STATISTIEK VOOR HET SECUNDAIR ONDERWIJS

Regressie: exploratieve methoden

Werktekst voor de leerling

Prof. dr. Herman Callaert

Hans Bekaert
Cecile Goethals
Lies Provoost
Marc Vancaudenberg

Inhoudstafel

DEEL 1. De ideeën achter de formules

1. Even opfrissen

Regressie is een veel gebruikte statistische techniek die kadert in de studie van het verband tussen twee veranderlijken. Het is belangrijk dat je regressie een juiste plaats geeft binnen het grotere kader van statistische methoden. Daarom is het goed om vooraf enkele teksten (die je misschien vroeger al bestudeerd hebt) te bekijken.

Bij studies naar samenhang tussen twee veranderlijken moet je vooraf vastleggen welke veranderlijke de **respons** is en welke veranderlijke je als **verklarende veranderlijke** behandelt. Soms kan je de respons en de verklarende veranderlijke eenvoudig bepalen uit de aard van het vraagstuk, zoals bij een onderzoek naar de samenhang tussen studieresultaten (respons) en het geslacht (verklarende veranderlijke). In andere gevallen, zoals bij de samenhang tussen de lengte van vaders en de lengte van zonen, kan je als respons zowel de lengte van de vaders als de lengte van de zonen nemen. Welke keuze je hierbij maakt is bepalend voor de conclusie die je uiteindelijk zal kunnen trekken. Let daar dus goed op bij de start van je onderzoek.

Welke statistische technieken je bij een onderzoek moet gebruiken hangt niet alleen af van de onderzoeksvraag maar ook van het type veranderlijke waarmee je te maken hebt. Een eenvoudige indeling is als volgt: ofwel is een veranderlijke numeriek continu (dikwijls behandel je een numeriek discrete veranderlijke met veel verschillende uitkomsten ook met continue methoden) ofwel komen haar waarden terecht in een beperkt aantal categorieën (nominaal, ordinaal of numeriek discreet met een beperkt aantal verschillende uitkomsten). Het eerste type noem je **continu**, het tweede **categorisch**. Zowel de respons als de verklarende veranderlijke kan continu of categorisch zijn. Dit leidt tot 4 verschillende combinaties die elk een eigen statistische aanpak vereisen. Als zowel de respons als de verklarende veranderlijke continu zijn, dan kan je gebruik maken van regressie.

Wanneer je een samenhang tussen twee veranderlijken ontdekt dan betekent dat niet noodzakelijk dat je te maken hebt met **oorzaak en gevolg**. In de meeste studies gaat het gewoon over dingen die samen optreden en waarbij je enkel kan zeggen dat je een **associatie** gezien hebt.

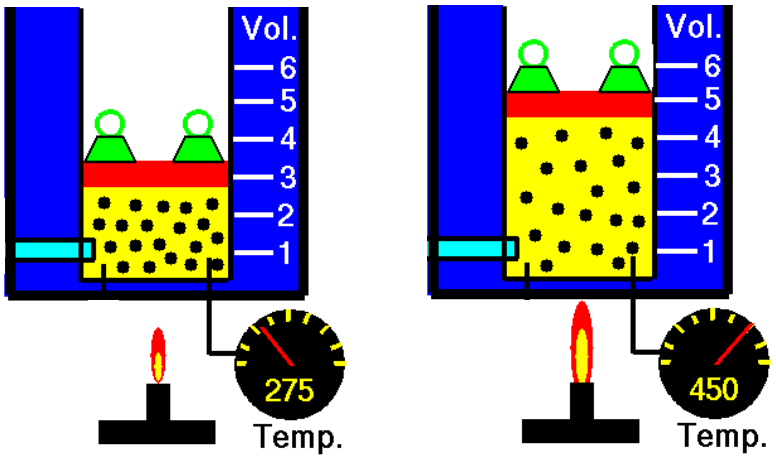
Meer informatie over de begrippen respons, verklarende veranderlijke, categorisch, continu, oorzakelijk verband en associatie kan je vinden in de studie over “Geboortegewicht en zwangerschapsduur” in de tekst “Studies naar samenhang: basisbegrippen” die je kan vinden op http://www.uhasselt.be/lesmateriaal-statistiek .

Een studie naar samenhang tussen twee continue veranderlijken begint bij het tekenen van een puntenwolk. Om de kernideeën van regressie goed te begrijpen bestudeer je eerst puntenwolken waarbij je globaal de indruk hebt dat de punten willekeurig verstrooid liggen rond een rechte (en niet rond een ander patroon zoals een parabool). Het **lineaire verband** dat je zo ontdekt kan je karakteriseren als **positief** of **negatief**, **sterk**, **matig** of **zwak**. Je kan die lineaire samenhang ook bestuderen met een getal: dat is de **correlatiecoëfficiënt**.
Meer informatie hierover vind je in de afzonderlijke tekst “Correlatie. Achtergrondinformatie” op http://www.uhasselt.be/lesmateriaal-statistiek .

2. Het verdwenen volume

2.1. Volume en temperatuur

De volumewet van Gay-Lussac zegt dat het volume van een (ideaal) gas recht evenredig is met de temperatuur wanneer je de druk en de massa constant houdt. Op bijgaande figuur zie je een illustratie van deze gaswet waarbij een vaste massa gas gevangen zit in een container waarop een vaste druk wordt uitgeoefend. Voor dit labexperiment werd een volume van 280 milliliter opgetekend bij een temperatuur van 275 Kelvin en vergrootte het volume tot 460 ml wanneer de temperatuur gelijk was aan 450 K.



Een analoog experiment werd uitgevoerd door een klas van 14 leerlingen die allemaal dezelfde opstelling gebruikten. Iedereen voerde de proef één keer uit en noteerde het volume (in ml) en de temperatuur (in K) met volgend resultaat:

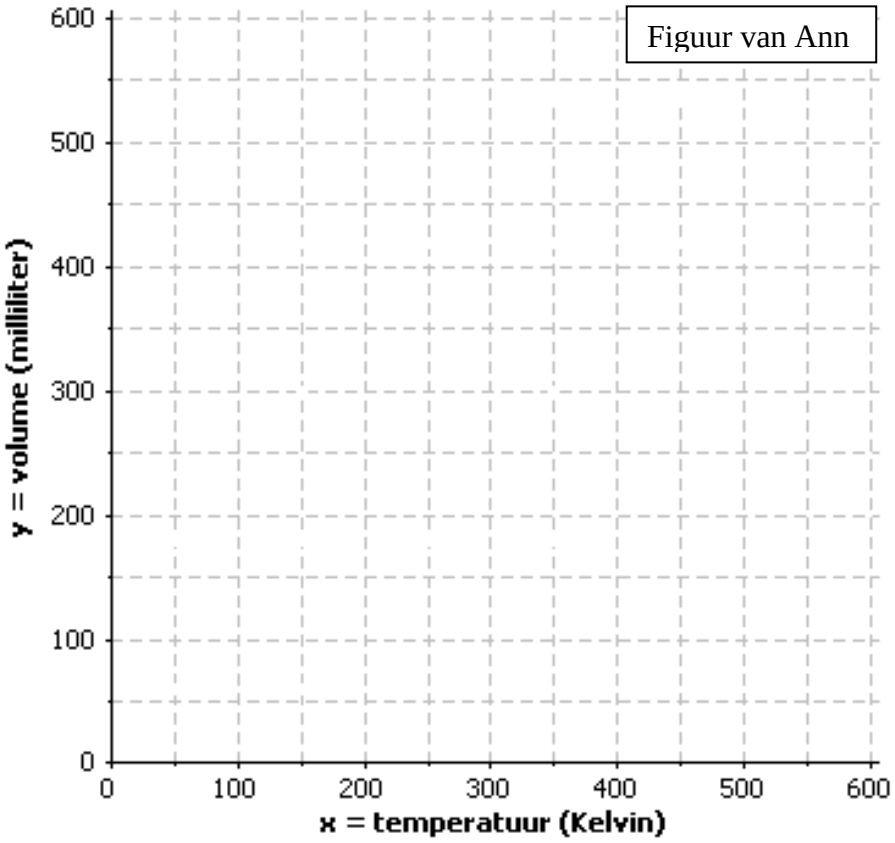
Temp (K)	Vol (ml)	Temp (K)	Vol (ml)	Temp (K)	Vol (ml)	Temp (K)	Vol (ml)	Temp (K)	Vol (ml)
250	180	450	420	50	60	350	420	250	420
150	180	150	60	150	300	350	540	50	180
350	300	350	180	450	540	150	420		

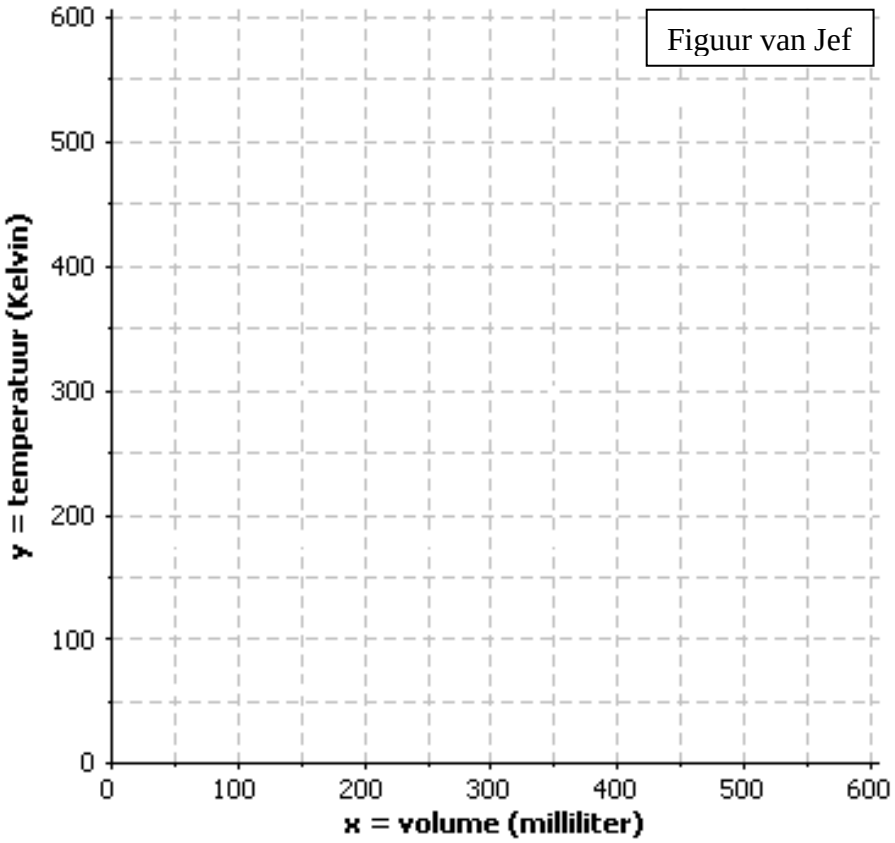
Deze opmetingen kan je vinden op http://www.uhasselt.be/lesmateriaal-statistiek waar je de bestanden TMP1.8xl en VOL1.8xl kan downloaden. Breng deze bestanden over naar je GRM als lijsten TMP1 en VOL1.

Op de volgende bladzijde staan twee figuren (figuur van Ann en figuur van Jef) die jij moet aanvullen in de opdrachten die je verder in dit deel ontmoet. Werk ordelijk zodat je dezelfde figuur voor meerdere opdrachten kan gebruiken.

Opdracht 1

Ann en Jef gaan de opmetingen van hun 14 medeleerlingen analyseren. Zij bemerken dat het hier gaat over bivariate gegevens omdat er per proef twee dingen tegelijk werden opgemeten: de temperatuur en het volume. Zowel temperatuur als volume kan je behandelen als continue numerieke veranderlijken.
Het eerste wat Ann en Jef doen is een figuur tekenen om de data grafisch voor te stellen. Ann besluit om de temperatuur uit te zetten op de x-as en het volume op de y-as. Jef doet het juist omgekeerd. Hij kiest ervoor om, per uitgevoerde proef, het volume uit te zetten op de x-as en de bijhorende temperatuur op de y-as. Teken jij nu ook de gepaste figuur, zowel voor Ann als voor Jef. Hoe noem je de figuur die je tekent bij bivariate continue gegevens?


Figuur van Ann
y = volume (milliliter)
x = temperatuur (Kelvin)
0
100
200
300
400
500
600


Figuur van Jef
y = temperatuur (Kelvin)
x = volume (milliliter)
0
100
200
300
400
500
600

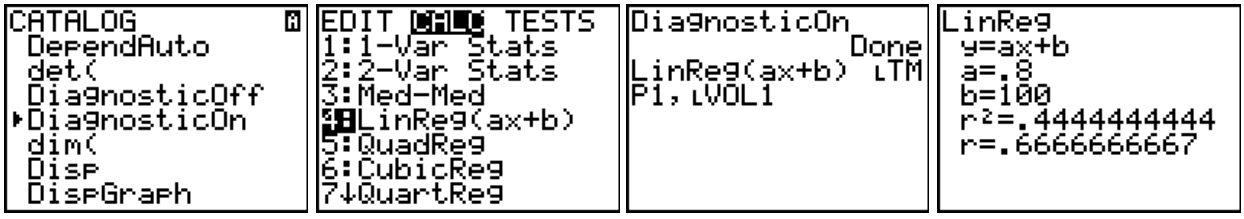
2.2. Zoek het verschil

Ann en Jef bestuderen de samenhang tussen temperatuur en volume. Zij zijn ervan overtuigd dat het daarbij geen verschil maakt welke grootheid je de naam x geeft en welke de naam y. Uiteindelijk gaat het toch gewoon over samenhang tussen twee dingen.

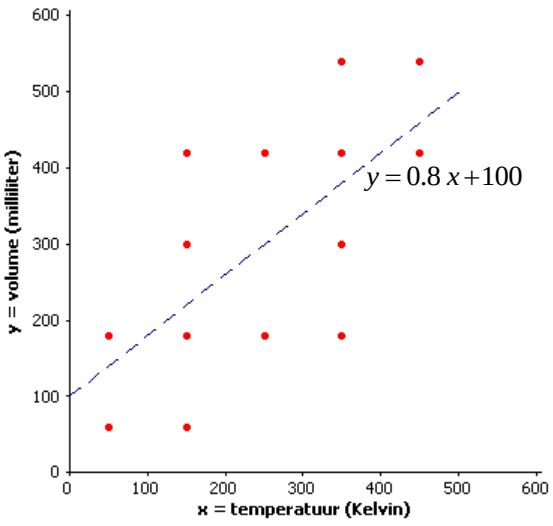
Er zijn oneindig veel manieren om een verband tussen twee veranderlijken te beschrijven maar hier ga je op zoek naar een **lineair** verband. Dat er zo'n verband is verwacht je wanneer de punten van de puntenwolk de indruk geven dat zij willekeurig verspreid liggen rond een rechte.

Ann en Jef hebben ooit gelezen dat de regressierechte de rechte is die het beste aansluit bij zo'n puntenwolk. Bovendien weten zij dat je een regressierechte y = a x + b rechtstreeks uit je GRM kan halen. Zij gaan dus op zoek naar die regressierechte om het lineair verband tussen hun twee veranderlijken te beschrijven. Zij blijven daarbij op hun standpunt wat betreft de naam (x of y) die zij kiezen voor de temperatuur en het volume.

Ann gaat als volgt te werk.



Om ook de correlatiecoëfficiënt *r* zichtbaar te maken tikt zij vooraf [2nd] [CATALOG], drukt op de groene letter D en loopt verder naar beneden tot aan DiagnosticOn en drukt tweemaal [ENTER]. Dan drukt zij [STAT], loopt naar CALC, drukt 4:LinReg(ax+b) en vervolledigt het commando zoals aangegeven. Dat gebeurt als volgt. Druk [2nd] [LIST], loop naar beneden tot je naast TMP1 staat en druk [ENTER]. Druk [,] en dan terug [2nd] [LIST], loop naar beneden tot je naast VOL1 staat en druk twee keer [ENTER]. Het commando dat op die manier is ingebracht, berekent een regressierechte y=ax+b op basis van een puntenwolk (x_i, y_i) waarvan de x-coördinaten in TMP1 staan (de x-veranderlijke is de temperatuur) en de y-coördinaten in VOL1 (de y-veranderlijke is het volume).



Ann heeft de regressierechte y = 0.8 x + 100 gevonden waarbij y verwijst naar het volume en x naar de temperatuur. Zij tekent deze regressierechte op haar figuur.

Opdracht 2

Teken op jouw figuur van Ann de gevonden regressierechte. Zoek daarna de regressierechte van Jef en teken die rechte op de figuur van Jef. Je kan hierbij de methode die Ann gebruikt heeft nabootsen. Noteer de coëfficiënten a en b tot op 3 decimalen.

