Statistiek voor tandartsen (VI)

Regressie en correlatie

A. H. B. Schuurs, tandarts[1]
H. J. Duivenvoorden, methodoloog[2]
M. A. van 't Hof, statisticus[3]

Uit [1]de vakgroep Cariologie en Endodontologie van het Academisch Centrum Tandheelkunde Amsterdam (ACTA), [2]de vakgroep Medische Psychologie en Psychotherapie van de Faculteit der Geneeskunde, Erasmus Universiteit te Rotterdam en [3]de Medisch Statistische Afdeling van de Faculteit der Medische Wetenschappen, Tandheelkunde van de Katholieke Universiteit te Nijmegen.

Trefwoorden: Statistiek

Datum van acceptatie: 22 januari 1992.

Adres: Dr. A.H.B. Schuurs, ACTA, Louwesweg 1, 1066 EA Amsterdam.

Samenvatting. In dit artikel wordt ingegaan op de betekenis en het toepassingsgebied van correlatiecoëfficiënten en wordt het begrip 'regressie' aan de hand van voorbeelden verduidelijkt. Een aantal kernbegrippen, zoals 'intercept', helling, onverklaarde variantie, scattergram, Pearson-toets en causale relaties, worden besproken. Besloten wordt met enkele praktische aanwijzingen.

SCHUURS AHB, DUIVENVOORDEN HJ, VAN 'T HOF MA. Statistiek voor tandartsen (VI). Regressie en correlatie. Ned Tijdschr Tandheelkd 1992; 99: 127-30.

1 Inleiding

Regressie en correlatie zijn 'broertje en zusje'. Met deze twee begrippen uit de statistiek krijgen we te maken als de relatie tussen twee variabelen moet worden beschreven. In feite gaat het om rechtlijnige verbanden tussen die variabelen. Als er sprake is van een kromlijnig verband (bijvoorbeeld de relatie tussen het aantal natuurlijke gebitselementen en leeftijd), is de gewone regressie-analyse niet zo efficiënt.

De rechte lijn, in het rechtlijnige verband, wordt vastgesteld met regressie-analyse. De correlatie (een getal) geeft aan hoe goed de waarnemingen aansluiten bij de gevonden rechte lijn.

Naast het beschrijven van de samenhang tussen twee variabelen kan de regressie-analyse ook worden gebruikt bij 'voorspellen'. De term 'voorspellen' houdt hier geen blik op de toekomst in; daarvoor is het woord 'prognose' beter op zijn plaats. Bedoeld wordt dat als de relatie tussen de variabelen X en Y bekend is, men op grond van een bepaalde waarde van X de bijbehorende waarde van Y kent, zonder deze te meten.

2 Voorspellen

In het tandheelkundige onderwijs zou op grond van het cijfer voor het morfologietentamen (X) kunnen worden voorspeld hoe goed de student in staat is om de tandvorm te modelleren (cijfer Y). Daartoe moet als eerste de relatie tussen de cijfers X en Y, die in het verleden zijn behaald, worden bepaald. Vervolgens kan het resultaat van die analyse worden gebruikt om te voorspellen.

De zin van het vaststellen van dergelijke relaties via een regressielijn is tweeërlei. Ten eerste is het vanuit theoretisch oogpunt prettig te weten dat kennis en praktijk met elkaar te maken hebben. Ten tweede zou er een praktische consequentie kunnen zijn. Men zou bijvoorbeeld kunnen overwegen studenten met een hoog cijfer voor kennis van de morfologie vrij te stellen van het practicum, maar dan moet de samenhang tussen beide vakken wel zeer sterk zijn.[1-3]

3 Regressie-analyse

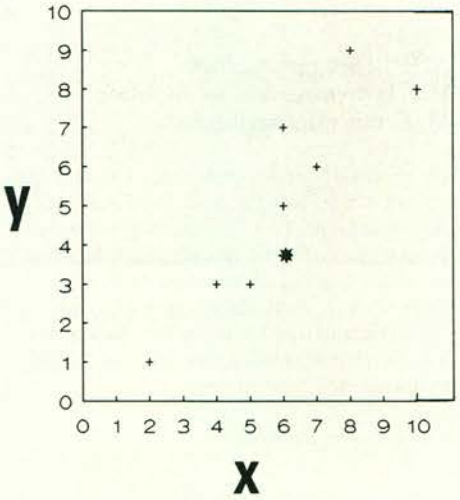
In tabel I lijken de cijfers X en Y per student min of meer 'samen te vallen', het-

Tabel I. Cijfers van 10 studenten voor morfologiekennis, het practicum modelleren en practicum scheikunde.

