J. A. KNOTTNERUS EN A. VOLOVICS

Correlatie en regressie

In deze aflevering komen aan de orde het verband tussen een categorische en een numerieke variabele, en het verband tussen twee numerieke variabelen. Voor het beschrijven van deze verbanden wordt in de regel gebruik gemaakt van het puntendiagram, de correlatiecoëfficiënt en de regressie-analyse. Vooral de correlatiecoëfficiënt wordt veel toegepast in dit tijdschrift.

Inleiding

In de vorige aflevering van deze serie kwam het verband tussen twee categorische (dichotome) variabelen aan de orde. Vaak zijn ook andere verbanden onderwerp van studie:
- het verband tussen een categorische en een numerieke variabele;
- het verband tussen twee numerieke variabelen.

Een voorbeeld van het eerste type is de relatie tussen hoofdpijn (categorische variabele) en hoge bloeddruk (numerieke variabele): van mensen met en zonder hoofdpijn worden de verdelingen van de bloeddrukwaarden vergeleken.[1] Een veel gebruikte maat voor het verband is dan het verschil tussen de gemiddelde waarden. Wil men beoordelen in hoeverre dit verschil al dan niet op toeval kan berusten, dan moet men daarnaast de standaarddeviaties van beide verdelingen kennen. Bij een grote spreiding van beide verdelingen is er immers meer kans dat een bepaald verschil op toeval berust, dan bij een veel kleinere spreiding.

Er kunnen ook meer groepen met elkaar vergeleken worden, bijvoorbeeld: nooit hoofdpijn, minder dan wekelijks hoofdpijn, en iedere week hoofdpijn. Door middel van *variantie-analyse* kan men eventuele verschillen bestuderen en statistisch toetsen.

Rijksuniversiteit Limburg, Postbus 616, 6200 MD Maastricht.
Prof. dr. J. A. Knottnerus, hoogleraar huisartsgeneeskunde, vakgroep Huisartsgeneeskunde; A. Volovics, statisticus, vakgroep Medische Informatica en Statistiek.
Correspondentie: Prof. dr. J.A. Knottnerus.

Voorbeelden van verbanden tussen numerieke variabelen zijn:
- de samenhang tussen de resultaten van de bloeddrukmeting tijdens het spreekuur en de intra-arteriële bloeddruk (thuis rustig zittend);[2]
- de samenhang tussen de uitslagen van verschillende reflectometers voor diabetische zelfcontrole;[3]
- de samenhang tussen het aantal sigaretten dat men rookt en de afname van de vitale capaciteit.

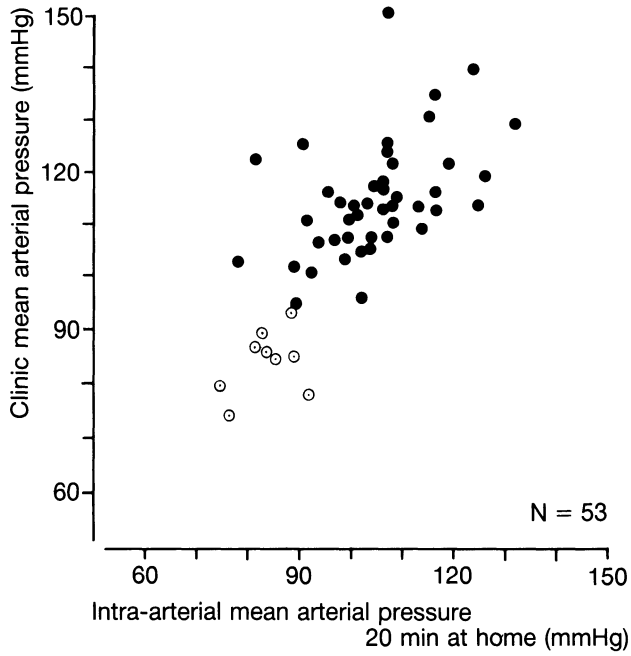
Voor het beschrijven van deze verbanden wordt in de regel gebruik gemaakt van het puntendiagram, de correlatiecoëfficiënt en regressie-analyse. In *Huisarts en Wetenschap* is vooral de correlatiecoëfficiënt erg populair, en het is nuttig wat langer stil te staan bij enkele adders die hier onder het gras liggen. Daarna maken we nog een aantal algemene opmerkingen over regressie.