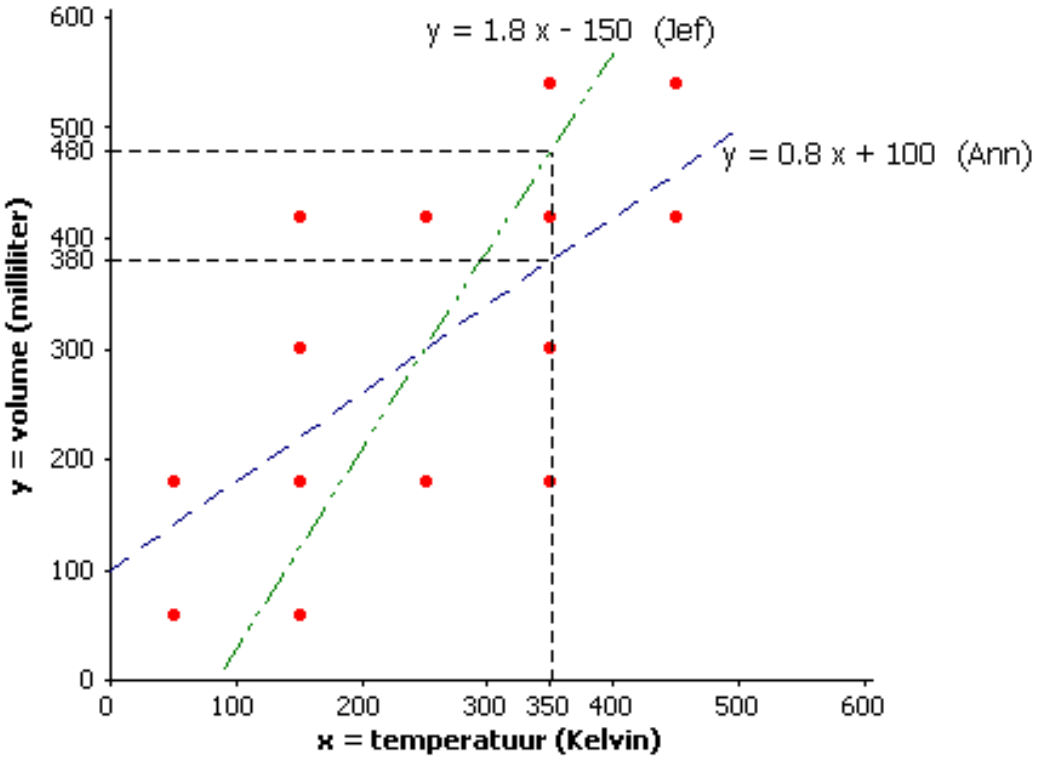
De regressierechte van Jef verschilt van die van Ann. Maar is dat een echt of een schijnbaar verschil? Om dat te weten te komen schrijf je de eenheden in "woorden" (om zeker geen verwarring tussen x en y te hebben).

Voor Ann wordt dit: **volume = 0.8 temperatuur + 100**.

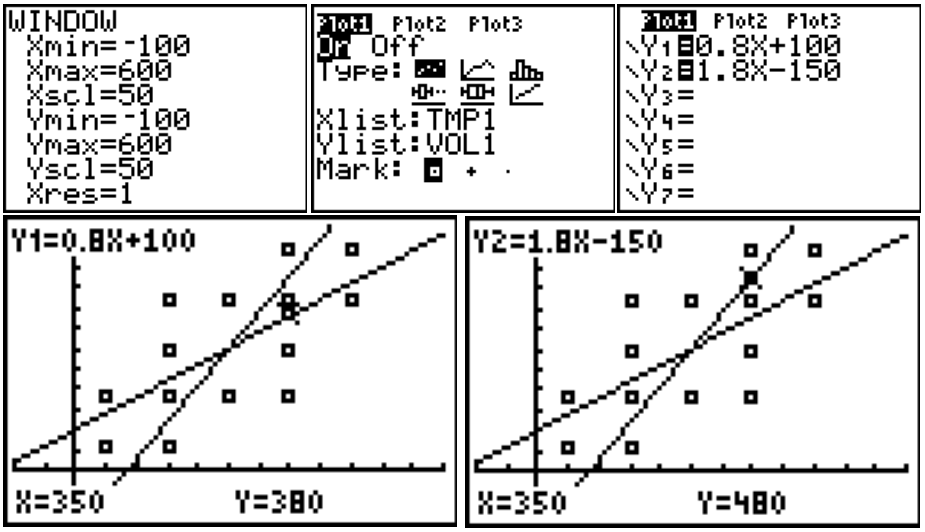
Voor Jef krijg je: temperatuur = 0.556 volume + 83.333. Dit kan je als volgt herschrijven: $\frac{temperatuur - 83.333}{0.556} = volume$. Aangezien $\frac{1}{0.556} \cong 1.8$ en $\frac{83.333}{0.556} \cong 150$ wordt de regressierechte van Jef ook geschreven als **volume = 1.8 temperatuur – 150**.

Opdracht 3

Teken de regressierechte van Jef **volume = 1.8 temperatuur – 150** op de figuur van Ann (waarbij temperatuur op de x-as en volume op de y-as staat). Duid grafisch aan welk volume Ann vindt bij een temperatuur van 350 K en welk volume dat voor Jef is. Bereken ook deze volumes uit de vergelijkingen. Vinden Ann en Jef hetzelfde volume bij eenzelfde temperatuur? Hoe verklaar je dit?



Je kan de regressierechten samen met de puntenwolk zichtbaar maken met je GRM. Doe dat volgens de schermafdrukken die je hier ziet.



Als je voor de opstelling in het labo de massa gas en de druk niet wijzigt en je zet de temperatuur op 350 K dan krijg je:

een volume van 380 ml volgens Ann
een volume van 480 ml volgens Jef.

Ann en Jef hebben voor hun berekeningen dezelfde experimentele data gebruikt. En toch is hier een volume van 100 ml "verdwenen". Waar is dat naartoe?

Besluit

Wanneer een samenhang tussen twee continue veranderlijken kan voorgesteld worden door een puntenwolk die rond een rechte verspreid is, dan is het zinvol om op zoek te gaan naar een lineaire samenhang tussen deze veranderlijken. In een statistische context is het daarbij nodig om rekening te houden met de variabiliteit van de opmetingen. Je werkt dan met een model waarbij de rol van de respons (y) en van de verklarende veranderlijke (x) helemaal niet symmetrisch is. De gevonden regressierechte mag niet op de klassieke manier als een wiskundige vergelijking geïnterpreteerd worden. Statistische regressie heeft een eigen manier van werken en een eigen interpretatie. Hoe dit juist in elkaar zit wordt verder in deze tekst uitgelegd.

3. Bouw je model

3.1. Vaders en zonen

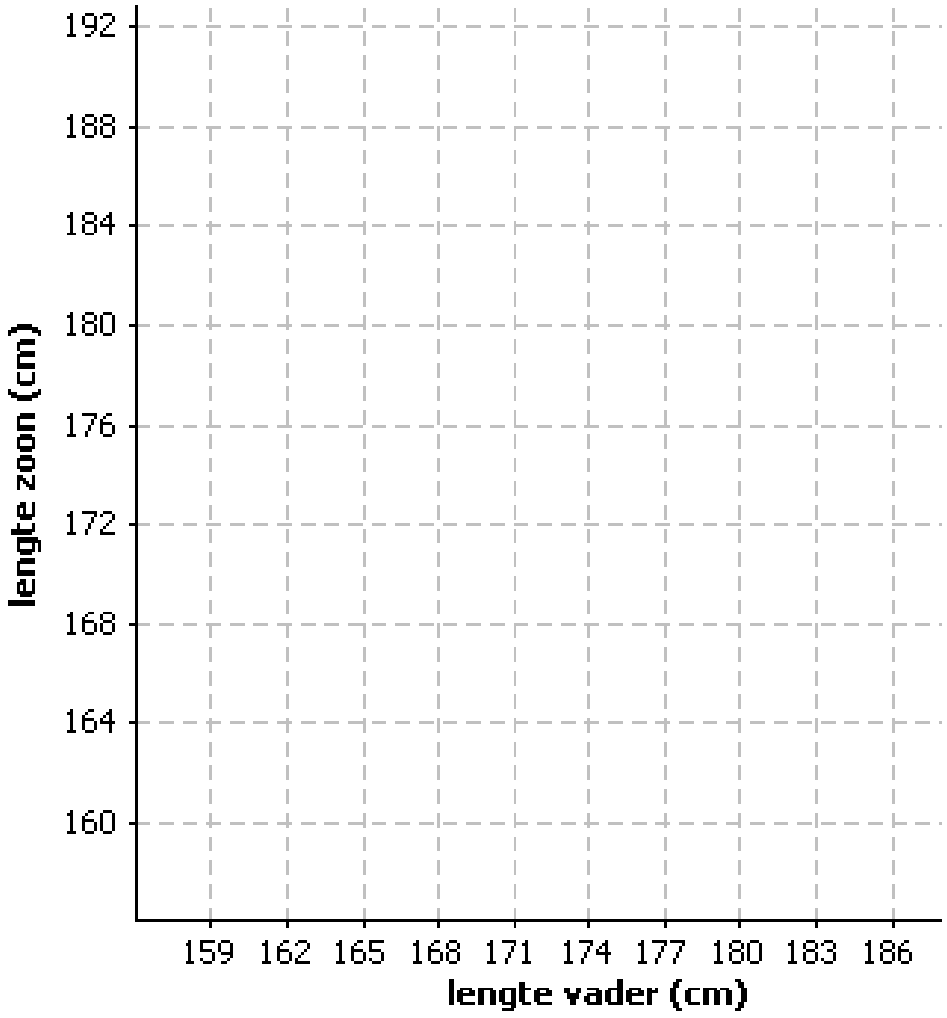
Een klassiek voorbeeld in de regressie gaat over het verband tussen de lengte van een vader en van zijn oudste volwassen zoon. We beginnen met zo'n voorbeeld. Het cijfermateriaal toont een steekproef van 29 gezinnen met volgend resultaat (de lengte is afgerond tot op de cm):

Volgnr. gezin	Lengte vader	Lengte zoon	Volgnr. gezin	Lengte vader	Lengte zoon	Volgnr. gezin	Lengte vader	Lengte zoon
1	168	188	11	177	184	21	165	168
2	162	172	12	168	180	22	180	180
3	177	176	13	180	172	23	168	172
4	174	164	14	183	184	24	165	184
5	180	188	15	174	172	25	162	164
6	159	168	16	174	180	26	162	180
7	174	188	17	159	160	27	171	184
8	165	176	18	183	192	28	168	164
9	171	168	19	171	176	29	177	192
10	165	160	20	177	168			

Deze opmetingen kan je vinden op http://www.uhasselt.be/lesmateriaal-statistiek waar je de bestanden VADER.8xl en ZOON.8xl kan downloaden. Breng deze bestanden over naar je GRM als lijsten VADER en ZOON.

Opdracht 4

Zoals gewoonlijk begin je met een dataset grafisch voor te stellen. Doe dat op de figuur hieronder waar de lengte van de vader als x-veranderlijke is gekozen en de lengte van de zoon als y-veranderlijke. Geeft de puntenwolk de indruk dat de punten verstrooid liggen rond een rechte? Is het hier zinvol om het verband voor te stellen met een rechte? Is dat verband positief of negatief en is het zwak, matig of sterk? Motiveer je antwoord.



3.2. Verklarende veranderlijke en respons

In de studie van de volumewet van Gay-Lussac heb je experimenteel vastgesteld dat de regressierechte van Jef verschilt van die van Ann. De reden hiervoor is dat Jef en Ann een verschillende keuze maakten voor hun verklarende veranderlijke (de x-veranderlijke) en voor hun respons (de y-veranderlijke).

In regressie spelen de verklarende veranderlijke en de respons een eigen rol die je niet zomaar mag omdraaien. Dat heeft te maken met de bedoeling van je studie.

Als je, op basis van de temperatuur een uitspraak wil doen over het bijhorende volume, dan neem je de temperatuur als verklarende veranderlijke en het volume als respons. Als je omgekeerd, op basis van het volume iets wil zeggen over de bijhorende temperatuur, dan neem je volume als verklarende veranderlijke en temperatuur als respons.

De keuze die je maakt voor de verklarende veranderlijke en voor de respons bepaalt ook de verdere manier van werken. Je behandelt de verklarende veranderlijke (de x-veranderlijke) als een nauwkeurig opgemeten grootheid die je (zo goed als) exact kent. De bijhorende respons (de y-veranderlijke) behandel je als een resultaat dat aan het toeval onderhevig is. Als je een volgende keer terug een opmeting doet bij exact dezelfde x-waarde dan verwacht je een bijhorende y-waarde die een beetje verschilt van de vorige.

In een labopstelling is het denkbaar dat je bijvoorbeeld heel nauwkeurig de gewichten kent die je aan een veer hangt maar dat je voor de lengte van de veer niet altijd hetzelfde resultaat vindt bij eenzelfde gewicht. Als je dan op basis van het gewicht een uitspraak wil doen over de lengte dan kies je het gewicht als verklarende veranderlijke en de lengte als respons. De boven beschreven werkwijze (de x-veranderlijke ken je exact en op de y-veranderlijke zitten meetfouten) lijkt hier logisch.

Maar ook in andere situaties blijft men bij regressie de verklarende veranderlijke behandelen als “exact gekend” en de respons als “aan het toeval onderhevig”. Bij de lengte van vaders en zonen zijn beide veranderlijken “toevallige” uitkomsten. Als je uit de populatie van alle gezinnen er lukraak één trekt, dan heb je een toevallig resultaat uit alle mogelijke koppels (x_i , y_i) met x_i = lengte vader en y_i = lengte zoon. De bedoeling van je studie is nu van cruciaal belang en bepaalt hoe je verder werkt. Als je op basis van de lengte van de vaders een uitspraak wil doen over de lengte van de zonen, dan kies je de lengte van de vader als verklarende veranderlijke en de lengte van de zoon als respons. Je behandelt dan de lengte van de vaders als exact opgemeten getallen en de lengte van de zonen als toevallige uitkomsten.

Om duidelijk het verschil te zien tussen de verklarende veranderlijke en de respons gebruik je in regressie een speciale terminologie waarbij de volgorde van de woorden belangrijk is. De algemene uitdrukking ziet er als volgt uit:

regressie van de respons **over** de verklarende veranderlijke

Nota.

- Vanaf nu zetten we de verklarende veranderlijke uit op de x-as en de respons op de y-as.
- In andere teksten kom je voor “verklarende veranderlijke” soms de woorden “regressor” of “covariabele” of “predictor” of “onafhankelijke veranderlijke” tegen. De “respons” wordt soms ook de “afhankelijke veranderlijke” genoemd.

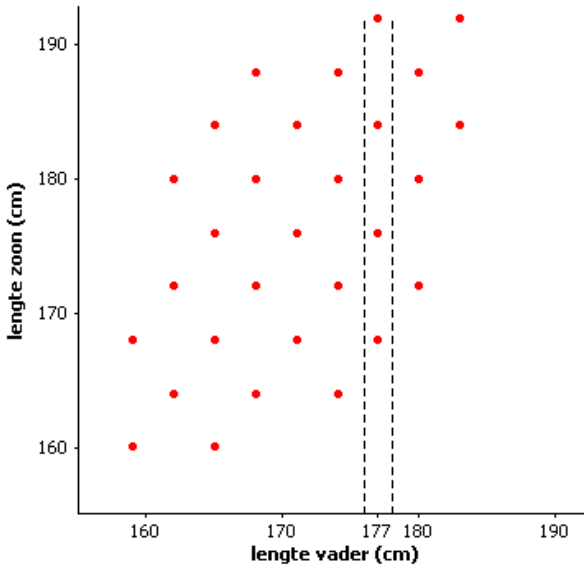
Opdracht 5

In een landbouwschool heeft men, bij een aantal percelen, genoteerd hoeveel meststof er werd gestrooid en wat de opbrengst was. Als men in deze context spreekt van "regressie van opbrengst over meststof", wat is dan de bedoeling van die studie? Wat is de verklarende veranderlijke en wat is de respons?

3.3. Gemiddelde respons

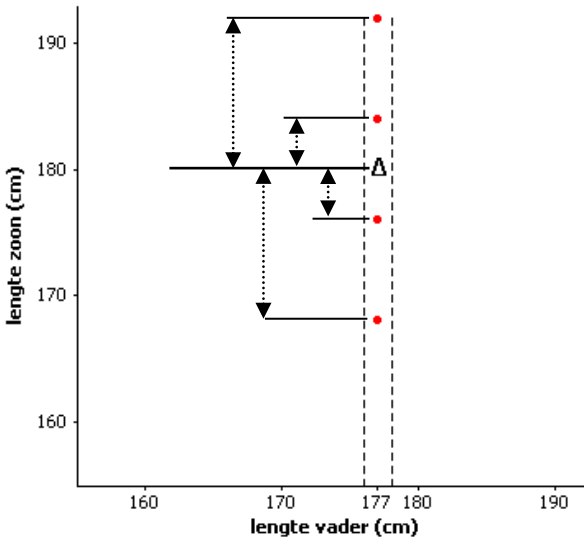
Nadat je besloten hebt welke grootheid je als verklarende veranderlijke neemt en welke als respons moet je bepalen hoe je het verband tussen die twee grootheden beschrijft. De verklarende veranderlijke behandel je als "exact" maar op de respons zit variabiliteit. Wat doe je daar mee?

Als voorbeeld vervolgen we onze studie waarbij we ervoor kiezen om op basis van de lengte van vaders iets te zeggen over de lengte van hun zonen. We doen dus een regressie van "lengte zoon" over "lengte vader".



Vaders die 177 cm groot zijn, hebben zonen die niet allemaal even groot zijn. Dat is nu eenmaal zo. Op de figuur zie je in de verticale strip punten (x_i, y_i) waarbij telkens $x_i = 177$ (lengte vaders) maar waarbij de y_i 's verschillende waarden hebben (lengte zonen).

Welke uitspraak ga je nu doen over de lengte van zonen waarvan de vader een lengte van 177 cm heeft? Een eenvoudig kengetal om globaal de lengte van al die zonen te beschrijven is het gemiddelde. In dit voorbeeld heb je (in de verticale strip) 4 zonen met lengte (in cm) 168 ; 176 ; 184 en 192. De gemiddelde lengte is hier 180 cm.