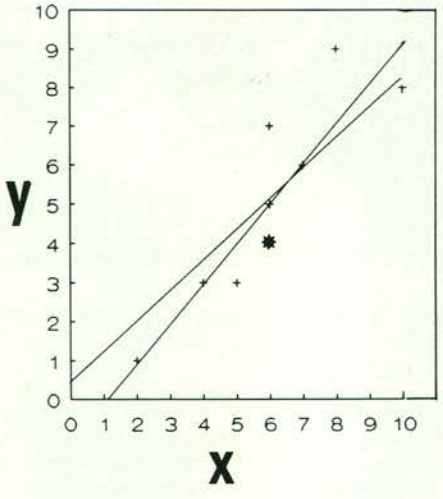
	Kennis morfologie X	Practicum modelleren Y	Practicum Scheikunde Z
Student			
A	4	3	9
B	7	6	4
C	8	9	3
D	6	7	4
E	6	5	8
F	6	4	10
G	2	1	5
H	10	8	4
I	5	3	9
J	6	4	4
Gemiddelde	6,0	5,0	6,0
sd	2,2	2,5	2,9

Tabel II. Kritieke benedenwaarden (afgerond op twee decimalen) van Pearson's r en van Spearman's rho voor $\alpha = 0{,}05$ (tweezijdige toetsing).

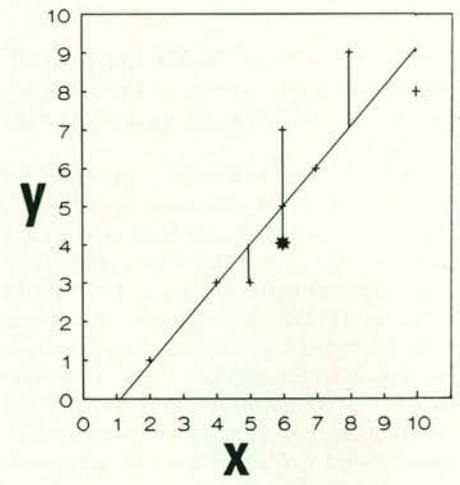
	r	rho
N		
5	0,88	1,00
6	0,81	0,89
7	0,75	0,79
8	0,71	0,74
9	0,67	0,68
10	0,63	0,65
15	0,51	0,52
20	0,44	0,45
25	0,40	0,40
30	0,36	0,36
50	0,28	0,28
100	0,20	0,20



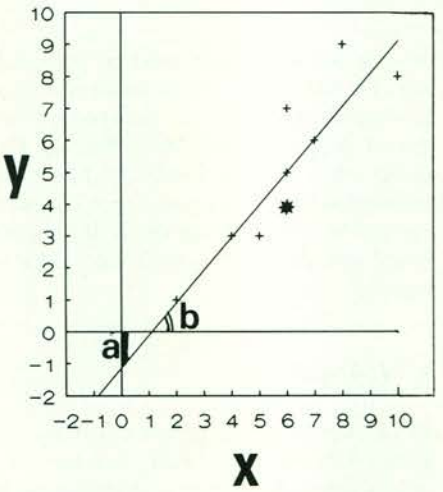
Afb. 1a. Spreidingsdiagram van de cijfers (tab. I) voor morfologiekennis en het practicum modelleren (* = twee samenvallende punten).



Afb. 1b. Welke lijn representeert de tien punten het beste? Er zijn veel meer 'passende' lijnen denkbaar (* = twee samenvallende punten).