Correlatie

Figuur 1, ontleend aan *Van Montfrans*, toont het verband tussen de gemiddelde intra-arteriële bloeddruk thuis gemeten (uitgezet op de X-as) en de gemiddelde bloeddruk gemeten op het spreekuur (uitgezet op de Y-as).[2] Onderscheid is gemaakt tussen 44 patienten met lichte hypertensie (●) en negen normotensieve proefpersonen (o). Te zien is dat er in de totale groep een positief verband bestaat tussen beide uitslagen, en dat dit verband bij benadering lineair is. De samenhang is echter lang niet perfect: in de regel is er een verschil tussen beide uitslagen, waarbij de spreekuurbloeddruk in de meeste gevallen hoger is, maar soms ook wat lager.

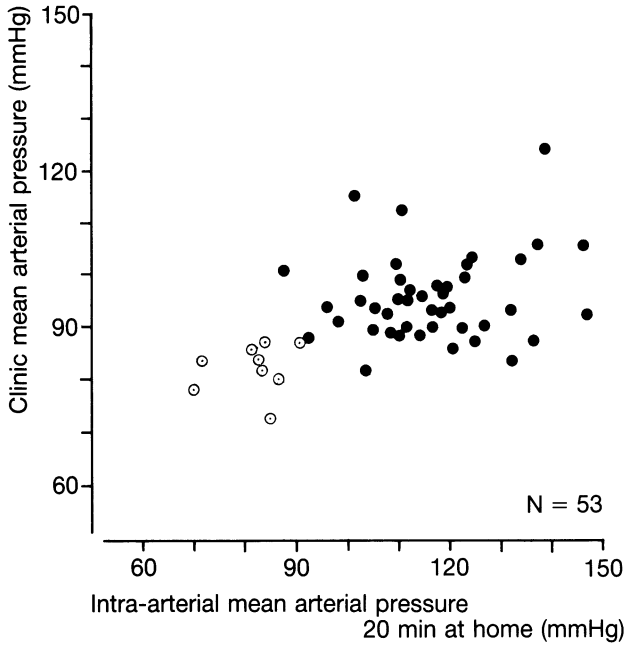
De correlatiecoëfficiënt (volgens Pearson) (r) is nu een indicator van de mate waarin de punten in het puntendiagram *op één rechte lijn* liggen en van de *richting* (positief of negatief) van de samenhang. Als r maximaal is, namelijk gelijk aan +1, is er sprake van een perfecte positieve lineaire correlatie. Als r daalt van 1 tot 0, is er in afnemende mate een niet-perfecte positieve lineaire

Figuur 1 *De spreekuur-bloeddruk uitgezet tegen de intra-arteriële bloeddruk bij de patiënt thuis.*

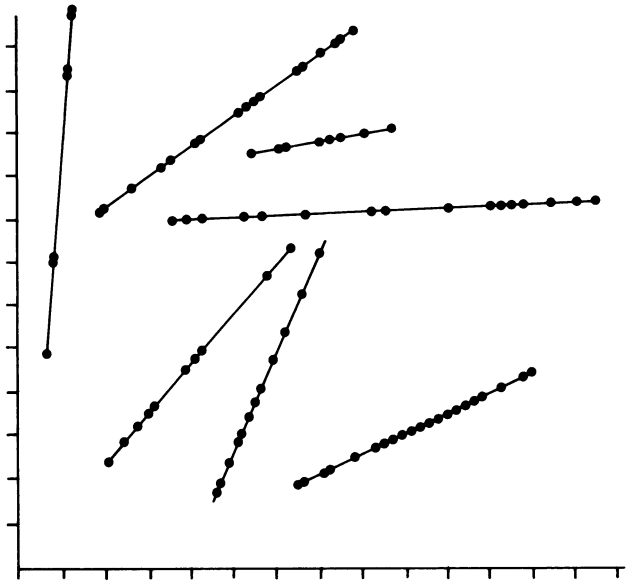


Bron *Van Montfrans.*[2]