Misschien heb je vroeger gezien dat het gemiddelde het getal is dat de som van de kwadratische afstanden minimaal maakt. Als je hier zou zoeken voor welk getal c de som

$$(168-c)^2+(176-c)^2+(184-c)^2+(192-c)^2$$

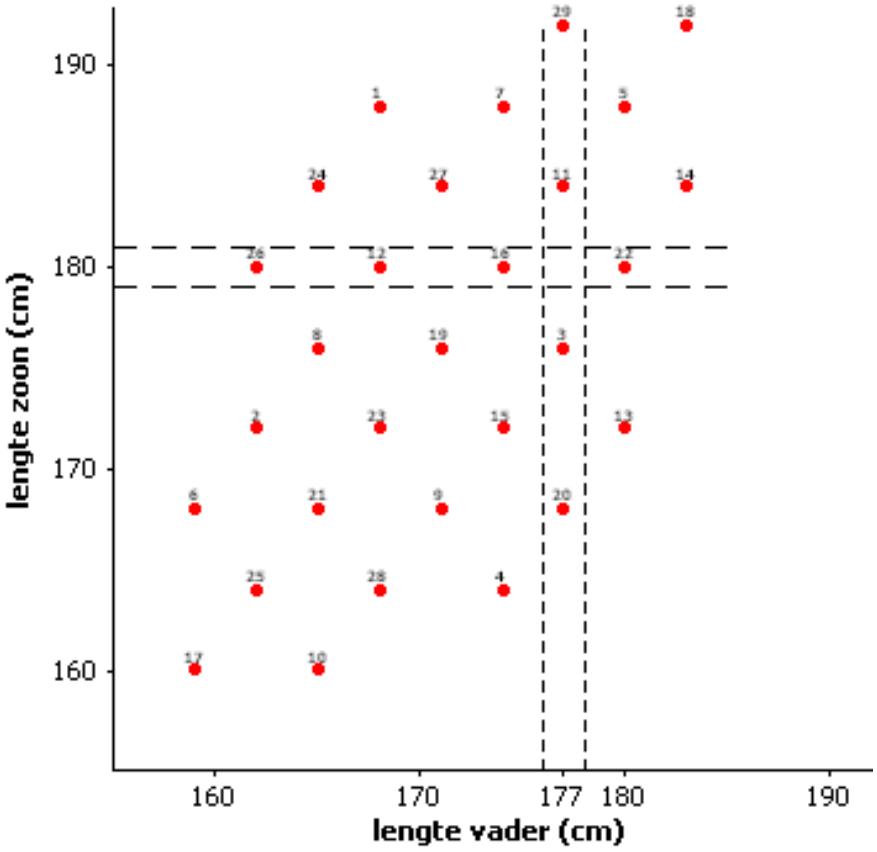
zo klein mogelijk is, dan is dat voor $c=\bar{y}=\frac{1}{4}\sum_{i=1}^{4} y_i = 180$. Op de figuur kijk je binnen de verticale strip naar **verticale afstanden** tussen meetpunten en hun gemiddelde [dat gemiddelde is op de figuur de y-coördinaat van het punt $\Delta = (177, \bar{y}) = (177, 180)$].

3.4. Niet omdraaien a.u.b.

In puntje 3.2 las je:" In regressie spelen de verklarende veranderlijke en de respons een eigen rol, die je niet zomaar kan omdraaien". Dat had te maken met de bedoeling van de studie. Inderdaad, maar hier zie je dat het ook te maken heeft met de manier van werken: bij een vaste waarde van de verklarende veranderlijke zoek je een bijhorende gemiddelde waarde van de respons.

Als je zou zeggen dat bij vaders van 177 cm zonen van 180 cm horen dan zou je daaruit waarschijnlijk ook besluiten dat bij zonen van 180 cm vaders van 177 cm horen. Maar het gaat hier niet over een verband tussen een lengte en een lengte, maar tussen een lengte en een gemiddelde lengte.

Als je een vaste lengte van vaders neemt (177 cm) dan zoek je een bijhorende gemiddelde lengte van zonen (180 cm). Die vind je in de verticale strip. Als je nu omgekeerd 180 cm als vaste lengte van zonen neemt, dan moet je in de horizontale strip zoeken naar de gemiddelde lengte van vaders. Die is daar 171 cm.



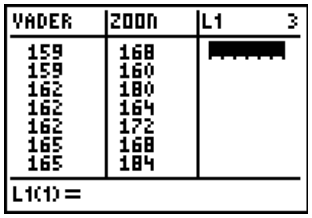
Je kan een gevonden relatie tussen "lengte" en "gemiddelde lengte" niet zomaar omdraaien. Dat zie je goed op de figuur. Elk gezin uit de steekproef is nu aangeduid met zijn volgnummer. Bij vaders van 177 cm horen zonen waarvan de gemiddelde lengte 180 cm is. Dat wordt bepaald door de gezinnen met volgnummer 3, 11, 20 en 29. Bij zonen van 180 cm horen vaders waarvan de gemiddelde lengte 171 cm is. Dat vind je uit de gezinnen met volgnummer 12, 16, 22 en 26. Er is hier geen symmetrie. Je haalt je informatie uit verschillende groepen gezinnen.

Opdracht 6

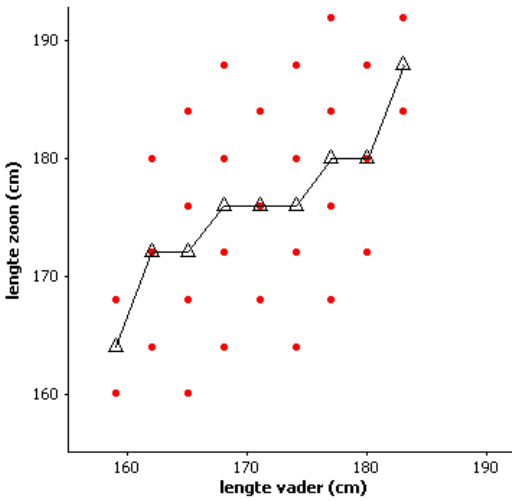
Voor elke verticale strip (elke vaste waarde van de verklarende veranderlijke) kan je het gemiddelde berekenen van de y-coördinaten van de punten in die strip (het gemiddelde van de bijhorende responsen). Doe dat nu en teken die gemiddelden op je figuur (gebruik een andere kleur of een ander symbool, bijvoorbeeld Δ).

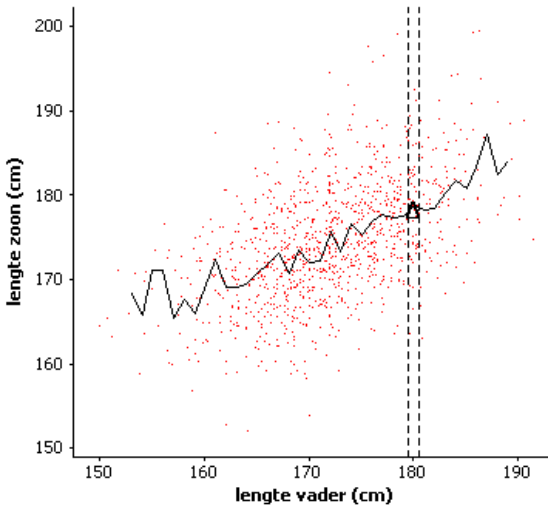
Tip: sorteer eerst op de lengte van de vader. Druk [STAT] en 1:Edit… en ga op de kop van lijst [L1] staan en druk [2nd][INS] en dan [2nd][LIST] en loop naar ZOON. Druk nu 2 keer [ENTER]. Blijf op de kop van lijst ZOON staan en druk [2nd][INS] en dan [2nd][LIST] en loop naar VADER. Druk nu 2 keer [ENTER] en dan [2nd][QUIT]. In het gewone scherm druk je nu [2nd][LIST], je loopt naar

OPS en drukt dan 1:SortA(. Vervolledig dit commando met [2nd][LIST] en loop naar VADER en [ENTER]. Druk dan de komma [,] en [2nd][LIST] en loop naar ZOON en druk 2 keer [ENTER]. Druk nu [STAT] en 1:Edit… . Je ziet nu alle vaders gesorteerd volgens lengte samen met de lengte van hun zonen.



3.5. De lijn der gemiddelden





Op de linkerfiguur zijn de gemiddelden, die je zopas in elke verticale strip hebt berekend, verbonden tot een "lijn der gemiddelden". Op de rechterfiguur zie je ook zo'n "lijn der gemiddelden" voor een grote dataset van 1078 gezinnen. Als je meer en meer opmetingen zou maken (en uiteindelijk zou werken met een model voor een oneindige populatie) dan zou je hier vinden dat "de lijn der gemiddelden" een rechte is. Die rechte geeft voor elke lengte van vaders (voor elke x-waarde) de gemiddelde lengte van de bijhorende zonen (het gemiddelde van alle bijhorende y-waarden). Dat is de populatie-regressierechte van "lengte van zonen" over "lengte van vaders". Wat die populatie-rechte is zal je nooit weten want je kan niet de hele populatie opmeten. Daarom probeer je die rechte te benaderen door een rechte die je wel kan vinden. Je gebruikt daarbij dingen die je kent, namelijk de opmetingen uit je steekproef.

4. De regressierechte

4.1. Wat zegt het model?

Bij de studie van de lengte van vaders en zonen heb jij tot nu toe vastgelegd:

1. dat je de lengte van zonen als respons neemt (en dus als opmetingen die **aan het toeval onderhevig** zijn) en de lengte van vaders als verklarende veranderlijke (waarvan je de waarden behandelt alsof ze **exact** gekend zijn)

2. dat je een verband zoekt tussen **de lengte** van vaders en **de gemiddelde lengte** van zonen

3. dat dit verband er moet uitzien als **een rechte**.

Op basis van de drie voorgaande eisen ga je nu op zoek naar de rechte die het best het verband tussen de lengte van vaders en de gemiddelde lengte van zonen weergeeft. Je gebruikt hierbij de puntenwolk van je opmetingen samen met je kennis dat de vergelijking van een rechte er uitziet als y = ax + b.

4.2. Wat is best?

Je hebt nu nog een criterium nodig voor "best". Welke rechte y = ax + b geeft het gezochte verband het best weer in jouw puntenwolk?

Het criterium dat je hier gebruikt, is niets anders dan een uitbreiding van het criterium voor het gemiddelde waarmee je hierboven hebt gewerkt. Toen ben je op zoek gegaan in een verticale strip naar een punt c dat "zo dicht mogelijk" tegen de meetpunten in die strip lag. Als eis heb je toen gesteld dat "de som van de kwadraten van de verticale afstanden tot het punt c" zo klein mogelijk moest zijn. Het beste punt c dat hieruit te voorschijn kwam, was het gemiddelde $\bar{y}$ (binnen de strip).

Nu ga je niet op zoek naar een beste punt in een verticale strip, maar naar een beste rechte voor de hele puntenwolk. Je zoekt een rechte y = ax + b die "zo dicht mogelijk" tegen alle punten van de puntenwolk ligt. Als eis stel je nu dat "de som van de kwadraten van de verticale afstanden tot de rechte y = ax + b" zo klein mogelijk moet zijn. De beste rechte die hieruit te voorschijn komt is de regressierechte die je noteert als $\hat{y} = a\,x + b$. Dat doe je vanaf nu altijd zo (en dat is dus een andere notatie dan wat Ann en Jef in het begin van deze tekst gebruikt hebben). De notatie $\hat{y}$ lees je als y-hoed.

4.3. De regressierechte

Als je de vergelijking $\hat{y} = a\,x + b$ ontmoet, dan weet je dat het hier niet om zomaar een rechte gaat, maar dat je te maken hebt met de **regressierechte** die hoort bij je puntenwolk. Het is de unieke rechte waarbij a en b zo bepaald zijn dat de som van de kwadraten van de verticale afstanden tot die rechte zo klein mogelijk is. Om die reden wordt de regressierechte ook "**de kleinste kwadraten rechte**" genoemd.

De vergelijking $\hat{y} = a\,x + b$ van de regressierechte gebruik je om bij een x-waarde (**lengte** van vader) de bijhorende $\hat{y}$ -waarde (**gemiddelde lengte** van zonen) te berekenen en niet omgekeerd.

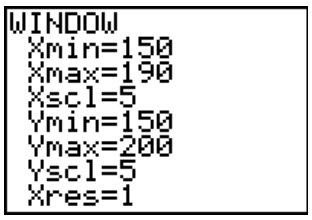
De betekenis van x, y, en $\hat{y}$ staat in onderstaande tabel nog eens samengevat. Let goed op het verschil tussen de punten van de puntenwolk [genoteerd als $(x_i\,,\,y_i)$] en de punten op de regressierechte [genoteerd als $(x_i\,,\,\hat{y}_i)$].

	horizontale as	verticale as
in woorden	de lengte van de vader neem je als verklarende veranderlijke (x-veranderlijke)	de lengte van de zoon neem je als respons (y-veranderlijke)
een opmeting $(x_i\,,\,y_i)$	x_i = lengte van de vader	y_i = lengte van de zoon
een punt $(x_i\,,\,\hat{y}_i)$ op de regressierechte	x_i = lengte van de vader	$\hat{y}_i$ = gemiddelde lengte van zonen

Hoe je met je GRM de regressierechte berekent, weet je al. Druk [STAT], loopt naar CALC, druk 4:LinReg(ax+b) en vervolledig het commando zoals aangegeven. Dat doe je als volgt. Druk [2nd] [LIST], loop naar beneden tot je naast VADER staat en druk [ENTER]. Druk [,] en dan terug [2nd] [LIST], loop naar beneden tot je naast ZOON staat en druk twee keer [ENTER]. Het commando dat op die manier is ingebracht, berekent een regressierechte $\hat{y} = a\,x + b$ (je GRM gebruikt de notatie y=ax+b) op basis van een puntenwolk (x_i , y_i) waarvan de x-coördinaten in de lijst VADER staan (de x-veranderlijke is de lengte van de vader) en de y-coördinaten in de lijst ZOON (de y-veranderlijke is de lengte van de zoon).

Opdracht 7

Bereken (met je GRM) de regressierechte zoals hierboven aangegeven. Teken die rechte op je figuur. Maak die rechte, samen met de puntenwolk, ook zichtbaar op je GRM. Gebruik de WINDOW instellingen zoals aangegeven.
Vul in: voor x = 175 is $\hat{y}$ = …… Zeg dit ook in woorden (gebruik de juiste betekenis van x en van $\hat{y}$) en teken dit verband op je figuur.



5. Uitkomsten voorspellen

In de vorige voorbeelden heb je met puntenwolken gewerkt waarbij er per vaste x-waarde meerdere y-waarden corresponderen (je hebt meerdere punten in een verticale strip). Zo'n voorbeelden helpen je om goed te begrijpen hoe regressie echt werkt. Je ziet dan dat het gaat om een verband tussen vaste x-waarden en gemiddelde y-waarden. Nu je dat weet kan je ook werken met puntenwolken waarbij er bij een bepaalde x-waarde slechts 1 y-waarde is opgemeten. Dat kom je vaak tegen. Alle begrippen die je over regressie geleerd hebt blijven onveranderd geldig.

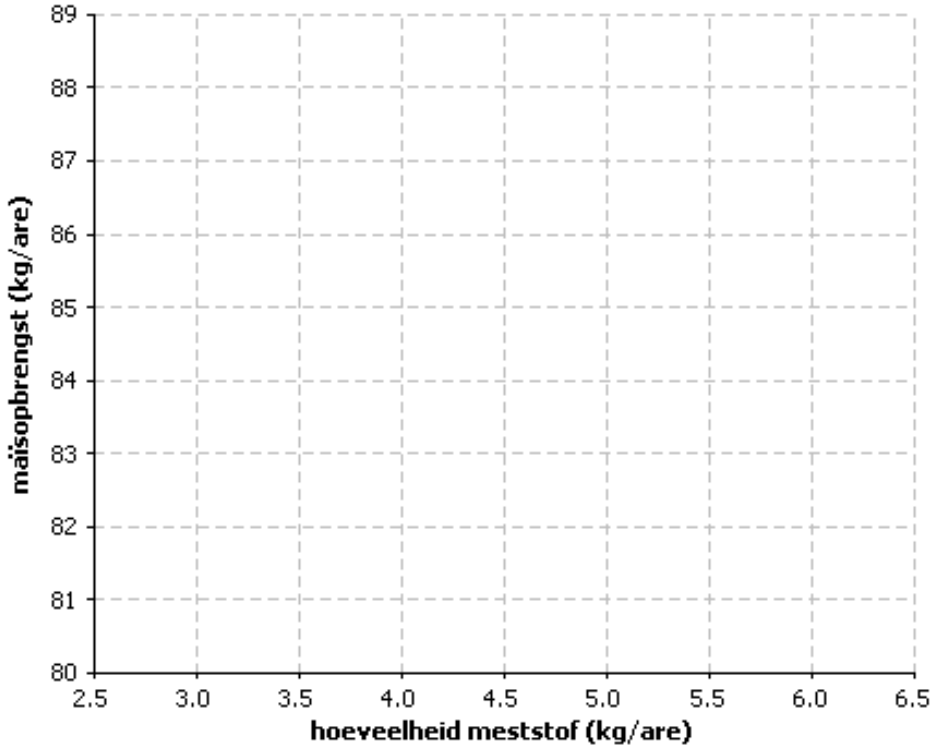
5.1. Opbrengst en meststof

In een landbouwschool heeft men op verschillende percelen een verschillende hoeveelheid meststof gestrooid bij het zaaien van maïs. Bij de oogst is, per perceel, zowel de hoeveelheid meststof als de opbrengst genoteerd. Het is de bedoeling om na te gaan hoe de opbrengst wijzigt naarmate men meer of minder meststof gebruikt. In deze studie is dus de hoeveelheid gestrooide meststof de verklarende veranderlijke en de maïsopbrengst is de respons.