Afb. 1c. De regressielijn loopt zodanig dat de gemiddelde gekwadrateerde verticale afstand van alle punten tot de lijn zo klein mogelijk is (* = twee samenvallende punten). De verticale lijnen geven de residuen aan.



Afb. 1d. Nogmaals de regressielijn voor de morfologiekennis en modelleren (tab. I): a = intercept (vet weergegeven deel van de verticale lijn); b = helling.

geen gemakkelijker is te zien door ze in een 'spreidingsdiagram'(scattergram) uit te zetten (afb. 1a); voor elke student werd op de X-as het cijfer voor de morfologiekennis (de onafhankelijke, voorspellende variabele) uitgezet en op de Y-as dat van de te voorspellen (afhankelijke) variabele, het practicumcijfer.

Als iedere student voor beide vakken identieke cijfers zou hebben behaald, loopt een rechte lijn onder een hoek van 45° met de X-as door alle punten (de coördinaten van X en Y). Afbeelding 1a laat zien dat het verband tussen X en Y redelijk rechtlijnig (lineair) is. Maar welke rechte lijn we ook tekenen (afb. 1b), er zullen altijd een aantal punten boven en onder de lijn liggen. De vraag is dus waar en hoe die lijn het beste kan worden getekend. Het ligt voor de hand zodanig te 'schipperen', dat de verticale afstand van ieder punt tot de lijn zo klein mogelijk is (afb. 1c). Een bekende formule voor de rechte lijn in een tweeassenstelsel is: $Y = a + b \times X$

Bovengenoemde formule voorspelt het meest waarschijnlijke verloop van de (gemiddelde) scores op variable Y, op grond van de scores op X. Men noemt a en b de regressiecoëfficiënten en de lijn 'regressielijn'. De letter a geeft de waarde weer van Y voor X = 0, het zogenoemde 'intercept' (afb. 1d). Coëfficiënt b geeft de helling van de regressielijn weer. De waarde van b vertelt met hoeveel eenheden Y toeneemt voor elke eenheid die X groter wordt. Als b positief is, is de waarde van Y groter naarmate X groter is: de lijn stijgt. Is b negatief, dan daalt de lijn. Achterhaald moet worden welke waarden a en b hebben, want dan kan de best passende lijn worden getrokken.

3.1 Berekening regressiecoëfficiënten a en b

Zoals zo vaak in de statistiek wordt ook hier gewerkt met kwadraten van de verschillen (van de punten tot de lijn, de zogenoemde residuen; afb. 1c). Door met de kwadraten van deze residuen te werken, wordt voorkomen dat de pluswaarden (de punten boven de lijn) de minwaarden (de punten onder de lijn) opheffen. De best passende lijn zal de gemiddelde kwadratische afstand tot de lijn zo klein mogelijk maken (kleinste kwadratenmethode). De berekening van a en b geschiedt meestal met computerprogramma's en daarom lijkt presentatie van de daartoe gebruikte formules overbodig. Invoering van de gegevens van tabel I in zulk een statistisch programma levert op:

$a = -1{,}14$ en: $b = 1{,}02$

Voor de voorspelde cijfers Y (aangeduid met $\hat{Y}$) geldt derhalve:

$\hat{Y} = a + b \times X = -1{,}14 + 1{,}02 \times X$, ofwel:

geschatte cijfer modelleren $= -1{,}14 + 1{,}02 \times$ cijfer morfologie.

Voor twee waarden van X zijn de bijbehorende, voorspelde waarden van Y berekend:

$X = 1 \rightarrow \hat{Y} = -1{,}14 + 1{,}02 \times 1 = -0{,}12$
$X = 10 \rightarrow \hat{Y} = -1{,}14 + 1.02 \times 10 = 9{,}16$

Door deze twee berekende waarden in het assenstelsel uit te zetten en met elkaar te verbinden, kan de regressielijn worden getekend (afb.2b). Als de lijn goed is berekend, loopt hij door het punt $\bar{X}$ en $\bar{Y}$. De gemiddelde score $\bar{X} = 6$. Dus $\bar{Y} = -1{,}14 + 1{,}02 \times 6 = 5{,}98$ hetgeen afgerond inderdaad gelijk is aan $\bar{Y}$ (zie tab. I).