Figuur 2 *Fictief puntendiagram, ontstaan door draaiing van de puntenwolk uit figuur 1 naar een minder steile hellingshoek.*



Figuur 3 *Zeven lineaire verbanden met verschillende hellingshoeken.*



correlatie. Een r van 0 betekent dat er in het geheel geen lineair verband is. Een negatieve correlatiecoëfficiënt houdt in dat er in zekere mate sprake is van een negatief lineair verband, met als laagste waarde -1. In het laatste geval liggen alle punten weer precies op één rechte lijn: een perfecte negatieve lineaire correlatie.

Bij de gegevens in *figuur 1* werd een correlatiecoëfficiënt van 0.68 berekend. Dit duidt dus op een positieve correlatie. Wat zegt dit getal ons nu, als we op grond van de spreekuurbloeddruk een uitspraak willen doen over de intra-arteriële druk? Niet meer dan: hoe hoger de spreekuurbloeddruk, des te hoger in het algemeen de intra-arteriële druk, met inachtneming van een nog tamelijk grote spreiding. Dat hadden we ook al uit de puntenwolk opgemaakt. Bovendien gaf die ons nog de extra informatie dat de spreekuurmeting meestal hoger uitvalt. Daarnaast geeft het puntendiagram meer gedetailleerde informatie over de spreiding en, last but not least, ook een indruk van de mate waarin de thuis gemeten intra-arteriële druk gemiddeld hoger is bij patiënten bij wie een hogere spreekuurbloeddruk gemeten wordt.

Over dit laatste, de 'helling' van het verband, laat de correlatiecoëfficiënt echter geen conclusie toe, behalve dan dat de helling positief of negatief is. De correlatiecoëfficiënt is verder slechts een indicator van de mate waarin de punten op één rechte lijn liggen. Zo past bij *figuur 2*, die niets anders is dan een draaiing van de punten uit *figuur 1* naar een minder steile 'helling', ook een correlatiecoëfficiënt van 0.68. En voor alle rechtlijnige verbanden in *figuur 3* geldt, onafhankelijk van de 'helling', een correlatiecoëfficiënt van +1.

Niet alleen de helling van een lineair verband, maar meer in het algemeen de vorm van verbanden kan sterk variëren en toch een zelfde correlatiecoëfficiënt opleveren. Sprekende voorbeelden hiervan zijn de soms bizarre puntendiagrammen in *figuur 4* waarbij telkens een correlatiecoëfficiënt van 0.70 berekend kan worden.[4]

Een ander probleem van de correlatiecoëfficiënt is dat waarden in de buurt van 0 nog niet uitsluiten dat er een verband is, zij het dan geen lineair verband. Een fraai U-vormig verband bijvoorbeeld kan een correlatiecoëficiënt van 0 of praktisch 0 opleveren (*figuur*