De resultaten waren als volgt:

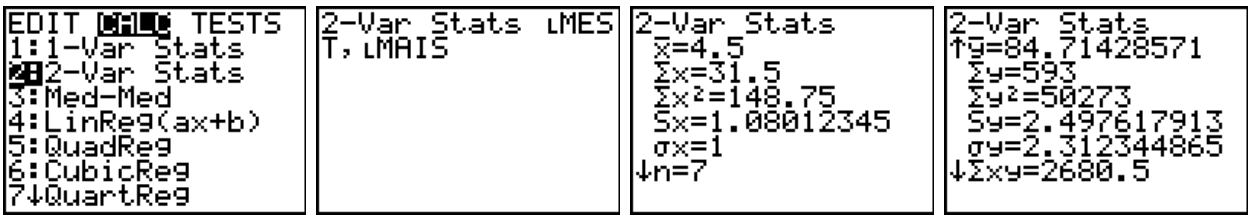
x_i = meststof (kg/are)	3.0	3.5	4.0	4.5	5.0	5.5	6.0
y_i = opbrengst (kg/are)	83	81	86	83	87	85	88

Deze data kan je vinden op http://www.uhasselt.be/lesmateriaal-statistiek waar je de bestanden MAIS.8xl en MEST.8xl kan downloaden. Breng deze bestanden over naar je GRM als lijsten MAIS en MEST. Je kan ze natuurlijk ook gewoon intikken.



Opdracht 8

Verzamel eerst enkele getallen die je bij deze studie nodig hebt. Het zwaartepunt van je puntenwolk is het punt $(\bar{x}, \bar{y})$ wat je bijvoorbeeld als volgt kan vinden. Druk [STAT], loop naar CALC en werk verder zoals aangegeven. Bereken daarna ook de regressierechte. Hoe je dat moet doen weet je nu al.



1. Teken op de figuur je puntenwolk en duid (met een speciaal symbool zoals een sterretje) ook het zwaartepunt aan. Schrijf ook op waaraan dit zwaartepunt gelijk is.

2. Als je naar de puntenwolk kijkt en naar de correlatiecoëfficiënt, is het dan zinvol om naar een lineair verband tussen meststof en opbrengst te zoeken? Is dit verband positief of negatief en is het zwak, matig of sterk?

3. In deze studie wordt een regressie opgesteld van …… over …… (vul in). Teken nu de regressierechte op je figuur. Schrijf ook haar vergelijking in de juiste notatie. Bemerk dat de regressierechte door het zwaartepunt loopt (dat is altijd zo, men kan dat wiskundig bewijzen).

5.2. Verwachte respons

De regressierechte $\hat{y} = 1.7\,x + 77$ geeft het verband tussen x = de hoeveelheid meststof en $\hat{y}$ = de gemiddelde maïsopbrengst (of de "verwachte" opbrengst). De regressierechte laat toe om te voorspellen welke gemiddelde respons $\hat{y}_i$ je verwacht bij een gekozen waarde x_i van de verklarende veranderlijke. Daarbij ben je niet verplicht om een x_i te kiezen uit de waarden die je gebruikt hebt bij je experiment. Maar je moet wel naar de context blijven kijken. De gevonden regressierechte $\hat{y} = 1.7\,x + 77$ is hier slechts zinvol binnen een bepaald gebied. In je experiment heb je hoeveelheden meststof gebruikt die gaan van 3 tot 6 kg/are. Binnen dat gebied en waarschijnlijk ook nog tot een beetje erbuiten (zoals van 2 tot 7 kg/are) geeft de gevonden rechte zinvolle verbanden. Maar je mag niet te ver gaan. Als extreem voorbeeld zou je hebben dat bij 100 kg/are meststof je een gemiddelde opbrengst van $\hat{y} = 1.7\,x + 77 = (1.7)\,(100) + 77 = 247$ kg/are zou hebben. Maar bij zo'n overbemesting gaan maïsplanten dood en heb je helemaal niets.

Bij regressiestudies ontmoet je verschillende uitspraken die, als je ze goed interpreteert, allemaal hetzelfde betekenen.

Als $\hat{y} = 1.7\,x + 77$ dan zegt men:

1. bij 5 kg/are meststof is de **gemiddelde opbrengst** 85.5 kg/are maïs
2. bij 5 kg/are meststof **verwacht je een opbrengst** van 85.5 kg/are maïs
3. bij 5 kg/are meststof **hoort een opbrengst** van 85.5 kg/are maïs

De eerste uitspraak geeft expliciet weer dat je met regressie te maken hebt waar een gemiddelde respons ŷ hoort bij een vaste x-waarde.

De tweede uitspraak gaat over wat er verwacht wordt. Dit moet je niet opvatten als een perfect verband waarbij je elke keer opnieuw, bij het toedienen van 5 kg/are meststof, exact 85.5 kg/are maïs zal oogsten. Het gaat over een verwachting van wat je gemiddeld zal te zien krijgen als je op vele percelen telkens 5 kg/are meststof zou strooien.

De derde uitspraak is de kortste maar ook de meest gevaarlijke. Als je die uitspraak gebruikt, geef dan heel goed aan dat het hier over regressie gaat waarbij "opbrengst" de respons is. Dan weet men dat het hier niet over een wiskundige vergelijking gaat waarbij uit de kennis van x de waarde van ŷ volgt en uit de kennis van ŷ de waarde van x. Want wat is "de kennis van ŷ"? Hoe moet je ŷ interpreteren? Je mag hier niet zeggen dat je uit de opmetingen kan besluiten dat je 5 kg/are meststof nodig hebt voor een maïsopbrengst van 85.5 kg/are. Je hebt geleerd dat je in regressie niet mag omdraaien! De speciale notatie y-hoed (ŷ) helpt je om daar telkens weer aan te denken.

Nota.
Om goed aan te geven dat je met regressie werkt, is het aanbevolen om de uitdrukkingen "gemiddelde opbrengst" of "verwachte opbrengst" te gebruiken wanneer "opbrengst" de respons is in je regressiestudie.

Opdracht 9

Duid op je figuur de punten (x_i , y_i) en $(x_i , \hat{y}_i)$ aan voor $x_i = 5$. Wat is hier de betekenis van x_i , van y_i en van $\hat{y}_i$? Zeg dit in woorden.
Welke maïsopbrengst verwacht je als je 3.8 kg/are meststof strooit? Verklaar je antwoord.

6. Nonsens regressierechten

Als je alleen maar naar wiskundige technieken kijkt dan kan je bij elke dataset (x_i , y_i) op zoek gaan naar een rechte die er “best” bij aansluit volgens het criterium der kleinste kwadraten.

Je weet echter dat een grafische voorstelling en de context van een studie essentiële elementen zijn bij elk statistisch onderzoek. Hieronder zie je voorbeelden van wat er kan gebeuren als je, zonder nadenken, zomaar op de knop LinReg(ax+b) drukt.

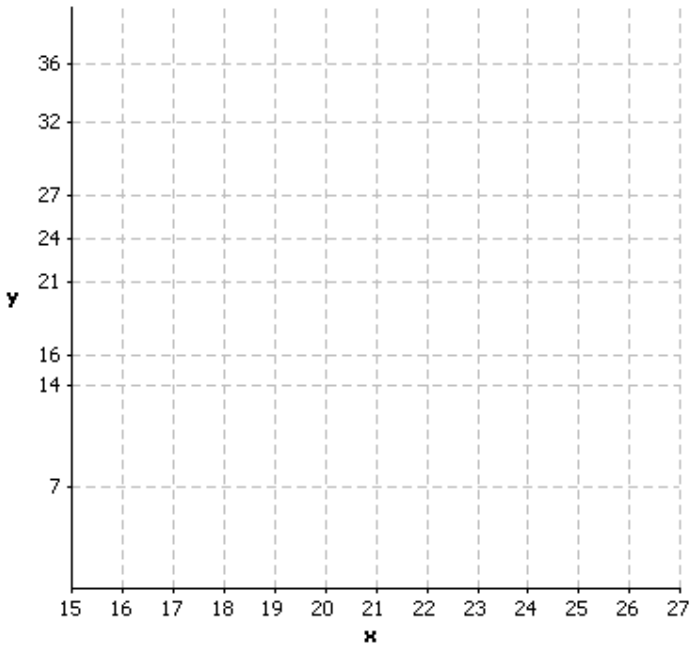
6.1. Lineair verband maar zinloze context

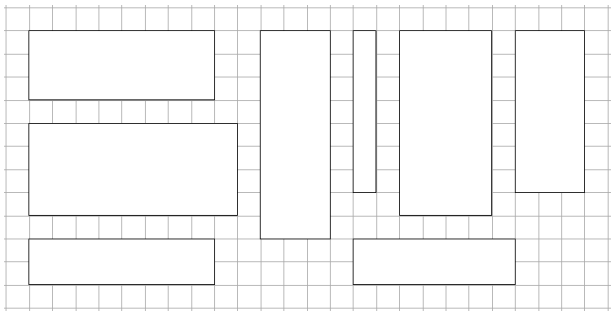
Hieronder staan de resultaten van 8 opmetingen (x_i , y_i).

x_i	22	18	20	26	24	16	24	20
y_i	24	14	21	36	32	7	27	16

Opdracht 10

1. Teken op de figuur hiernaast de puntenwolk van die 8 opmetingen.
2. Geeft de puntenwolk de indruk gespreid te zijn rond een rechte?
3. Hoe zou je, op zicht, het lineaire verband karakteriseren? Positief of negatief, zwak, matig of sterk?
4. Wordt je uitspraak bevestigd door de correlatiecoëfficiënt? Om die te berekenen tik je eerst de x-coördinaten in de lijst [L1] en de y-coördinaten in [L2]. Daarna gebruik je de commando's die je hierboven geleerd hebt, om de correlatiecoëfficiënt te bepalen samen met de regressierechte.
5. Schrijf de vergelijking van de regressierechte in de juiste notatie en teken die rechte op je figuur.





Alles wat je tot nu toe gedaan hebt is correct wanneer je alleen maar naar getallen kijkt. Maar eigenlijk zijn die getallen afkomstig van een leerling die rechthoekjes aan het tekenen was. De acht rechthoeken zie je hiernaast. Per rechthoek heeft die leerling de omtrek en de oppervlakte berekend. Dat zijn de acht x-coördinaten (omtrek) en de acht y-coördinaten (oppervlakte) die je hierboven gekregen hebt. Alles is genoteerd in maatgetallen, zonder eenheden (cm voor omtrek en cm^2 voor oppervlakte).

Opdracht 11

Schrijf in woorden wat de gevonden regressierechte betekent in de context van de rechthoeken van die leerling. Is dit zinvol? Welke oppervlakte verwacht je wanneer je een rechthoek hebt waarvan de omtrek 10 cm is? Probeer die eens te tekenen.

6.2. Zinvol verband maar niet lineair

Heel wat verbanden in economie, biologie, psychologie,… zijn zinvol te modelleren met behulp van wiskundige functies. Maar die functies hoeven niet lineair te zijn. Hun grafiek kan er bijvoorbeeld uitzien als een parabool.

Een puntenwolk opgesteld door Anscombe (die dataset vind je ook in de afzonderlijke tekst over correlatie) start met de volgende gegevens:

x_i	4	5	6	7	8	9	10	11	12	13	14
y_i	3.10	4.74	6.13	7.26	8.14	8.77	9.14	9.26	9.13	8.74	8.10

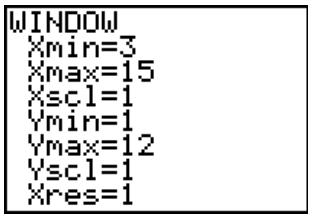
Deze data kan je vinden op http://www.uhasselt.be/lesmateriaal-statistiek waar je de bestanden ANSX.8xl en ANSY.8xl kan downloaden. Breng deze bestanden over naar je GRM als lijsten ANSX en ANSY.

Opdracht 12

Onderstel dat je voor de dataset van Anscombe geen grafiek maakt maar dat je met je GRM rechtstreeks de correlatiecoëfficiënt en de regressierechte berekent. Wat zou je dan op basis van de correlatiecoëfficiënt zeggen over de samenhang tussen x en y? Welke y-waarde verwacht je dan als x = 9 ?

Opdracht 13

Teken de puntenwolk samen met de regressierechte op je GRM (kijk hoe je dat vroeger gedaan hebt en gebruik de WINDOW instellingen zoals aangegeven). Ga tevens op de regressierechte staan bij x = 9. Is de regressierechte een goed model voor deze gegevens? Kan je dat weten als je alleen maar de berekeningen van opdracht 12 uitvoert?



7. Samenvatting

Hieronder staan enkele stappen waarop je moet letten bij het opstellen van een regressierechte. Daarmee is over regressie lang niet alles gezegd. Je kan allerlei problemen ontmoeten. Wat doe je met "uitzonderlijke" punten, die ver buiten het globale patroon vallen? Wat doe je als de puntenwolk helemaal niet rond een rechte verstrooid is? Wat doe je met de steekproef van je medeleerling die bij dezelfde studie andere opmetingen had en dus ook een andere regressierechte?
Op al deze vragen zijn er antwoorden, maar dat vergt een diepere studie van regressie.

In deze tekst heb je kennis gemaakt met regressie. Je hebt daarbij zeer belangrijke basisbegrippen ontmoet. Die moet je bij elke regressiestudie goed in het oog blijven houden. Zij zijn samengevat in de onderstaande punten.

Bij een studie waar je een regressierechte opstelt, doorloop je de volgende stappen:

1. Bepaal wat de bedoeling van de studie is en wat hieruit volgt voor de keuze van de respons en van de verklarende veranderlijke.

2. Denk aan de context van de studie. Misschien hebben vroegere studies een lineair verband gevonden of volgt er uit de aard van het probleem dat een lineair verband verantwoord is.

3. Doe je opmetingen (x_i, y_i) waarbij x_i de waarde van de verklarende veranderlijke is en y_i de waarde van de respons.

4. Teken een puntenwolk en controleer of de globale vorm van de puntenwolk de onderstelling van een lineair verband niet tegenspreekt.

5. Wanneer je (na punt 2 en 4) ervan overtuigd bent dat een lineair verband zinvol is, voer dan de regressie van y over x uit en stel de regressierechte $\hat{y} = a\,x + b$ op.

6. Op basis van de gevonden regressierechte doe je uitspraken over de gemiddelde (of de verwachte) waarde $\hat{y}$ van de respons bij een bepaalde waarde x van de verklarende veranderlijke. Je weet dat je dergelijke uitspraken niet mag omdraaien. Je mag ook geen x-waarden nemen die ver buiten het gebied liggen van de x-waarden die je in de studie gebruikt hebt (gevaar van extrapolatie).

7. Je gevonden regressierechte (en dus ook je conclusie) is slechts een benadering van de echte regressierechte voor de populatie. Je hebt immers gewerkt met een steekproef. Een nieuwe steekproef geeft een andere puntenwolk en een andere regressierechte. Uitspraken over lineaire verbanden in een populatie horen thuis in de verklarende statistiek.

8. Data snooping

Het verband tussen de respons en de verklarende veranderlijke is een onderstelling die je vooraf maakt. Het maakt deel uit van de populatie-eigenschappen. Zo kan je er bijvoorbeeld vooraf van overtuigd zijn dat de gemiddelde respons op een rechte $\alpha x + \beta$ ligt (Griekse letters voor een populatiekarakteristiek). Die rechte schat je dan met behulp van je steekproef. Dat doe je door de rechte $\hat{y} = ax + b$ te zoeken die volgens het criterium der kleinste kwadraten het best aansluit bij je puntenwolk.

Het kan zijn dat de puntenwolk er helemaal niet uitziet als "punten gespreid rond een rechte", maar veeleer als een parabool. Dat brengt je op het idee dat, in de populatie, de gemiddelde respons zich gedraagt als $\alpha x^2 + \beta x + \gamma$. En dus ga je nu op zoek naar een parabool $ax^2 + bx + c$ die het best aansluit bij je puntenwolk.

Bovenstaande manier van werken is normaal in het kader van exploratief onderzoek. Je probeert een idee te krijgen over het gedrag van een populatie. Zo'n idee kan komen uit theoretische eigenschappen, of uit vroeger onderzoek, of uit een pilootstudie die je hebt opgezet, enz.

Als je nadien (met verklarende statistiek) populatie-eigenschappen wil "bewijzen" dan mag je daarvoor niet dezelfde gegevens gebruiken die je "op het idee gebracht hebben" over de vorm van de te zoeken functie. Zo'n manier van werken noemt men "data snooping". Je gebruikt data om te raden hoe een populatie-eigenschap er uitziet en daarna gebruik je dezelfde data om die populatie-eigenschap "te bewijzen". Dat mag niet. Je moet dan een nieuwe steekproef trekken en het toeval zijn rol laten spelen. Zo werkt statistiek.

DEEL 2. De formules achter de ideeën

Zoals in vorig deel ontmoet je ook hier ideeën en formules waarover je verdere informatie kan vinden in de afzonderlijke tekst “Correlatie. Achtergrondinformatie”.
Deze tekst is beschikbaar op http://www.uhasselt.be/lesmateriaal-statistiek/ .

9. Houd het simpel: standaardiseer

Een grafische voorstelling is een zeer belangrijk onderdeel van een regressiestudie. Dat betekent dat je altijd een puntenwolk moet tekenen en dan goed kijken hoe die eruit ziet. Hier is soms een probleem. De keuze van de lengte van de eenheden op de assen kan tot een figuur leiden die je op het verkeerde been zet. Heel wat grafische pakketten maken zelf een keuze bij het tekenen van een puntenwolk. Dikwijls gebruiken zij daarbij een “landscape” formaat waarbij de eenheid op de x-as langer is dan op de y-as. Maar ook als je zelf zo’n figuur moet tekenen, wat doe je dan? In deel 1 zijn we gestart met het opmeten van temperatuur en volume bij de studie van een gas. Als je temperatuur op de x-as uitzet, neem je dan stappen van 1 cm om 100 Kelvin voor te stellen? En neem je 1 cm op de y-as per 10 milliliter of per 100 milliliter of per …?