Voor een student die het cijfer 3,5 voor morfologie heeft behaald, is nu met de regressielijn te voorspellen dat hij een cijfer voor modelleren zal behalen van $-1{,}14 + 1{,}02 \times 3{,}5 = 2{,}4$. Of dit een goede voorspelling is, hangt af van de hoogte van de correlatiecoëfficiënt.

4 Correlatie

De relatie tussen de waarden van twee variabelen is in één getal samen te vatten, de zogenoemde Pearson's correlatie, aangeduid met r. [1-3]

4.1 Pearson's correlatie (r)

4.1.1 Berekening

Pearson's r kan met verschillende formules, die tot elkaar zijn te herleiden, worden berekend. In essentie wordt daarbij uitgegaan van het gemiddelde produkt van de afwijkingen van de waarden van X en Y ten opzicht van hun gemiddelde. Zulk een afwijking wordt ook wel 'moment' genoemd. Het woord moment slaat niet op de tijd, maar stamt uit de mechanica (denk aan het moment van een kracht).

Er bestaan computerprogramma's waarmee r kan worden berekend, reden om ook deze formule(s) hier weg te laten. De coefficiënt r voor de variabelen morfologiekennis en modelleren (tab. I) blijkt te zijn: r = 0,89. Deze positieve correlatie houdt in dat naarmate het cijfer voor morfologiekennis hoger is, het punt voor het modelleren ook hoger is, en omgekeerd.

Berekening van r voor het modelleren en scheikunde (tab. I) levert de waarde r = − 0,57 op, een negatieve en minder hoge correlatie. Het minteken duidt erop dat naarmate het cijfer voor het practicum hoger is, het cijfer voor scheikunde lager is, en omgekeerd. De lagere correlatie met scheikunde impliceert dat het cijfer voor modelleren minder goed te voorspellen is uit het cijfer voor scheikunde.

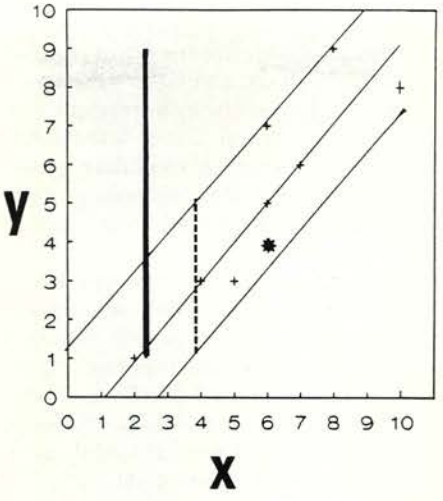
4.1.2 Significantie van r

In veel gevallen is het interessant om na te gaan of het verband tussen twee variabelen significant is. Hiervoor is de Pearson-correlatietoets onwikkeld. Getoets wordt of de correlatiecoëfficiënt r in de populatie al dan niet gelijk is aan nul. De toets kan worden uitgevoerd met behulp van de berekende correlatie en het aantal paren waarnemingen.

Er bestaan tabellen waarin de kritieke waarden r voor een gegeven aantal punten (N) kan worden afgelezen. Dergelijke tabellen bevatten voor N = 5... 100 (en meer) de kritieke waarden van r voor bijvoorbeeld $\alpha = 0,05$. Tabel II bevat een vereenvoudigde significantietabel voor r. Voor het voorbeeld r = 0,85 en N = 10 vindt men achter N = 10 in de eerste kolom van tabel II de waarde 0,63; deze waarde moet r ten minste hebben om significant te zijn op het niveau van $\alpha = 0,05$ (bij tweezijdige toetsing; een minteken mag worden genegeerd voor de bepaling van de significantie). In ons voorbeeld geldt dat de correlatie r = 0,89 significant is en r = − 0,57 niet. De Pearson-correlatietoets is terug te brengen op de t-toets. Deze zal in een later artikel worden besproken.