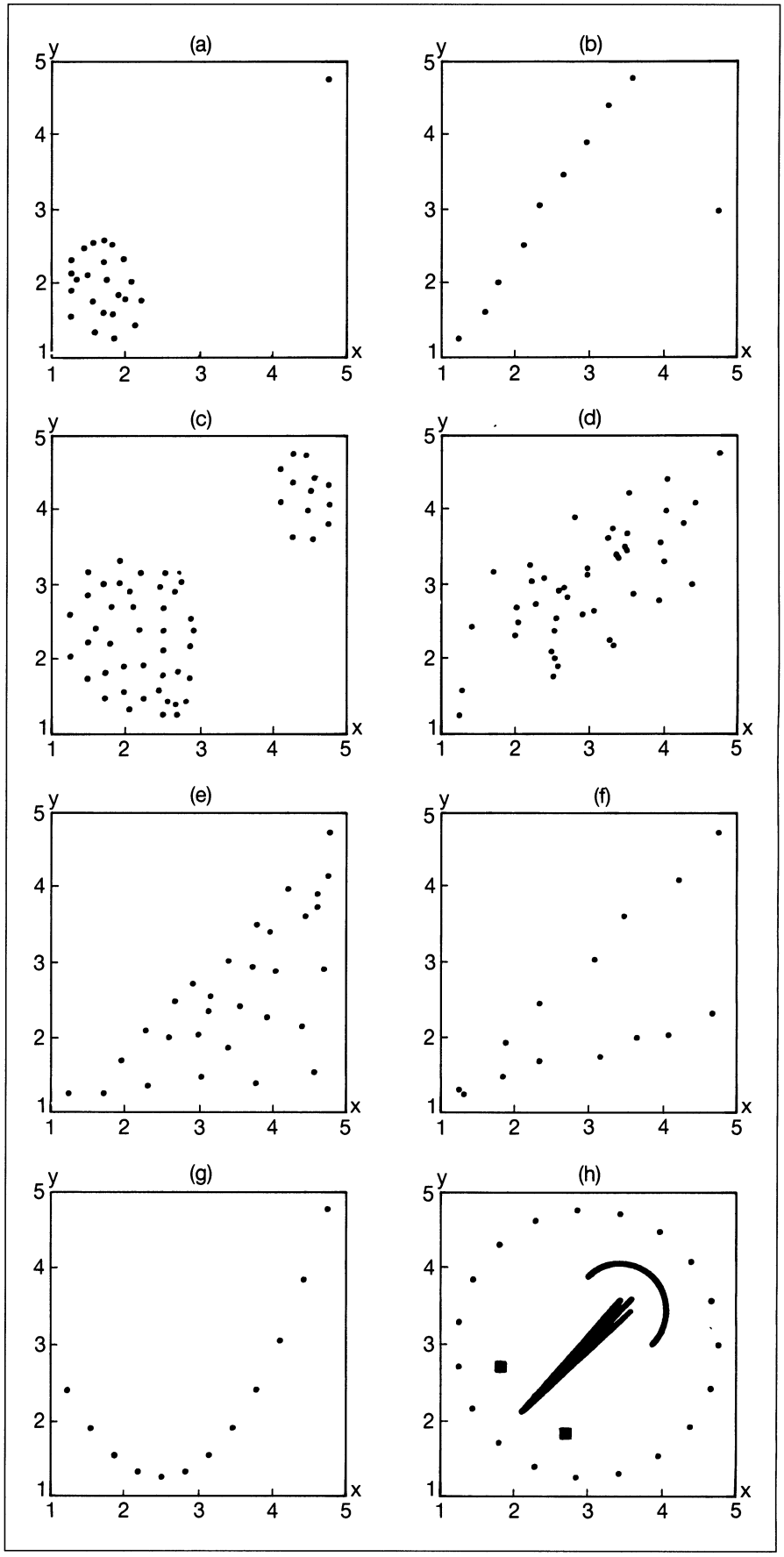
5). Als men op deze waarde zou afgaan, zouden belangrijke relaties over het hoofd gezien kunnen worden.

Tenslotte is de correlatiecoëfficiënt erg gevoelig voor de manier waarop de onderzochte personen zijn geselecteerd. Zo kan men verschillende groepen samenvoegen tot één grote groep, of juist een grote groep in subgroepen onderverdelen, en vervolgens geheel andere waarden van de correlatiecoëfficiënt in de nieuw samengestelde groepen vinden. In *figuur 1* bijvoorbeeld zou men onder normotensieven en hogebloeddrukpatiënten apart veel lagere correlatiecoëfficiënten vinden. En ook in *figuur 4c* is te zien, dat er van noemenswaardige correlatie geen sprake meer zou zijn, als men de beide puntengroepen apart zou beschouwen.