9.1. Gezichtsbedrog

Een puntenwolk teken je meestal een beetje op het gevoel en je tracht ervoor te zorgen dat ze duidelijk is. In veel gevallen is dat voldoende als eerste stap. Toch weet je niet of de figuur je echt goed toont wat er in je dataset te beleven valt. Probeer dat maar eens uit in de volgende opdracht.

Opdracht 14

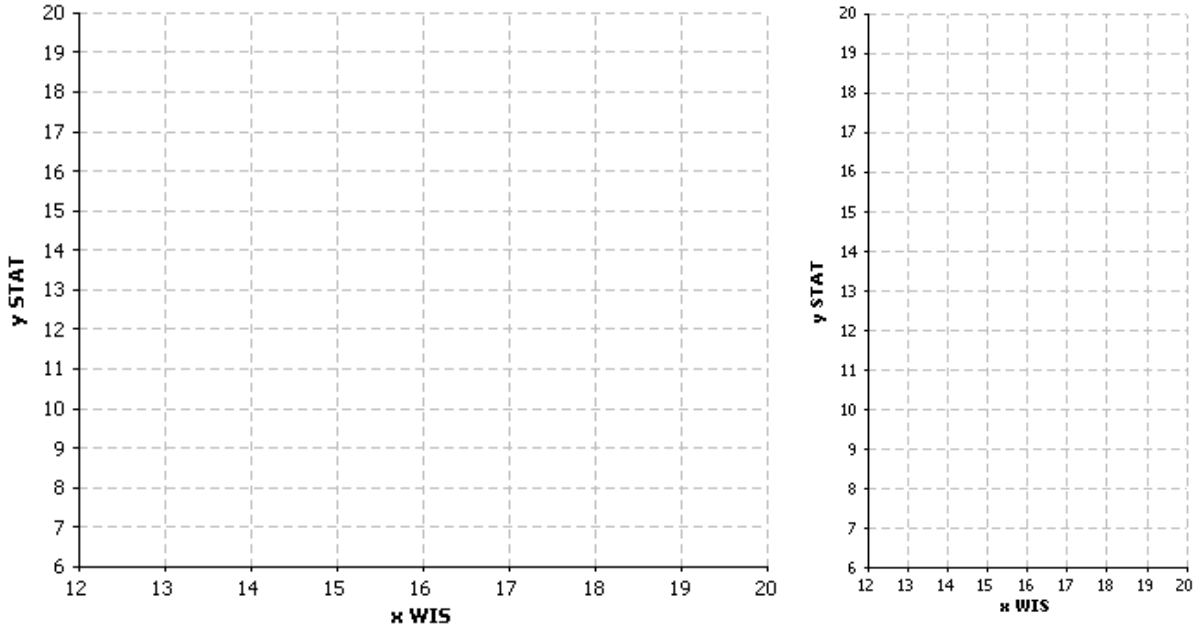
Een onderzoek tracht, op basis van de punten die je haalt op wiskunde in je laatste jaar secundair, te voorspellen welke punten je zal halen op het vak statistiek in je eerste jaar hoger onderwijs.

Een steekproef van 13 studenten leverde volgende resultaten (wis = punt op wiskunde in het secundair onderwijs, stat = punt op statistiek in het hoger onderwijs):

wis x_i	18	13	15	17	16	16	13	19	16	14	19	17	15
stat y_i	13	11	7	15	13	9	7	19	17	13	15	19	11

Deze data kan je vinden op http://www.uhasselt.be/lesmateriaal-statistiek waar je de bestanden WIS.8xl en STAT.8xl kan downloaden. Breng deze bestanden over naar je GRM als lijsten WIS en STAT.

Teken deze puntenwolk op de twee figuren hieronder.
Als je nu “op zicht” een uitspraak moet doen over deze puntenwolken, geven zij dan een even sterke stijging weer? En geven zij de indruk dat de samenhang even sterk is? Beschrijf wat je ziet en probeer dit ook te verklaren.

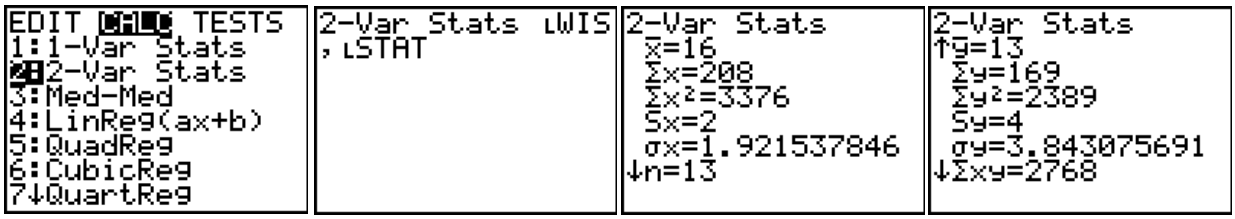


De twee puntenwolken zijn afkomstig van dezelfde dataset en toch zou je ze “op zicht” niet op dezelfde manier interpreteren. Als je een voorbeeld wil zien waar het helemaal fout loopt, doe dan maar eens het experiment dat beschreven is op blz. 23–24 van “Correlatie. Achtergrondinformatie” en lees daarna de oplossing op blz. 27–28 in die tekst.

Zomaar “op het gevoel” de fysische lengte van een eenheid op de assen kiezen is blijkbaar niet zo’n goed idee. Dus ga je op zoek naar een assenstelsel dat ongevoelig is voor “lengteproblemen” zodat je tenminste daardoor niet meer in verwarring geraakt. Je doet dit door over te stappen op z-scores. Dat is een zeer krachtige techniek die niet alleen figuren standaardiseert, maar die je ook toelaat om appels met peren te vergelijken, of met examenpunten, of met je BMI, …. Hoe dat werkt, lees je hieronder.

9.2. z-scores

Elke dataset kan je op een eenvoudige manier transformeren naar een andere dataset die bestaat uit z-scores. Je hebt daarvoor het gemiddelde en de standaardafwijking nodig. Voor de behaalde examenpunten uit de vorige opdracht vind je die kengetallen eenvoudig als volgt:



Een z-score is een getal dat aangeeft hoeveel standaardafwijkingen een observatie boven (positief) of onder (negatief) het gemiddelde ligt.

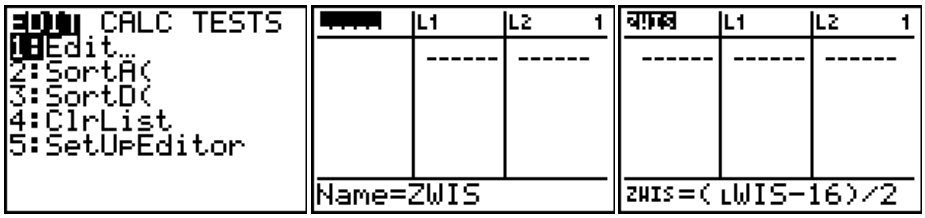
WIS	STAT
$\bar{x} = 16$	$\bar{y} = 13$
$s_x = 2$	$s_y = 4$

Bij de examenpunten op wiskunde was het gemiddelde $\bar{x} = 16$ en de standaardafwijking $s_x = 2$. Het eerste opgemeten examenpunt was $x_1 = 18$. De z-score is hier gelijk aan 1 want $18 = 16 + 2 = \bar{x} + (1)s_x$ .Op dezelfde manier is de z-score van $x_2 = 13$ gelijk aan -1.5 want 13 ligt 1.5 standaardafwijkingen onder het gemiddelde aangezien $13 = 16 - 3 = \bar{x} - (1.5)s_x$.

Bij elke x_i hoort een z-score z_{x_i} met $z_{x_i} = \dfrac{x_i - \bar{x}}{s_x}$.

Bemerk dat een z-score eenheidsloos is omdat zowel x_i als $\bar{x}$ als s_x allemaal dezelfde eenheid hebben. Door het quotiënt te maken vallen die eenheden weg.

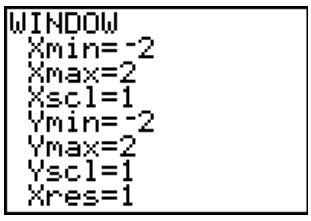
Met je GRM bepaal je de z-scores van de examenpunten op wiskunde als volgt. Druk [STAT], 1:Edit.. en ga op de kop van lijst [L1] staan. Druk [2nd] [INS] zodat je een nieuwe lijst krijgt waar je onderaan het scherm de naam ZWIS invult. Druk dan twee keer [ENTER] en vul in zoals aangegeven. Om de naam van de lijst WIS in te vullen druk je [2nd] [LIST], loop dan naar de lijst WIS en druk [ENTER]. Als alles ingevuld is druk je [ENTER]. De lijst ZWIS bevat nu de z-scores van de lijst WIS.



Opdracht 15

Maak de lijst ZWIS en ZSTAT zodat je nu lijsten hebt met z-scores van de examenpunten op wiskunde en op statistiek. Bereken (met GRM) het gemiddelde en de standaardafwijking van de lijsten met z-scores. Wat leid je hieruit af voor het zwaartepunt van de puntenwolk van de z-scores? Kan je verklaren waarom dit altijd het geval is (gebruik de definitie van z-scores samen met een eigenschap van het gemiddelde)?

ZWIS	ZSTAT
$\bar{z}_x = \ldots..$	$\bar{z}_y = \ldots.$
$s_{z_x} = \ldots.$	$s_{z_y} = \ldots.$



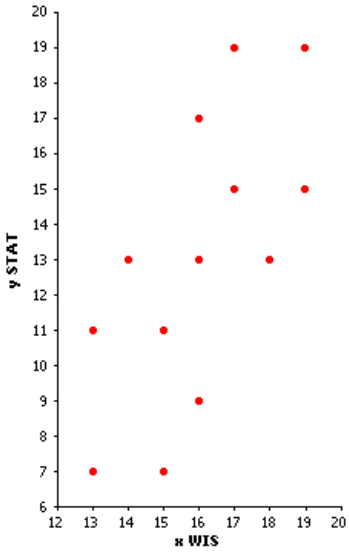
Teken nu die puntenwolk. Dat doe je met 2nd [STAT PLOT]. De z-scores wiskunde zet je op de horizontale as en de z-scores statistiek op de verticale. Gebruik de instellingen voor WINDOW zoals aangegeven. Controleer bovendien dat alle functies Y= af staan. Druk dan TRACE. Ligt het zwaartepunt van de puntenwolk waar je het had verwacht?

Besluit

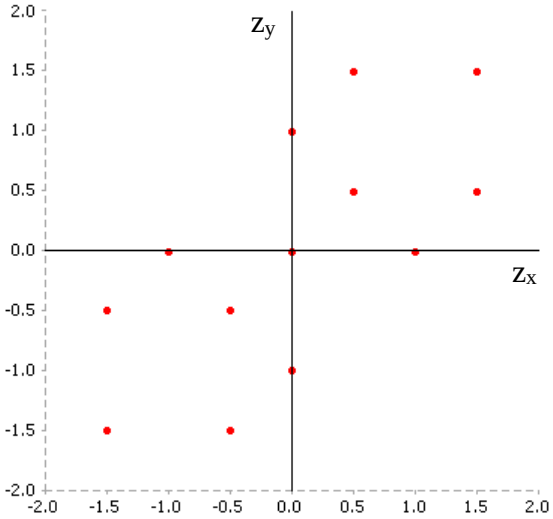
Elke bivariate dataset kan je standaardiseren door over te stappen op z-scores. Als je dan een vaste afstand (op je papier of computerscherm) kiest van bijvoorbeeld 2 cm per eenheid, zowel op de horizontale als op de verticale as, dan krijg je "gestandaardiseerde puntenwolken" die je "fysisch" op elkaar kan leggen om ze te vergelijken. Of de oorspronkelijke opmetingen dan over temperatuur, gewicht, lengte, examenpunten of wat dan ook gaan, dat heeft allemaal geen belang. De corresponderende z-scores zijn eenheidsloos.

Overstappen van oorspronkelijke gegevens op z-scores: $z_x = \frac{x - \bar{x}}{s_x}$ *en* $z_y = \frac{y - \bar{y}}{s_y}$

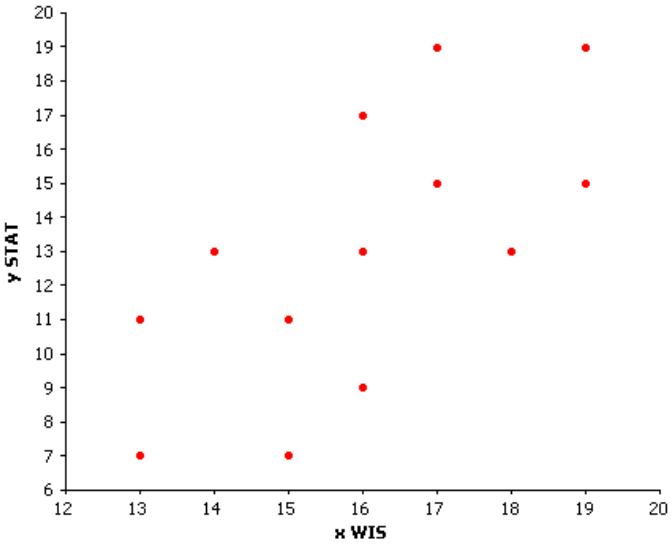
Overstappen van z-scores op oorspronkelijke gegevens: $x = \bar{x} + z_x \cdot s_x$ *en* $y = \bar{y} + z_y \cdot s_y$



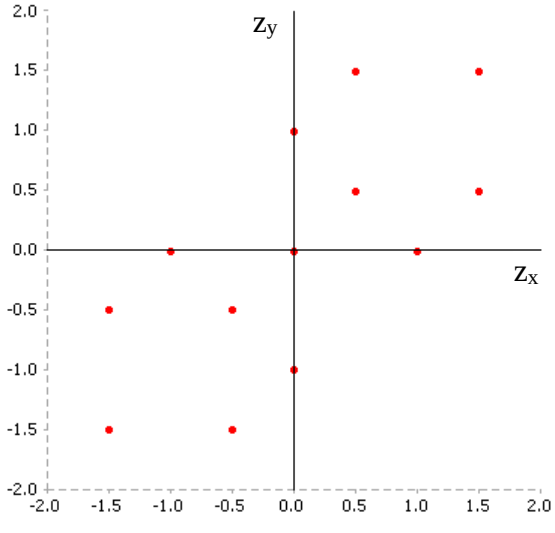
Examenpunten:
oorspronkelijke gegevens



Examenpunten:
getransformeerd naar z-scores



Examenpunten:
oorspronkelijke gegevens



Examenpunten:
getransformeerd naar z-scores

9.3. De regressierechte

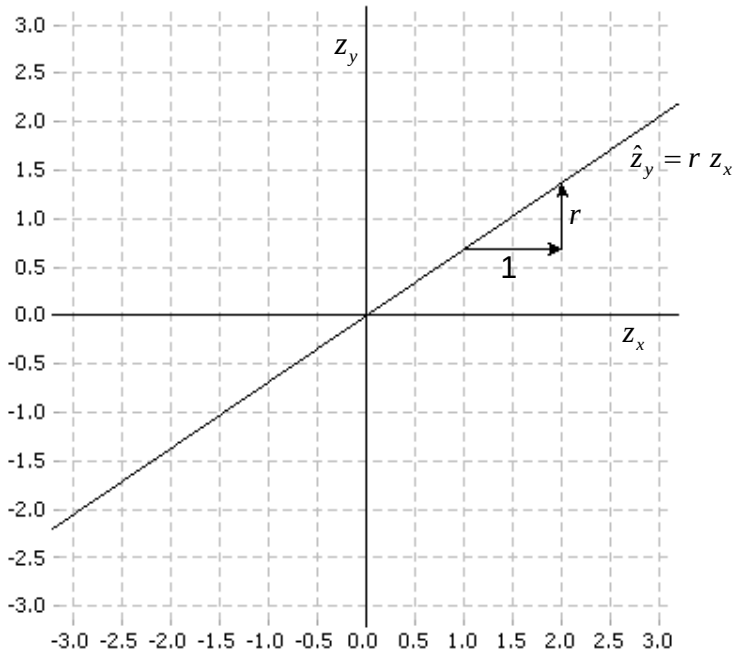
Werken met z-scores heeft zijn voordeel bij de grafische voorstelling van een dataset. Je kan dan op een gestandaardiseerde manier informatie halen uit puntenwolken en je kan ze (zonder gezichtsbedrog) vergelijken met andere puntenwolken.

Ook de vergelijking van de regressierechte is heel eenvoudig als je overstapt op z-scores. Je krijgt dan een rechte die door de oorsprong gaat (want bij z-scores is de oorsprong het zwaartepunt van de puntenwolk) en waarvan de richtingscoëfficiënt gelijk is aan de correlatiecoëfficiënt r (dat kan wiskundig bewezen worden).

> In z-scores wordt de vergelijking van de regressierechte gegeven door:
>
> $$\hat{z}_y = r \cdot z_x$$
>
> waarbij r de correlatiecoëfficiënt is.

Bemerk dat hier de notatie $\hat{z}_y$ wordt gebruikt om duidelijk aan te geven dat het over regressie gaat.

