfde, alleen veel sneller.



Via 'Change Settings' kan je de instellingen veranderen: de waarde van α en β, de standaardafwijking van de normale verdelingen van de *y*-waarden, de grootte van de steekproef, het aantal steekproeven ... Het zou een mooie opdracht voor leerlingen kunnen zijn om te onderzoeken welke factoren de nauwkeurigheid van de schattingen beïnvloeden.

Regressielijn en verklarende statistiek

Omdat de regressielijn die je vindt, afhangt van de steekproef, moet je het resultaat met de nodige omzichtigheid hanteren. Een andere steekproef zou immers wel eens een heel ander resultaat kunnen opleveren. Bijvoorbeeld: een positieve helling die dicht bij 0 ligt, zou kunnen veranderen in een negatieve helling. Bij een echte regressie wordt daarom met betrouwbaarheidsintervallen gewerkt: een betrouwbaarheidsinterval voor de helling α, voor het intercept β, voor de voorspelling van een verwachtingswaarde en van een individuele lengte ... Ook worden hypothesen getest, bijvoorbeeld: is de helling significant verschillend van 0?

In deze paragraaf (en in de applet) zijn wij eerder experimenteel te werk gegaan. We hebben de variabiliteit in de schattingen voor α en β onderzocht door heel veel steekproeven te nemen. In werkelijkheid leidt men op een theoretische manier af hoe de schattingen voor α en β verdeeld zijn. Men vertrekt hierbij van bepaalde veronderstellingen over de stochastische veranderlijken Y_x die de verdeling van de *y*-waarden bij een zekere *x*-waarde beschrijven. Men veronderstelt bijvoorbeeld

- dat de stochastische veranderlijken Y_x allemaal normaal verdeeld zijn (zoals in ons voorbeeld)
- dat er getallen α en β bestaan zo dat de verwachtingswaarde van Y_x gelijk is aan $\alpha x + \beta$ voor alle x
- dat de standaardafwijking bij al deze normale verdelingen steeds dezelfde is (wat in ons voorbeeld niet het geval was)

- dat de stochastische veranderlijken Y_{x_1} en Y_{x_2} bij verschillende x-waarden x_1 en x_2 onafhankelijk zijn
- ...

In de standaardsituatie die men bestudeert, kan men dan bewijzen dat de schattingen die via het kleinste-kwadraten-criterium bepaald worden normaal verdeeld zijn (vandaar de normale dichtheden die getekend worden in de applet). De verwachtingswaarden voor deze schattingen zijn α, respectievelijk β. Men kan ook de standaardafwijking van deze schattingen berekenen. Op basis hiervan kan men dan formules opstellen voor betrouwbaarheidsintervallen en voor het toetsen van hypothesen.

Bibliografie

[1] H. Callaert, *Statistische Intelligentie – module 4, De samenhang ontdekken: exploratie van bivariaat cijfermateriaal, Deel 1a. Correlatie*, Handleiding voor leerkrachten, zie: www.scholennetwerk.be

[2] G. Delaleeuw, *Begeleid zelfstandig leren en werken in de derde graad met de TI-83 (84) Plus*, op het T^3-symposium 2004, zie: www.t3vlaanderen.be

[3] P. Holmes, *Correlation: From Picture to Formula*, Teaching Statistics, Volume 23, Number 3, Autumn 2001

[4] T. N. Lam, *Estimating fuel consumption from engine size*, Journal of Transportation Engineering, 111 (1985), pp. 339-357

[5] J. Roels e.a., *Wiskunde vanuit toepassingen*, Leuven (Aggregatie HSO Wiskunde – K.U. Leuven), 1990

[6] A. J. Rossman, *Television, Physicians and Life Expectancy*, Journal of Statistics Education v.2 n.2, 1994, zie: www.amstat.org/publications/jse

[7] B. Van Deyck, *Een eerste kennismaking met regressie,* Leerlingentekst bij licentiaatsverhandeling, K.U.Leuven

[8] K. L. Weldon, *A Simplified Introduction to Correlation and Regression*, *Journal of Statistics Education* v.8 n.3, 2000, zie: www.amstat.org/publications/jse

[9] *The World Almanac and Book of Facts*, Pharos Books (New York), 1993

Websites waarnaar verwezen werd:

[10] http://standards.nctm.org/document/eexamples/chap7/7.4/index.htm

[11] http://www.ruf.rice.edu/~lane/stat_sim/reg_by_eye/index.html

We verwijzen tot slot nog naar een aantal boeken over statistiek die deze loep mee geïnspireerd hebben.

[12] F. Daly, D. J. Hand, M. C. Jones, A. D. Lunn, K. J. McConway, *Elements of statistics*, Addison-Wesley, 1995, ISBN 0-201-42278-6

[13] D. S. Moore, G. P. McCabe, *Statistiek in de praktijk*, 3de herziene druk, Academic Service, 2001, ISBN 90-395-1420-8

[14] A. J. Rossman, B. L. Chance, *Workshop Statistics*, 2nd edition, Key College Publishing, 2001, ISBN 1-93 0190-03-4

Johan, Jan, Pedro