Het komt erop neer dat de correlatiecoëfficiënt op zichzelf van zeer betrekkelijke waarde is. Indien men het puntendiagram heeft bestudeerd en eventueel gepubliceerd of beschreven, kan hij een nuttige samenvattende toevoeging zijn, als er tenminste geen sprake is van een duidelijk niet-lineaire samenhang. De zogenaamde correlatiematrix, een tabel met correlatiecoëfficiënten voor de samenhang van een groter aantal verschillende variabelen, is op zich nauwelijks informatief en soms zelfs misleidend, tenzij men een aanduiding geeft van de vorm van de betreffende verbanden en de wijze van selectie van de onderzochte groep nauwkeurig beschrijft.[6] Het is al met al niet mogelijk om de grootte van de correlatiecoëfficiënt eenduidig te interpreteren als men het bijbehorende puntendiagram niet kent: verbanden van geheel verschillende aard kunnen dezelfde correlatiecoëfficiënt opleveren. Volstrekt niet-lineaire verbanden kunnen nog redelijk hoge waarden opleveren (*figuur 4*), en zeer lage waarden sluiten een fraai, niet-lineair verband niet uit. Sterke argumenten voor een lineair verband heeft men pas bij waarden boven 0.80 à 0.85.

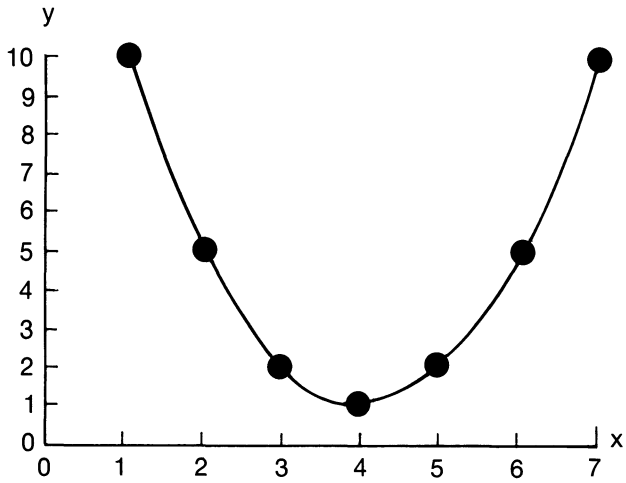
Regressie

Bij regressie-analyse is het doel om te komen tot een beschrijving van de mate waarin een bepaalde (afhankelijke) va-



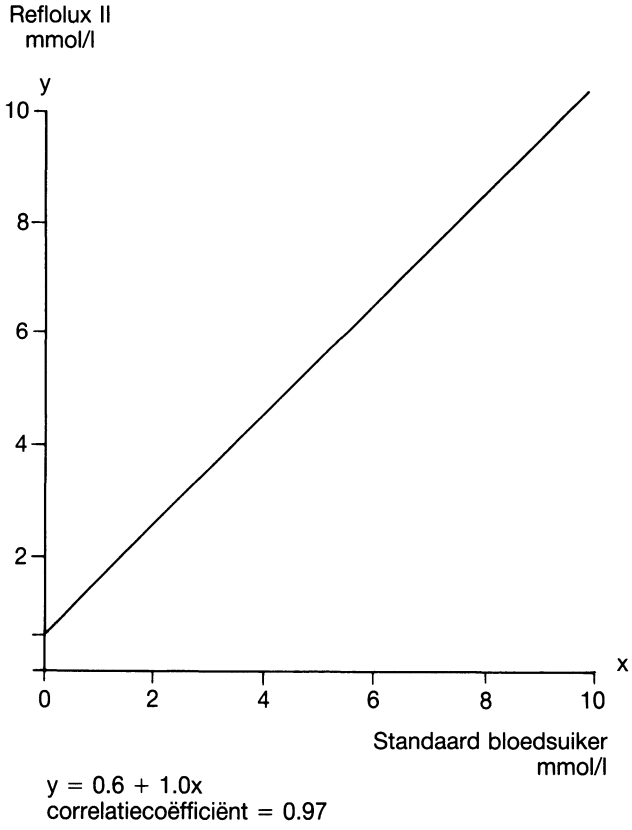
Figuur 4 *Acht puntendiagrammen, elk met een correlatiecoëfficiënt van 0.70.*[4]

Figuur 5 *U-vormig verband met een correlatiecoëfficiënt van 0.0.*



Bron *Colton.*[5]

Figuur 6 *Lineaire regressiefunctie voor het verband tussen bloedsuiker bepaald door de Reflolux II-meter, en de standaardbloedsuiker.*



Bron *Dujardin e.a.*[3]

riabele verandert, indien één of meer andere (onafhankelijke) variabelen andere waarden aannemen. Deze beschrijving heeft de vorm van een algebraïsche functie, en dient goed te passen bij de voor analyse beschikbare empirische gegevens. Het bepalen van zo'n passende functie noemt men wel het 'fitten' van een regressiefunctie of -lijn.