Vroeger heb je de betekenis van de richtingscoëfficiënt (rico) van een rechte geleerd. Als je in een willekeurig punt op de rechte staat, dan zegt de rico hoeveel eenheden je verticaal naar boven of beneden gaat als je een stap van één eenheid vooruitzet in de horizontale richting. In het vlak van de z-scores ziet de regressierechte er dus als volgt uit:

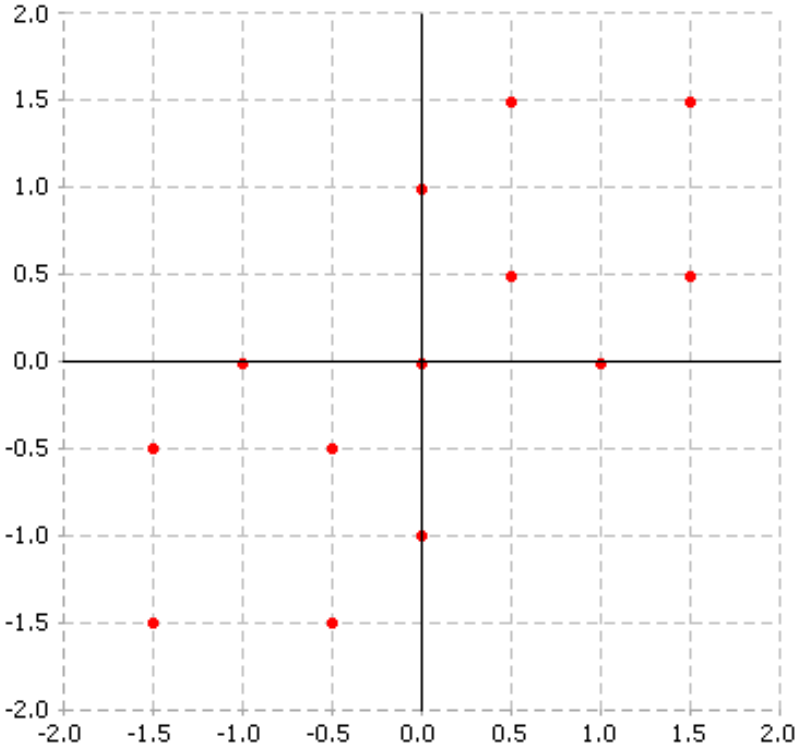


De figuur op de volgende bladzijde heb je nodig bij de volgende opdracht. Zij helpt je om het verband te zien tussen de oorspronkelijke opmetingen en hun z-scores.

Opdracht 16

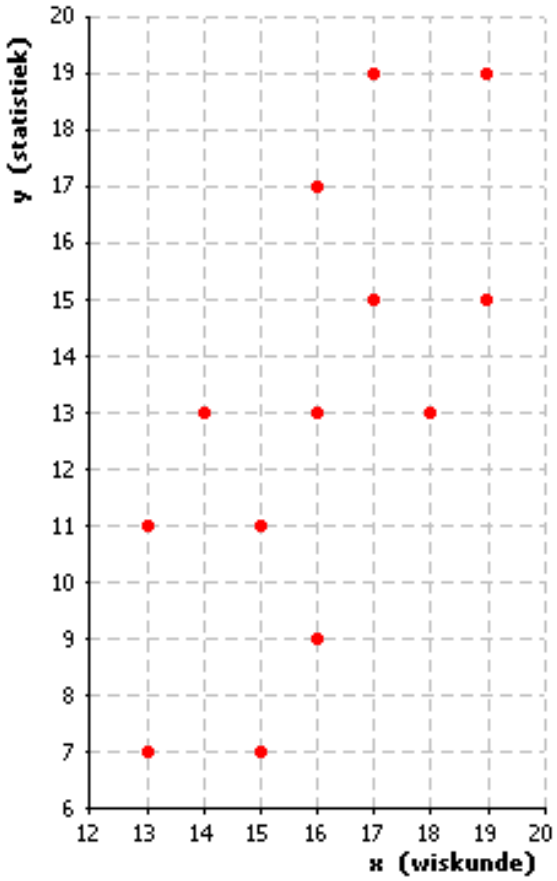
Zoek (met je GRM) de regressierechte in z-scores (dus voor de lijsten ZWIS en ZSTAT). Wat merk je voor de waarden van b, a en r? Teken die rechte op je figuur (met juiste notatie). Transformeer nu het voorschrift van de regressierechte in z-scores naar het voorschrift van de regressierechte in de oorspronkelijke veranderlijken. Vul daarna de volgende waarden in: $\overline{x} = 16,\ s_x = 2,\ \overline{y} = 13,\ s_y = 4,\ r = 0.67$ en teken nu ook deze regressierechte. Zeg in woorden wat de betekenis is van de rico van deze rechte en stel dat ook grafisch voor op je figuur.

z-scores: gestandaardiseerde puntenwolk en regressierechte



==========

oorspronkelijke opmetingen: puntenwolk en regressierechte

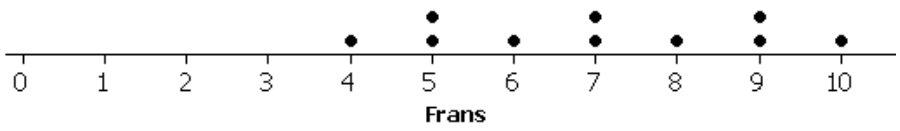


10. Gebruik de juiste meetlat

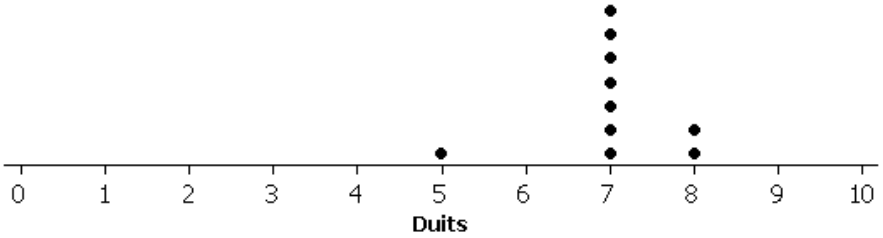
10.1. 5 op 10 gehaald… en dan?

Pol heeft op de toets Frans 5 op 10 gehaald. Het klasgemiddelde was 7. Op de toets Duits haalde Pol ook 5 en ook op die toets was het gemiddelde van de klas 7.
Besluit: De prestatie van Pol was, in vergelijking met zijn klas, twee keer dezelfde. Of niet soms?

Dit verhaal krijgt een heel andere wending als je er een eenvoudig puntendiagram bij tekent.



Bij de toets Frans liggen de punten nogal gespreid. Twee leerlingen haalden een 5, er was ook een leerling met een 4 maar er waren er ook met 9 en 10. Voor de punten van die 10 leerlingen is het gemiddelde 7 en de standaardafwijking is 2.



De toets op Duits ziet er helemaal anders uit. Iedereen haalde daar een 7 of een 8, behalve…. Pol, die had een 5. Bij deze toets is het gemiddelde 7 en de standaardafwijking is 0.8.

Een getal uit een dataset zomaar vergelijken met het gemiddelde vertelt niet het hele verhaal. Soms geeft dit zelfs een verkeerd beeld. De variabiliteit rond dat gemiddelde speelt ook een rol. Bij Frans behaalde Pol een punt dat 1 standaardafwijking onder het gemiddelde ligt, want 5 = 7 – (1)(2). Bij Duits scoorde Pol 2.5 standaardafwijkingen onder het gemiddelde want 5 = 7 – (2.5)(0.8).

De standaardafwijking van een dataset is dikwijls een goede meetlat om punten uit die dataset te vergelijken met hun gemiddelde. Zo houd je ook rekening met de variabiliteit van de gegevens. Als je de standaardafwijking als meetlat neemt dan heeft Pol "–1" op Frans en "–2.5" op Duits. In vergelijking met zijn medeleerlingen is zijn prestatie op Duits veel lager dan op Frans. Uit de oorspronkelijke punten van Pol kan je dat niet te weten komen maar uit zijn z-scores wel. Op Frans haalde Pol een z-score van –1 en op Duits was zijn z-score –2.5.

10.2. Waarom 175 meer kan zijn dan 179

Extra inzicht in regressie krijg je als je de regressierechte op een "statistische" manier bekijkt. Je maakt daarbij gebruik van de klassieke kengetallen: gemiddelde, standaardafwijking en correlatiecoëfficiënt. Als je de regressierechte op die manier begrijpt dan hoef je de wiskundige vergelijking niet meer te onthouden, je kent ze dan automatisch.

De vergelijking van de regressierechte (functievoorschrift in woorden)

Bij een x-waarde die "een aantal" x-standaardafwijkingen boven (of onder) het x-gemiddelde ligt hoort een $\hat{y}$-waarde die "r keer dat aantal" y-standaardafwijkingen boven (of onder) het y-gemiddelde ligt. Hierbij is r de correlatiecoëfficiënt.

Vanuit de eenvoudige gestandaardiseerde vergelijking $\hat{z}_y = r \cdot z_x$ ga je naar $\frac{\hat{y}-\overline{y}}{s_y} = r \cdot \frac{x-\overline{x}}{s_x}$ of verder naar $\hat{y} = r\frac{s_y}{s_x}x + (\overline{y} - r\frac{s_y}{s_x}\overline{x})$. Die vergelijking kan je herschrijven als $\hat{y} = \overline{y} + r(\frac{x-\overline{x}}{s_x})s_y$.

Stel nu $\frac{x-\overline{x}}{s_x} = k$ waarbij k positief (of nul) is als $x \geq \overline{x}$ en negatief anders. Je hebt dan dat $x = \overline{x} + k\, s_x$. Dat betekent dat je naar een x-waarde kijkt die k x-standaardafwijkingen s_x boven (of onder) het x-gemiddelde $\overline{x}$ ligt. De bijhorende $\hat{y}$–waarde wordt dan gegeven door $\hat{y} = \overline{y} + r\,k\,s_y$. Dat is een $\hat{y}$-waarde die $r \cdot k$ y-standaardafwijkingen s_y boven (of onder) het y-gemiddelde $\overline{y}$ ligt.

Als voorbeeld bekijk je de studie van de lengte van vaders en zonen waarbij de kengetallen eruit zagen zoals in de tabel.

Vader	Zoon
$\bar{x}$ =171	$\bar{y}$ =176
s_x =7.13	s_y =9.50
$r = 0.52$	

Een analoge studie werd vroeger door Francis Galton (1857–1936) uitgevoerd. Zoals zijn neef Darwin was Galton geboeid door genetica en daarbij bestudeerde hij allerlei kenmerken, ondermeer de lengte van ouders en hun kinderen. Hierbij ontdekte hij dat zeer grote vaders ook wel grote zonen hebben, maar dat die toch niet zo extreem groot zijn. En zeer kleine vaders hebben kleine zonen, maar toch weer niet zo klein. Dit fenomeen noemde hij "regression towards mediocrity" waarbij hij (enigszins denigrerend) aangaf dat biologische verschijnselen vanuit extreme waarden terugvallen naar de middelmaat.
Misschien kan je beter "regression towards the mean" gebruiken, wat "een terugval naar het gemiddelde" betekent.

In onze studie zijn de zonen van 179 cm inderdaad niet meer zo extreem groot als hun vaders van 175 cm. Voor $x_i = 175$ is $z_{x_i} = \frac{x_i - \bar{x}}{s_x} = \frac{175-171}{7.13} = 0.56$. Een vader van 175 cm steekt dus 0.56 standaardafwijkingen uit boven het gemiddelde. Opdat de zonen van die vaders ook 0.56 standaardafwijkingen groter zouden zijn dan het gemiddelde moet hun lengte (gemiddeld) gelijk zijn aan $\bar{y} + (0.56)\, s_y = 176 + (0.56)\,(9.50) = 181$ cm. Maar dat is niet wat het onderzoek ons leert. De gemiddelde lengte van die zonen is de overeenkomstige $\hat{y}_i$-waarde op de regressierechte. Als je te maken hebt met een x_i die 0.56 x-standaardafwijkingen s_x boven het x-gemiddelde $\bar{x}$ ligt dan valt de bijhorende $\hat{y}_i$-waarde slechts $r \cdot (0.56) = (0.52)(0.56) = 0.29$ y-standaardafwijkingen s_y boven het y-gemiddelde $\bar{y}$. Dus is $\hat{y}_i = \bar{y} + 0.29\, s_y = 176 + (0.29)(9.50) = 178.8 \cong 179$. Die zonen zijn gemiddeld slechts 179 cm in plaats van 181 cm. Als groep zij zijn "teruggevallen" naar het gemiddelde ($\bar{y}$ =176 cm).

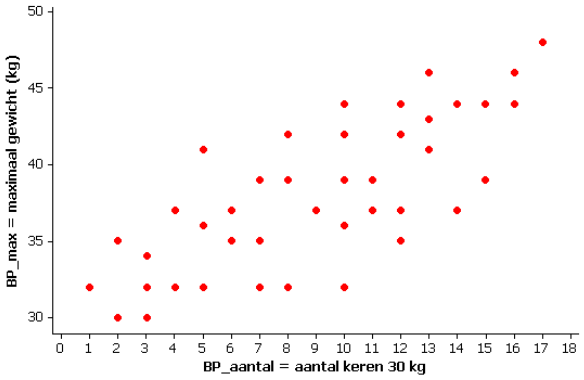
Opdracht 17

De bench press (bankdrukken) is een oefening die bij krachttraining wordt gebruikt. Je moet dan, terwijl je op een bank ligt, een staaf met gewichten langzaam naar je borst brengen om hem daarna weer helemaal omhoog te duwen tot je armen gestrekt zijn.

Het maximale gewicht [BP_max] dat bij bench press kan opgeduwd worden, wordt soms gebruikt als een parameter om te bepalen hoe sterk een atleet is. Hierover zijn er al heel wat studies gedaan bij mannen maar nog relatief weinig bij vrouwen.

Bij het bepalen van BP_max moet je echt tot het uiterste gaan, met risico op blessures. Daarom is het interessant om een verband te zoeken tussen BP_max en een minder gevaarlijke oefening. Daarbij tel je hoeveel keren na elkaar je een vast gewicht van 30 kg kan opduwen [BP_aantal]. Hieruit probeer je dan te bepalen wat je BP_max zou zijn.

Bij een (goed getrokken) steekproef van 41 jonge vrouwelijke atletes (tussen 14 en 17 jaar) werden zowel BP_max als BP_aantal opgemeten. In de puntenwolk zie je (x_i , y_i) = (BP_aantal , BP_max).



1. Is het zinvol om het verband tussen BP_aantal en BP_max voor te stellen door een rechte? Waarom?
2. Is het verband tussen BP_aantal en BP_max stijgend of dalend, zwak, matig of sterk?
3. Als je weet dat het maximale gewicht dat die atletes konden opduwen gemiddeld 38 kg was en je zou lukraak een atlete uit die leeftijdsgroep kiezen, wat voorspel je dan voor haar BP_max? Waarom?
4. Als men je bovendien zegt dat die atlete 12 keer na elkaar een gewicht van 30 kg kan opduwen, wat voorspel je dan voor haar BP_max? Je mag hierbij gebruik maken van de volgende informatie: $\bar{x} = 9$, $s_x = 4.42$, $\bar{y} = 38$, $s_y = 4.87$, $r = 0.75$. Voer een volledige studie uit, gebruik de juiste notatie en leg uit hoe je het gevonden resultaat moet interpreteren.

11. De correlatiecoëfficiënt

De correlatiecoëfficiënt is een maat voor de sterkte van een ***lineair*** verband. Hij komt voor in de vergelijking van de regressierechte die je, op een statistische manier (met kengetallen), kan schrijven als $\hat{y} = \bar{y} + r \cdot \frac{x - \bar{x}}{s_x} s_y$. De rol die de correlatiecoëfficiënt speelt in het kader van een regressiestudie kan je goed begrijpen als je enkele voorbeelden bekijkt. Je vertrekt hierbij telkens van hetzelfde probleem. Vlaamse meisjes van 17 hebben een gemiddelde lengte van 166 cm met een standaardafwijking van 6 cm. Als je lukraak een Vlaams meisje van 17 zou kiezen, hoe groot zal zij dan zijn (in cm)?

Zonder verdere informatie is je beste gok 166 cm. Dat is wat je “als lengte verwacht” of wat je “als uitkomst voorspelt” of wat “het gemiddelde van die groep is”.

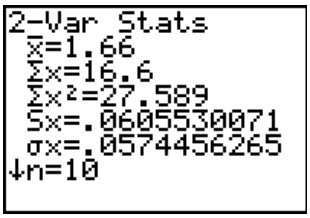
Nu ga je extra informatie toevoegen en kijken of die informatie je helpt om de lengte van dat meisje beter te voorspellen. Als voorbeeld gebruik je een dataset van 10 meisjes waarbij 4 veranderlijken werden opgemeten: de lengte in cm, de lengte in meter tot op twee decimalen, de punten op het vak Nederlands (op een maximum van 20) en het gewicht in kg.

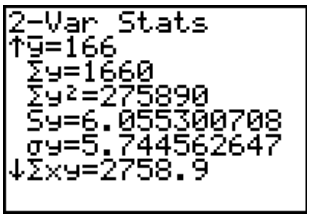
volgnummer	lengte in cm	lengte in m	score Nederlands	gewicht in kg
1	163	1.63	14	62
2	167	1.67	16	53
3	173	1.73	8	71
4	165	1.65	4	56
5	161	1.61	10	50
6	169	1.69	12	65
7	175	1.75	14	68
8	157	1.57	14	48
9	159	1.59	6	51
10	171	1.71	6	61

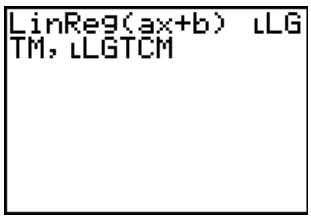
Deze data kan je vinden op http://www.uhasselt.be/lesmateriaal-statistiek waar je de bestanden LGTCM.8xl, LGTM.8xl, SCNED.8xl en GEWKG.8xl kan downloaden. Breng deze bestanden over als lijsten in je GRM.