4.1.3 Een hoge, significante waarde van r

In de praktijk wordt wel als vuistregel gehanteerd dat r ten minste 0,40 moet zijn om



Afb. 2. De totale spreidingsbreedte van Y (vette, ononderbroken lijn) is groter dan de spreidingsbreedte van de waarden Y rond de regressielijn (korte gestippelde lijn).

enige waarde te hebben, maar er is pas sprake van een sterke correlatie als r een waarde van 0,80 nadert.[1] Toch is het moeilijk om in het algemeen aan te geven wat een 'goede' correlatie is.

Om r te interpreteren het volgende. De waarde van r kan maximaal +1 zijn (de punten liggen dan precies op een perfect stijgende lijn) en minimaal −1 (perfect dalende lijn). Een r = 0,89 duidt op een weliswaar hoge, maar niet perfecte correlatie. Wat betekent r = 0,89 nu? In ieder geval niet dat er een 89% samenhang aanwezig is.

Het kwadraat van r is gemakkelijker te interpreteren. Als r^2 wordt vermenigvuldigd met 100, verkrijgt men het percentage van de verklaarde variantie. Voor de correlatie morfologie-modelleren geldt: $0,89^2 \times 100 = 79\%$ verklaarde variantie. De term 'verklaard' slaat op 'verklaard door de andere variabele' en houdt geen 'inhoudelijke' verklaring in. De spreiding van de waarden Y rond de regressielijn is namelijk kleiner dan de spreiding rond het gemiddelde van Y (afb. 2). De spreiding rond het gemiddelde wordt 'totale spreiding' genoemd en de spreiding rond de regressielijn heet ook wel 'onverklaarde (of residuele) spreiding'. Voor het practicumcijfer geldt:

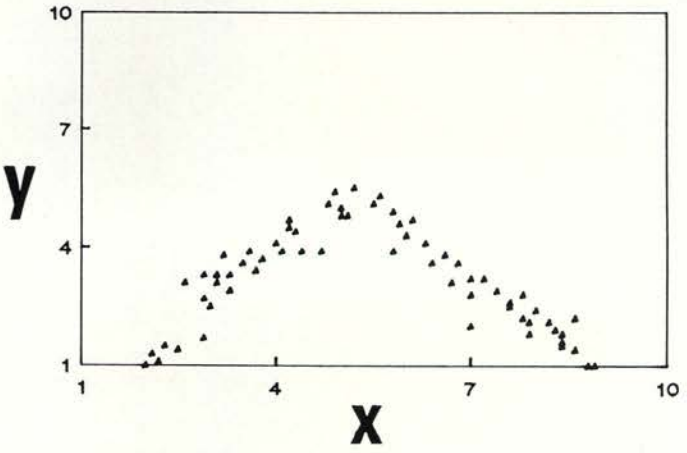
totale variantie = 6,2 (= $2,5^2$; zie tab. I)
onverklaarde variantie = 1,5 (geschat uit afb. 2)

Het verschil tussen de totale en onverklaarde variantie is 6,2 − 1,5 = 4,7. De verklaarde variantie is $4,7/6,2 \times 100 = 79\%$. De verklaarde variantie is niet eenvoudig uit afbeelding 2 af te lezen. Voor praktisch gebruik in de dagelijkse praktijk moet r toch wel 0,90 zijn.

4.1.4 Een hoge of lage r

In wetenschappelijk, explorerend onderzoek kan echter een lage r van bijvoorbeeld 0,25 van belang zijn. Een lage correlatie houdt niet per se in dat de relatie oninteressant zou zijn; zulk een lage waarde kan bijdragen aan nieuwe theorievorming.[4]

Een kromlijnig verband tussen twee variabelen, hetgeen in een scattergram (afb. 3) tot uiting komt, resulteert in een lage r, terwijl de samenhang in feite sterk kan zijn. r^2 is in zulke gevallen een onderschatting van de verklaarde variantie. Een correlatie is alleen dan goed interpreteerbaar als er sprake is van een rechtlijnig verband tussen de variabelen. In een scattergram zal dan een ellips vormige verzameling van de punten te zien zijn: in zulke gevallen zal een rechtlijning verband aanwezig zijn en slechts dan geeft de correlatie coëfficiënt een juiste indruk van de samenhang tussen de variabelen. (Bij twijfel aan een rechtlijnig verband kan men met kunstgrepen toch een lineaire regressielijn vinden; in plaats van met de waarden van Y zelf te werken, kan bijvoorbeeld de logaritme van Y worden genomen.)