Regressie-analyse kent verschillende toepassingen:

- In de eerste plaats het beschrijven van de aanwezigheid en de vorm van bepaalde relaties. Een voorbeeld hiervan is de beschrijving van het verband tussen de leeftijd en de kans op een somatische diagnose in aansluiting op onverklaarde moeheidsklachten.[7] In dit kader is ook te noemen de beschrijving van het mogelijke dosis-respons effect in onderzoek naar een etiologisch verband of de effectiviteit van een geneesmiddel. Een voorbeeld is het aantal grammen dat een neonaat gemiddeld minder weegt per door de moeder tijdens de zwangerschap gerookte sigaret.
- In de tweede plaats kan men beogen op grond van de afgeleide functie (het regressiemodel) voorspellingen te doen ten aanzien van de afhankelijke (respons- of outcome-)variabele in gevallen waarin men alleen de onafhankelijke variabelen (predictoren of determinanten) kent. Een voorbeeld hiervan bevat het recente artikel van *Dujardin e.a.*, waarin wordt onderzocht of reflectometers voor diabetische zelfcontrole een nauwkeurige schatting kunnen leveren van de bloedsuikerwaarde die verwacht mag worden bij meting in een klinisch-chemisch laboratorium.[3] Voorwaarde voor zo'n voorspelling is dan wel dat de populatie op grond waarvan de regressiefunctie werd afgeleid, voldoende representatief is voor de groep of de personen waarbij de 'voorspelling' gedaan wordt.

De regressiefunctie wordt in de regel bepaald op grond van de resultaten van een onderzoek. Als er bij bestudering van het puntendiagram geen aanleiding is een niet-lineair verband te vermoeden, dan kan in principe gekozen worden voor het 'lineair model'. In de meest eenvoudige vorm hiervan is slechts sprake van één afhankelijke (respons-)variabele y en één onafhankelijke variabele (predictor) x:

$$y = A + Bx$$

Dit wil zeggen: in de gevallen dat er sprake is van één zelfde waarde van x, is y het verwachte gemiddelde van de bijbehorende waarden van y. Een exacte voorspelling van y is uiteraard niet mogelijk. A is een constante, die nodig is omdat bij een x-waarde van 0 niet per se een y-waarde van 0 hoort. B is de regressiecoëfficiënt, die aangeeft hoeveel y verandert per eenheid van x. In *figuur 6* wordt een voorbeeld gegeven van een lineaire regressiefunctie, ontleend aan *Dujardin e.a.*[3] Volgens zo'n functie is door een puntenwolk een rechte lijn te trekken die voor iedere waarde van x telkens de schatting voor y aangeeft. De regressiecoëfficiënt B bepaalt de 'helling' van de lijn, en deze kan weer positief of negatief zijn. A en B volgen uit de regressie-analyse (met sommige pocketcalculators en zeker per computer snel uit te voeren).

Bij de regressie-analyse is een probleem, dat men in een onderzoek in het algemeen te maken heeft met meetonnauwkeurigheid, biologische variatie en variaties ten gevolge van de selectie van een bepaalde onderzoeksgroep. Opnieuw uitvoeren van de regressie-analyse op nieuwe gegevens, op een ander tijdstip of bij andere vergelijkbare personen verkregen, zal dan ook meestal tot op zekere hoogte andere waarden voor A en B opleveren. Het verdient dan ook aanbeveling om met behulp van statistische procedures, zoals significantietoetsen of een bepaling van betrouwbaarheidsintervallen, na te gaan in hoeverre de regressielijn de waarnemingen adequaat beschrijft en hoe nauwkeurig en betrouwbaar uitspraken en voorspellingen op grond van de regressielijn zijn.