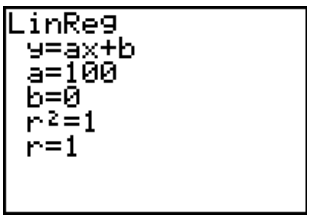
11.1. Een perfect lineair verband

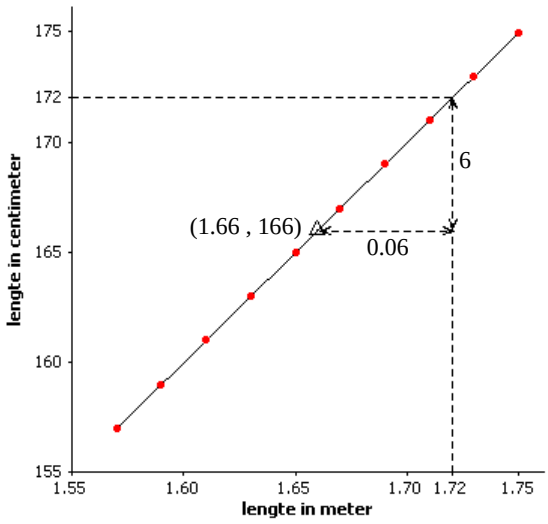
Welke respons $\hat{y}$ (verwachte lengte in centimeter) voorspel je op basis van de waarde van de verklarende veranderlijke (lengte in meter)? Daarvoor kijk je naar de puntenwolk en bepaal je (met je GRM) de regressierechte van "lengte in centimeter" over "lengte in meter". Ook de kengetallen die je nodig hebt om de regressierechte "statistisch" te interpreteren bepaal je met de GRM.











Alle meetpunten liggen hier op een rechte. Er is een perfect lineair en positief verband ($r = 1$).

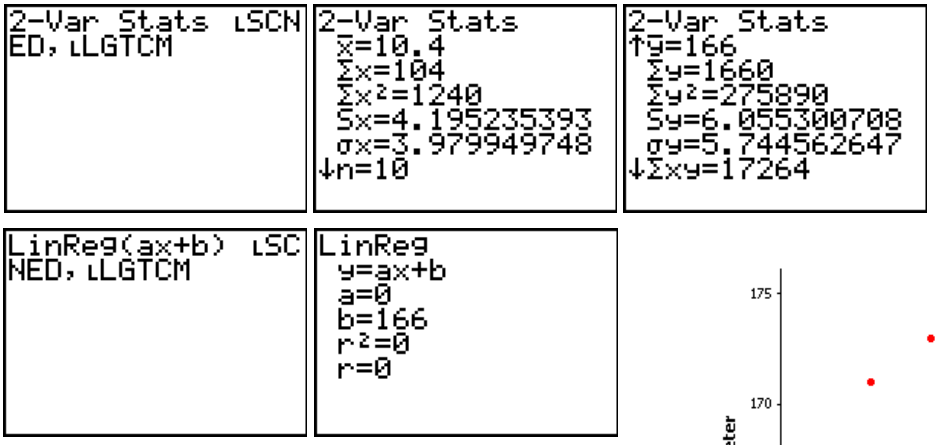
Bij de groep van meisjes waar de verklarende veranderlijke (lengte in m) gelijk is aan x=1.72 is de respons (lengte in cm) voor iedereen y=172 zodat het gemiddelde van die groep ook 172 is. Dat gemiddelde (de waarde $\hat{y}$ op de regressierechte) laat je toe om de lengte (in cm) foutloos te voorspellen zodra je de waarde van de verklarende veranderlijke (lengte in m) kent.

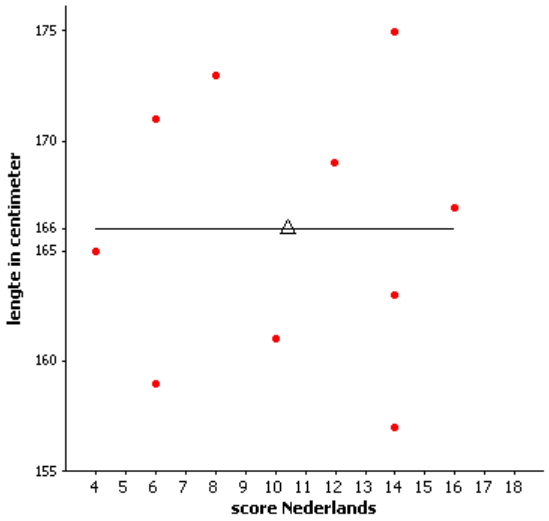
Voor de regressierechte geldt: "Bij een x-waarde die k x-standaardafwijkingen boven (of onder) het x-gemiddelde ligt hoort een $\hat{y}$-waarde die $r \cdot k$ y-standaardafwijkingen boven (of onder) het y-gemiddelde ligt". Hier is $r = 1$. Als je bijvoorbeeld in de x-richting één standaardafwijking $s_x = 0.06$ bij het gemiddelde $\bar{x} = 1.66$ optelt dan krijg je $x = 1.72$. De bijhorende $\hat{y}$-waarde op de regressierechte vind je dan door in de y-richting één standaardafwijking $s_y = 6$ bij het gemiddelde $\bar{y} = 166$ op te tellen, wat $\hat{y} = 172$ oplevert. Dat zie je geïllustreerd op bijgaande figuur.

11.2. Geen lineair verband

Om de lengte (in cm) van dat meisje te voorspellen zou je als verklarende veranderlijke ook eens de punten op het vak Nederlands kunnen nemen. Zou dat helpen?

We zoeken (met de GRM) weer de klassieke kengetallen en tekenen ook een figuur. Bemerk dat je hier een regressie uitvoert van "lengte in cm" over "score op Nederlands".





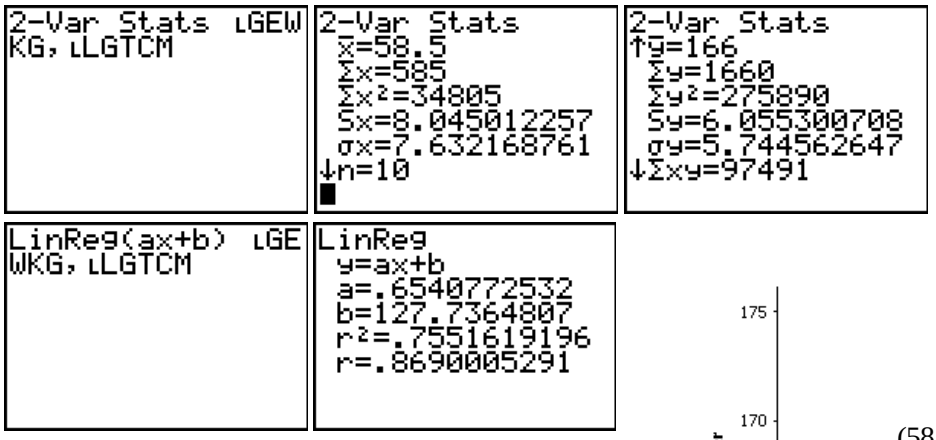
De punten liggen totaal willekeurig verspreid en de correlatiecoëfficiënt r is gelijk aan nul. Er is geen lineair verband tussen de lengte van 17-jarige meisjes en de punten die zij op het vak Nederlands halen.

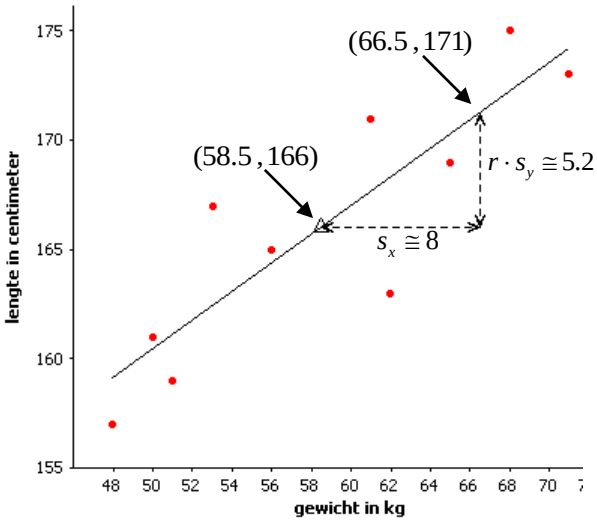
De vergelijking van de regressierechte, die zoals altijd door het zwaartepunt $(\bar{x}, \bar{y}) = (10.4, 166)$ gaat, is hier gelijk aan $\hat{y} = 166$.

Als $r = 0$ dan is de vergelijking van de regressierechte $\hat{y} = \bar{y}$. In deze vergelijking komt de verklarende veranderlijke x niet voor. Informatie over het aantal standaardafwijkingen s_x dat de score op Nederlands van dat meisje boven de gemiddelde score $\bar{x}$ ligt, helpt je niet om voor zo'n meisje je voorspelling dat haar lengte 166 cm is te verbeteren.

11.3. Enig lineair verband

Om de lengte (in cm) te voorspellen kan je ook het gewicht (in kg) als verklarende veranderlijke nemen. Je verwacht dat er "globaal" een samenhang is tussen "groter zijn" en "meer wegen".





De puntenwolk ligt globaal rond een rechte en de correlatiecoëfficiënt r is gelijk aan 0.87. Je hebt hier te maken met een lineair verband dat positief is en tamelijk sterk.

Als je (in de x-richting) bij het gewicht een volledige standaardafwijking $s_x \cong 8$ verder gaat dan het gemiddelde $\bar{x} = 58.5$ dan mag je (in de y-richting) voor de bijhorende verwachte lengte geen volledige standaardafwijking $s_y \cong 6$ toevoegen aan het gemiddelde $\bar{y} = 166$. Je mag maar een fractie van s_y toevoegen en die fractie wordt gegeven door de correlatiecoëfficiënt r. De correlatiecoëfficiënt geeft aan hoe sterk het ***lineaire*** verband is. Bij perfect lineair verband ($r = 1$) mag je in de y-richting een volledige standaardafwijking s_y nemen. Wanneer er helemaal geen lineair verband is ($r = 0$) dan mag je bij $\bar{y}$ niets toevoegen. In de andere gevallen zoals hier neem je $r \cdot s_y$. Dit is geïllustreerd op de bijgaande figuur. Daar is ook de regressierechte getekend waarvan de vergelijking afgerond gegeven is door (GRM):

$$\hat{y} = ax + b \cong 0.65\,x + 127.7$$

Deze vergelijking kan je ook anders schrijven, met behulp van kengetallen:

$$\hat{y} = \bar{y} + r \cdot \left(\frac{x - \bar{x}}{s_x} \right) \cdot s_y \quad of \quad \hat{y} = \bar{y} + 0.87 \cdot \left(\frac{x - 58.5}{8} \right) \cdot s_y \quad of \quad \hat{y} = 166 + 0.87 \cdot \left(\frac{x - 58.5}{8} \right) \cdot 6$$

Je weet dat een punt op de regressierechte je vertelt wat de gemiddelde respons $\hat{y}$ is die hoort bij een bepaalde x-waarde van de verklarende veranderlijke. Dat betekent hier dat de groep van meisjes die allemaal 66.5 kg wegen een lengte hebben die gemiddeld gelijk is aan 171 cm.
Als men je nu zegt dat het gekozen meisje 66.5 kg weegt, wat is dan je beste gok voor haar lengte? Inderdaad, dat is niet meer 166 cm (de gemiddelde lengte van alle 17-jarige meisjes) maar wel 171 cm (de gemiddelde lengte van alle 17-jarige meisjes die 66.5 kg wegen).

DEEL 3. Herhalingsopdrachten

12. Tia Hellebaut

12.1. Olympisch goud

Herhalingsopdracht 1

Sinds de Olympische spelen van 1928 in Amsterdam is hoogspringen voor vrouwen een Olympische discipline. In 2008 won Tia Hellebaut de gouden medaille met een sprong van 2.05 m. Zij is de eerste Belgische vrouw die in atletiek een Olympische medaille behaalt. De resultaten van alle gouden medailles in deze discipline zie je hieronder. Bij het land zie je de Engelstalige afkorting. Tijdens de tweede wereldoorlog werden geen Olympische spelen georganiseerd zodat er geen gegevens zijn voor 1940 en 1944.

Hoogte	Jaar	Naam	Land
1.59	1928	Ethel Catherwood	CAN
1.65	1932	Jean Shiley	USA
1.60	1936	Ibolya Csák	HUN
1.68	1948	Alice Coachman	USA
1.67	1952	Esther Brand	RSA
1.76	1956	Mildred McDaniel	USA
1.85	1960	Iolanda Balas	ROM
1.90	1964	Iolanda Balas	ROM
1.82	1968	Miloslava Rezková	CZE
1.92	1972	Ulrike Meyfarth	FRG
1.93	1976	Rosemarie Ackermann	GDR
1.97	1980	Sara Simeoni	ITA
2.02	1984	Ulrike Meyfarth	FRG
2.03	1988	Louise Ritter	USA
2.02	1992	Heike Henkel	GER
2.05	1996	Stefka Kostadinova	BUL
2.01	2000	Yelena Yelesina	RUS
2.06	2004	Yelena Slesarenko	RUS
2.05	2008	Tia Hellebaut	BEL

Zorg dat je GRM de lijsten HOOG en JAAR bevat. Je kan die gewoon intikken of je kan ze ook halen op http://www.uhasselt.be/lesmateriaal-statistiek .

1. Stel dat je bovenstaande informatie wil gebruiken om te schatten hoe hoog de atlete met de gouden medaille zou gesprongen hebben als er in 1940 wel Olympische spelen hadden plaatsgevonden. Welke veranderlijke kies je dan als verklarende veranderlijke en welke als respons?

2. Gebruik je vorig antwoord om de gepaste puntenwolk te tekenen. Doe dit met je GRM waarbij je ZoomStat gebruikt om automatisch een goede vensterinstelling te krijgen. Ligt de puntenwolk globaal gespreid rond een rechte (behalve enkele punten op het einde – daar komen we op terug in herhalingsopdracht 2)? Is, op zicht, het verband positief of negatief en is het zwak, matig of sterk?

3. Vul in: ik ga een regressie van ……..…… over …….... uitvoeren.

4. Bereken (met GRM) de regressierechte en de correlatiecoëfficiënt. Schrijf, in de juiste notatie, de vergelijking van de gevonden regressierechte. Die is van de vorm $ax+b$. De richtingscoëfficiënt a moet je opschrijven tot op 6 decimalen nauwkeurig. De intercept b schrijf je tot op 2 decimalen. Bevestigt de correlatiecoëfficiënt je antwoord in punt 2? Waarom?

5. Teken (met GRM) gelijktijdig de puntenwolk en de regressierechte. Start met de toets [Y=] en breng de vergelijking in zoals je die hierboven hebt opgesteld. Teken de figuur, ga op de regressierechte staan en tik 1940 en [ENTER]. Welke bijhorende waarde krijg je op de regressierechte? Zeg in woorden wat je nu gevonden hebt.

6. Zou je de gevonden rechte gebruiken om te voorspellen hoe hoog de beste atlete zal springen op de volgende Olympische spelen (hoe hoog zou dat zijn)? Geef een korte motivatie voor je antwoord (een uitgebreide motivatie zie je in herhalingsopdracht 2).

12.2. Een trendbreuk?

Herhalingsopdracht 2

Deze opdracht is een vervolg van herhalingsopdracht 1.

Misschien heb je bij het oplossen van herhalingsopdracht 1 ontdekt dat een voorspelling van de prestatie die de gouden medaille zal neerzetten op de volgende Olympische spelen toch wel hoog uitvalt. Als je goed naar de opmetingen kijkt dan bespeur je een verandering van patroon over de jaren heen. Bij de eerste 10 Olympische spelen (van 1928 tot 1972) ging de grens voor Olympisch goud van 1.59 m naar 1.92 m wat een stijging van 33 cm is. Bij de laatste 10 Olympische spelen ging men van 1.92 m (Ulrike Meyfarth) naar "ongeveer" 2.05 m (2.06 m voor Yelena Slesarenko en 2.05 m voor Tia Hellebaut). Dat is slechts 13 cm. Zou dat op een trendbreuk kunnen wijzen? Als dat zo is dan moet je misschien je analyse aanpassen. Dat kan op verschillende manieren. Hier mag je het eenvoudig houden en met een regressierechte blijven werken maar doe dat eens met alleen maar de resultaten van de laatste 10 spelen.

Gebruik de figuur op de volgende bladzijde om de onderstaande vragen te beantwoorden.

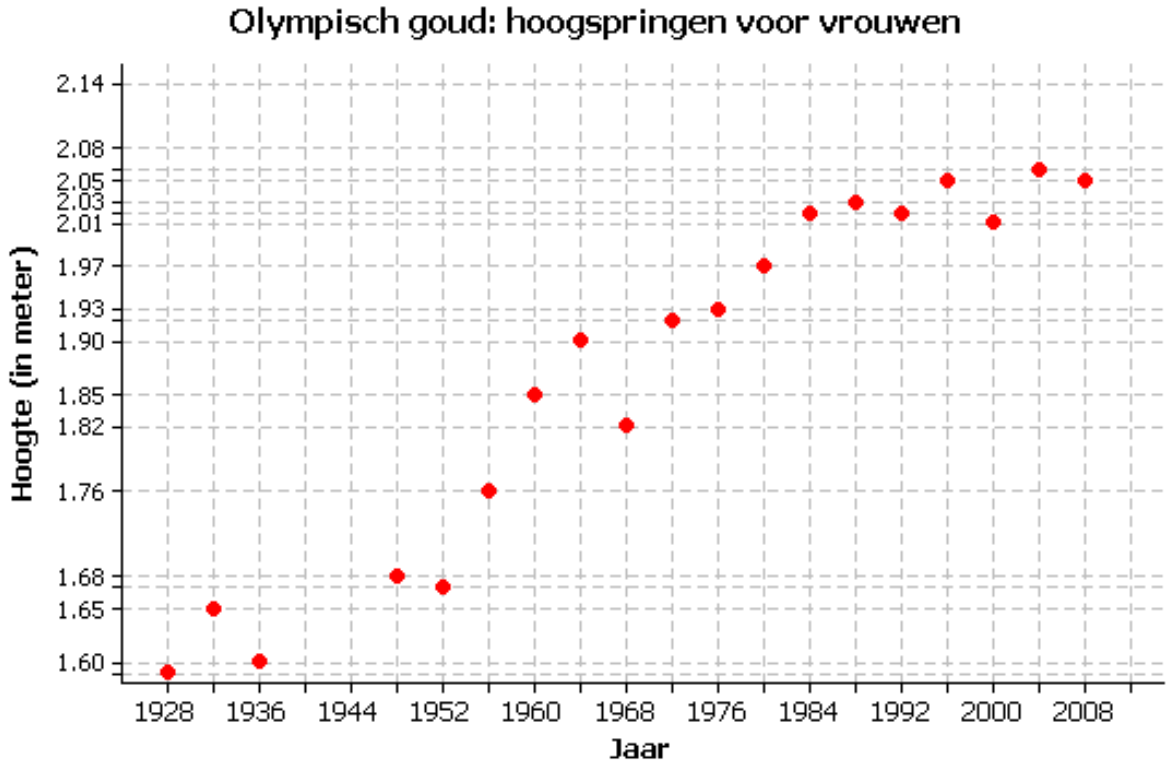
1. Teken de regressierechte gebaseerd op *alle* opmetingen. Neem daarbij x-waarden die lopen van 1928 tot 2012. Schrijf er ook de vergelijking van die rechte bij.

2. Kopieer de lijst JAAR naar [L1] en laat de eerste 9 getallen weg. Kopieer de lijst HOOG naar [L2] en laat ook daar de eerste 9 getallen weg. Gebruik nu de lijsten [L1] en [L2] om de regressierechte te bepalen, gebaseerd op de resultaten van *de laatste 10* Olympische spelen.

3. Teken de nieuwe regressierechte in een ander kleur of in stippellijn. Schrijf er ook haar vergelijking bij. Gebruik dezelfde figuur als daarnet waarbij je nu x-waarden neemt die lopen van 1972 tot 2012.

4. Welke recordhoogte voorspel je nu voor de volgende Olympische spelen? Lijk je dat een logischer antwoord dan wat je in punt 6 van vorige opdracht gevonden hebt? Waarom?

5. Zou je de nieuwe regressierechte gebruiken om te voorspellen hoe hoog men op de Olympische spelen van 2060 zal springen? Hoe hoog zou dat zijn? Geef uitleg bij je antwoord.


Olympisch goud: hoogspringen voor vrouwen
Hoogte (in meter)
2.14
2.08
2.05
2.03
2.01
1.97
1.93
1.90
1.85
1.82
1.76
1.68
1.65
1.60
1928
1936
1944
1952
1960
1968
1976
1984
1992
2000
2008
Jaar

13. Hoe oud is die boom?

13.1. Jaarringen

Herhalingsopdracht 3

Als je de leeftijd van een boom wil weten dan hoef je alleen maar de jaarringen te tellen. Het enige wat je daarvoor moet doen is die boom omzagen….

Als je een boom niet wil omzagen dan moet je een andere manier zoeken om de ouderdom te bepalen. Sommigen boren een kleine cilinder tot in de kern van de stam, halen dat hout eruit en tellen daarop de jaarringen.

Als je de boom helemaal niet wil beschadigen dan meet je gewoon de omtrek van de stam. Je doet dat op een vooraf afgesproken hoogte (dikwijls neemt men een hoogte van 1.30 m). Een verband met de ouderdom is afhankelijk van de boomsoort en van het klimaat. In deze tekst werken we met bomen van eenzelfde soort die onder dezelfde klimatologische omstandigheden zijn gegroeid.

Een onderzoeker heeft bij 22 gevelde bomen nauwkeurig de omtrek van de stam gemeten en de jaarringen geteld. Zijn collega zegt dat je beter met de oppervlakte van de stamdoorsnede werkt. Gelukkig weten biologen ook wel dat je geen boom moet doorzagen om die oppervlakte te kennen. Als je de omtrek ($2\pi r$) deelt door 2π dan heb je de straal r . De oppervlakte vind je dan uit πr^2. De (afgeronde) resultaten zien er als volgt uit:

Gemeten omtrek (cm)	Berekende oppervlakte (cm^2)	Jaarringen= ouderdom (jaar)	Gemeten omtrek (cm)	Berekende oppervlakte (cm^2)	Jaarringen= ouderdom (jaar)
14	16	4	82	535	21
15	18	5	96	733	20
18	26	8	102	828	22
35	97	8	105	877	28
35	97	10	106	894	30
44	154	14	113	1016	34
52	215	8	122	1184	30
59	277	10	123	1204	35
62	306	13	124	1224	38
78	484	16	132	1387	40
81	522	18	130	1345	42

Zorg dat je GRM de lijsten OMTRK, OPPVL en OUDBM bevat. Je kan die halen op http://www.uhasselt.be/lesmateriaal-statistiek .

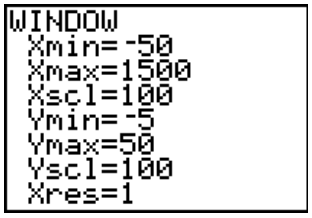
1. Het is de bedoeling van deze studie dat je kan schatten hoe oud een boom is en dat je dit doet door gebruik te maken van een regressierechte van de vorm $\hat{y} = ax + b$.

   Om te beginnen moet jij beslissen of je de ouderdom zal schatten aan de hand van de omtrek van de stam, ofwel of je daarvoor de oppervlakte van een (denkbeeldige) dwarsdoorsnede zal gebruiken. Hoe kom je hier te weten wat je moet kiezen? Doe wat nodig is om een goede keuze te maken en motiveer die keuze. De enige gegevens waarover je beschikt zijn (zoals altijd in de statistiek) de opgemeten data. Zij staan in de tabel hierboven. Vooraleer je de volgende punten (2, 3 en 4) oplost controleer je even met je leerkracht of je de goede keuze gemaakt hebt.

2. Wat is voor jou de respons en wat is de verklarende veranderlijke? Welke regressie ga je uitvoeren: een regressie van …….. over ……..…… (vul in).

3. Voer de regressie uit (met GRM) en schrijf de vergelijking van de regressierechte in de juiste notatie. Hoeveel is de correlatiecoëfficiënt? Bevestigt die wat je "op zicht" gezien had bij die puntenwolk? Wat was dat?

4. Als jij een boom van dezelfde boomsoort zou tegenkomen en je zou voor de opgemeten omtrek 79 cm vinden, hoe oud schat je die boom dan? Zoek dit met je GRM en gebruik de vensterinstellingen zoals aangegeven. Teken de puntenwolk samen met de regressierechte. Ga dan op de regressierechte staan, geef de juiste x-coördinaat in en druk ENTER. Wat is de bijhorende waarde op de regressierechte?



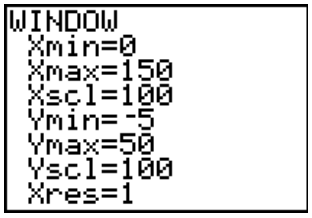
5. Interpreteer het getal dat jij gevonden hebt (de geschatte ouderdom) in het kader van regressiestudies. Wat betekent dat getal eigenlijk? Kijk naar de vroegere studie over vaders en zonen en gebruik wat je daar geleerd hebt om nu een analoge uitleg te geven over de geschatte ouderdom van die boom.

13.2. Residu's

Herhalingsopdracht 4 (facultatief)

Het is niet altijd eenvoudig om in een puntenwolk een patroon te ontdekken. Als je op de y-as de ouderdom van de boom zet en op de x-as de oppervlakte van een doorsnede van de stam op een hoogte van 1.30 m dan lijkt de puntenwolk rond een rechte gespreid. Maar als je op de x-as werkt met de omtrek, dan lijkt de puntenwolk gebogen. Je kan dat iets duidelijker zien wanneer je tegelijkertijd ook de regressierechte tekent die bij die puntenwolk hoort.

1. Teken de puntenwolk en de regressierechte wanneer je de omtrek zou gebruiken om de ouderdom te schatten. Bereken eerst de regressierechte en de correlatiecoëfficiënt. Bevestigt de correlatiecoëfficiënt wat je op zicht waarneemt in de puntenwolk? Welk besluit trek je hieruit? Teken nu je figuur en gebruik daarbij de vensterinstellingen zoals aangegeven. Vergelijk je figuur met de figuur die je in punt 4 van vorige opdracht hebt getekend. Zie je een duidelijk verschil? Welk?



2. Systematische afwijkingen van meetpunten ten opzichte van een gekozen model (hier is dat model de regressierechte) kan je goed bestuderen met "residu's". Een residu $e_i = y_i - \hat{y}_i$ geeft aan hoeveel een opmeting y_i afwijkt van de verwachte modelwaarde $\hat{y}_i$. Hierbij kijk je verticaal (in de richting van de y-as) en houd je rekening met het teken (positief of negatief).

Residu e_i	= opmeting – model
	= geobserveerd – verwacht
	= $y_i - \hat{y}_i$

Als je een goed model hebt, dan verwacht je opmetingen die willekeurig rond het model schommelen en dus residu's die zowel positief als negatief zijn en willekeurig rond nul schommelen. Als de residu's een bepaald patroon vertonen, dan wijst dat erop dat je geen goed model gekozen hebt.

Je GRM berekent automatisch residu's als je een regressierechte bepaalt. Die worden opgeslagen in de lijst RESID. Teken nu een figuur met residu's. Begin met de omtrek als verklarende veranderlijke. Daar heb je zopas mee gewerkt en dus is de lijst RESID die nu in je toestel staat de juiste. Teken de puntenwolk met OMTRK op de x-as en RESID op de y-as (en zet alle functies [Y=] af). Gebruik [ZOOM] en loop naar 9:ZoomStat en druk [ENTER] en dan [TRACE]. Zie je een duidelijk patroon? Welk? Herhaal daarna de hele procedure voor de situatie waarbij je de oppervlakte als verklarende veranderlijke kiest. Is er nu bij de residu's nog een patroon te ontdekken?

14. Tsjirpende krekels

14.1. Hoe warm is het?

Herhalingsopdracht 5

Misschien is je eerste ervaring met krekels, toen je in een tentje in het zuiden van Frankrijk niet in slaap kon geraken door hun doordringend getsjirp, niet zo positief.

Krekels hebben nochtans een wonderbaarlijke eigenschap. De frequentie van hun getsjirp (het aantal tsjirpen per minuut) hangt samen met de temperatuur. Een perfect verband is het niet, maar de Amerikaanse sneeuwboomkrekel (Snowy Tree Cricket = Oecanthus fultoni) is toch behoorlijk goed geijkt. Een regel die je bij die krekel kan gebruiken is als volgt: tel het aantal tsjirpen per minuut, deel dat door 8, tel daar 4.2 bij en dan weet je hoe warm het is (de temperatuur in graden Celsius). Straf hé !

Surf naar http://entomology.ifas.ufl.edu/walker/buzz/585a.htm om het tsjirpen bij verschillende temperaturen te horen.

Om na te gaan of die regel zinvol is en om er een correcte interpretatie aan te geven moet je je onderzoek baseren op reëel feitenmateriaal. Hieronder staan de data die door een onderzoeker in de Verenigde Staten werden opgemeten in 2007.

Geteld aantal tsjirpen per minuut	Opgemeten temperatuur in graden Celsius	Geteld aantal tsjirpen per minuut	Opgemeten temperatuur in graden Celsius
145	21	126	21
101	18	89	17
186	26	174	26
94	15	176	27
70	14	82	16
127	20	102	16
151	24	140	22
116	21	60	12
116	17	148	23
172	25	106	17
140	23	82	13
132	22	132	19
74	11	111	18
62	13	121	19
50	10	50	10
59	11		

Zorg dat je GRM de lijsten TSJRP en TEMP bevat.
Je kan die halen op http://www.uhasselt.be/lesmateriaal-statistiek

1. Wat is hier de onderzoeksvraag? Welke keuze maak je voor de respons en voor de verklarende veranderlijke? Welke regressie ga je uitvoeren?

2. Teken de puntenwolk en gebruik daarbij gepaste vensterinstellingen. Geeft de figuur de indruk dat er een lineair verband is? Is dat positief of negatief, zwak, matig of sterk?

3. Bereken de regressierechte en schrijf ze in de juiste notatie. Bevestigt de correlatiecoëfficiënt de analyse die je op zicht gemaakt hebt? Waarom?

4. Teken de puntenwolk samen met de regressierechte. Liggen de meetpunten willekeurig gespreid rond die rechte of ontdek je een bepaald patroon? Wat betekent dit?

5. Zoek grafisch hoe warm het is als de krekels 2 keer per seconde tsjirpen. Is het dan exact zo warm? Verklaar de betekenis van de temperatuur die je gevonden hebt in het kader van regressiestudies.

6. Is de regel om de temperatuur te bepalen zinvol? Is het een exacte regel? Hoe moet je die begrijpen? Is die regel altijd geldig?

14.2. Residu's

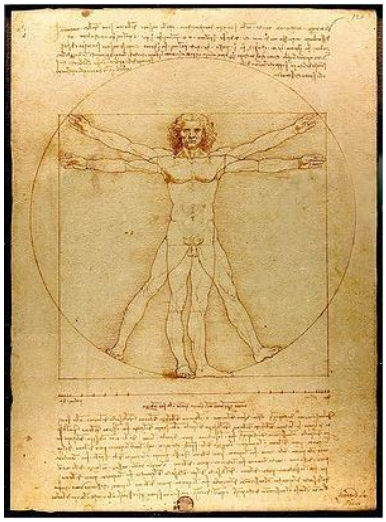
Herhalingsopdracht 6

Facultatief (als je herhalingsopdracht 4 gemaakt hebt): Teken de plot van de residu's. Welke punten worden hier eigenlijk getekend (wat staat er op de horizontale as en wat op de verticale)? Bevestigt de figuur het antwoord dat je in puntje 4 van vorige opdracht gegeven hebt? Waarom?

15. Had Da Vinci gelijk?

15.1. Het verzamelen van de data

Herhalingsopdracht 7



Leonardo Da Vinci was een kunstenaar en een wetenschapper. Hij kende heel veel van anatomie en beschreef allerlei soorten verhoudingen van het menselijk lichaam. Die kennis is handig als je mensen wil schilderen of beeldhouwen. Zo stelde hij ondermeer:

- dat je geknield nog drie kwart van je totale lichaamslengte groot bent
- dat de lengte van je hand een negende is van je lichaamslengte
- dat de spanwijdte van je volledig uitgestrekte armen gelijk is aan je lichaamslengte.

Da Vinci baseerde zich voor die verhoudingen op een "ideaal lichaam" van een volwassen man. Zouden die beweringen ook opgaan voor Vlaamse jongeren van 16 à 18 jaar? Dat ga je nu onderzoeken en je zal een regel proberen op te stellen voor de bewering van Da Vinci dat je uit de spanwijdte van je volledig uitgestrekte armen kan schatten hoe groot je bent.

Om de data te verzamelen moet je eventueel samenwerken met een andere klas van de derde graad zodat je toch bij minstens een vijftiental leerlingen opmetingen kan doen. Vul die in in de onderstaande tabel en breng ze naar de lijsten SPAN en LGT van je GRM. Alle opmetingen doe je tot op een centimeter nauwkeurig.

Volgnr. leerling	Spanwijdte armen	Lengte	Volgnr. leerling	Spanwijdte armen	Lengte
1			11		
2			12		
3			13		
4			14		
5			15		
6			16		
7			17		
8			18		
9			19		
10			20		

15.2. Lengte en spanwijdte

Herhalingsopdracht 8

Deze opdracht sluit aan bij de vorige en werkt met de data die je daar hebt verzameld.

1. Wat is hier de onderzoeksvraag? Welke keuze maak je voor de respons en voor de verklarende veranderlijke? Welke regressie ga je uitvoeren?

2. Teken de puntenwolk en gebruik daarbij ZoomStat om een gepaste vensterinstelling te genereren. Geeft de figuur de indruk dat er een lineair verband is? Is dat positief of negatief, zwak, matig of sterk?

3. Bereken de regressierechte en schrijf ze in de juiste notatie. Bevestigt de correlatiecoëfficiënt de analyse die je op zicht gemaakt hebt? Waarom?

4. Teken nu de regressierechte samen met de puntenwolk. Liggen de punten willekeurig gespreid rond de rechte of ontdek je een bepaald patroon? Wat betekent dit?

5. Zoek grafisch hoe groot iemand is wanneer de spanwijdte van zijn/haar volledig uitgestrekte armen 176 cm is. Verklaar de betekenis van het resultaat dat je gevonden hebt in het kader van regressiestudies.

6. Is jouw gevonden regel om de lengte te bepalen van Vlaamse jongeren van 16 à 18 jaar bruikbaar? Is het een exacte regel? Hoe moet je die begrijpen? Is die regel altijd geldig?

7. (*Moeilijk*) Had Da Vinci gelijk? Leg uit.