Grootheden die daarbij gebruikt worden, zijn de standaarddeviatie van de afwijkingen van de waarnemingen ten opzichte van de regressielijn en de daaruit afgeleide 'standaardfout' van de coëfficiënten (SE = standard error). Buiten de range van de in het onderzoek waargenomen waarden van x, bij hantering van andere meetmethoden of bij anders samengestelde onderzoeksgroepen is zo'n statistische analyse echter evenmin een garantie voor de geldigheid van de gevonden lineaire relatie.

Op het grondmotief van de meest eenvoudige vorm van lineaire regressie bestaan allerlei variaties.

Als het puntendiagram op een niet-lineaire samenhang duidt, moet gezocht worden naar andere functies. Bij het vermoeden van een U-vormig verband kan bijvoorbeeld een kwadratische functie de meest passende zijn:

$$y = A + B_1x + B_2x^2$$

Hierbij is, naast de constante A, sprake van twee regressiecoëfficiënten: één voor de lineaire en één voor de kwadratische 'component'. Indien men bestudeert in hoeverre y wordt beïnvloed door meer dan één onafhankelijke variabele tegelijk, kan men deze ook tegelijk in de regressievergelijking brengen, bijvoorbeeld:

$$y = A + B_1x_1 + B_2x_2 + \ldots + B_ix_i.$$

Men spreekt dan van multivariate regressie, en in dit geval van multivariate lineaire regressie, omdat in het rechterlid alleen lineaire componenten voorkomen.

De voordelen van de regressiefunctie ten opzichte van de correlatiecoëfficiënt bij het beschrijven van het verband tussen twee variabelen zijn:

- De regressiefunctie is gemakkelijk te interpreteren: de vorm van het verband is beschreven en met behulp van de functie is uit de onafhankelijke variabele de gemiddelde waarde van de afhankelijke te schatten. Bij correlatieberekening komt zo'n schatting in het geheel niet aan de orde. Uit *figuur 6* volgt bijvoorbeeld dat de Reflolux II reflectometer de beste schatting van de gemiddelde bloedsuiker (y) geeft, als men bij iedere gemeten waarde (x) 0.6 mmol/l (A) optelt. Het feit dat de correlatiecoëfficiënt zeer hoog is (0.97), geeft alleen aan dat de relatie redelijk nauwkeurig door een rechte lijn is te beschrijven.
- De regressiefunctie is niet per se gebonden aan lineaire relaties, en kan andere vormen beschrijven. De correlatiecoëfficiënt heeft dan echter geen betekenis meer.

Verband en oorzaak

Het is wellicht overbodig om te zeggen dat het vinden van een verband nog niet zegt dat er sprake is van bijvoorbeeld een causale relatie, laat staan over de richting daarvan. Bij etiologische vraagstellingen is er voor de bewijsvoering meer nodig dan een gevonden verband. Met name de aard en de kwaliteit van de studieopzet, maar ook de reeds tevoren bestaande kennis en een zorgvuldige interpretatie van de uitkomsten behoren hiertoe. Wie zou intensivering van het roken willen bepleiten op grond van de waarneming dat de toename van het tabaksgebruik in Nederland gepaard is gegaan met een verbeterde levensverwachting?

[1] Knipschild PG, Visser A, Knottnerus JA. Hoofdpijn en hoge bloeddruk. Huisarts Wet 1984; 27: 441-5.

[2] Van Montfrans GA. Spreekkamer bloeddrukmeting 2 [Ingezonden]. Huisarts Wet 1985; 28: 139.

[3] Dujardin BCG, Roijers AFM, Schoorl M. Laboratoriumonderzoek van reflectometers voor diabetische zelfcontrole. Huisarts Wet 1987; 30: 207-9.

[4] Chambers JM, Cleveland WS, Kleiner B, Tukey PA. Graphical methods for data analysis. 3rd ed. Belmont: Duxbury Press, 1987.

[5] Colton T. Statistics in medicine. Boston: Little, Brown and Company, 1974.

[6] Armitage P. Statistical methods in medical research. New York: Wiley, 1971.

[7] Knottnerus JA, Starmans R, Vissers A. Diagnostische conclusies naar aanleiding van onverklaarde moeheid. Huisarts Wet 1987; 30: 9-12.