Afb. 3. De 'puntenwolk' laat een kromlijnig (curvilineair) verband tussen variabelen X en Y zien, bijvoorbeeld de correlatie tussen werking van een geneesmiddel tegen de dosis afgezet (X= dosis, Y = welbevinden).

4.2 Andere correlatiecoëfficiënten

Pearson's r mag worden berekend voor ordinaal en op hoger niveau gemeten variabelen, mits voldaan is aan enkele eisen, waaronder een normale verdeling (geen uitbijters).[5] In geval van steekproeven met uitschieters kan de nonparametrische correlatiecoëfficiënt (rho van Spearman) worden berekend. Deze is gebaseerd op het toekennen van rangnummers, waardoor de invloed van uitschieters wordt bijgetrokken (een extreem hoge waarde krijgt dan weliswaar een hoog, maar wel aansluitend rangnummer, en hoge rangnummers moeten er toch zijn). De significantie van rho wordt opgezocht in de Spearman-tabel (tab. II). Uit deze tabel blijkt dat bij een wat grotere N ook de tabel voor Pearson's r kan worden gebruikt.

5 Slot

Mits verstandig toegepast blijken regressie- en correlatie-analyse zeer nuttige en vaak gebruikte technieken te zijn. In deze bijdrage is de bespreking van beide technieken beperkt tot de relatie tussen twee variabelen, maar uitbreiding tot meer variabelen is mogelijk. De volgende punten worden benadrukt:

1. Het nut van regressie-analyse is niet beperkt tot voorspellen. Zo kunnen door inspectie ook extreme waarden ('outliers') worden opgespoord. De analyse is eveneens zinvol voor beschrijven van verbanden, maar met oorzaakgevolg conclusies moet men oppassen. Als bijvoorbeeld snoepen en DMF-getal positief correleren, is het onzinnig te concluderen dat een hoger DMF-getal snoepen bevordert. Een correlatie zegt dus niets over een causaal verband tussen de variabelen, hoewel daarvan wel sprake kan zijn. Op statistische gronden kan men geen uitspraak over oorzaakgevolg doen.
2. De correlatie is geen maat voor de helling van de regressielijn. Als bijvoorbeeld voor 50-70-jarigen de correlatie tussen leeftijd en aantal natuuurlijke gebitselementen voor Nederland −0,40 is en voor de Verenigde Staten −0,15, dan kan niet worden geconcludeerd dat in Nederland het gebitsverval sterker is dan in de VS. Wel mag dan worden gezegd dat de 'Nederlandse' punten relatief strakker langs de regressielijn liggen dan de 'Amerikaanse' langs de Amerikaanse regressielijn. Hoe beide regressielijnen ten opzichte van elkaar lopen, vertelt de correlatiecoëfficiënt niet.
3. Als een tandheelkundig artikel gebaseerd is op waargenomen correlatiecoëfficiënten, dan is de inhoud van het artikel alleen goed te interpreteren als de besproken relaties lineair zijn. Dit moet dan ook blijken (of geverifieerd zijn) aan de hand van plaatjes (scattergrammen).

Summary

STATISTICS (VI); REGRESSION AND CORRELATION

Key word: Statistics

This article describes and illustrates the meaning and application of both regression and correlation. Important terms, such as intercept, slope, variance explained, scatterplot, Pearson test, and causal relationship are introduced. The interpretation of both statistics are presented.

Literatuur

[1]GUILFORD JP, FRUCHTER B. Fundamental statistics in psychology and education. Tokyo: McGraw-Hill Kogakusha, 1978: ch. 15.
[2]BULMAN JS, OSBORN JF. Analysing the association between two variables. Br Dent J 1989; 166: 303-7.
[3]SNEDECOR GW, COCHRAN WG. Statistical methods. Ames, Iowa: The Iowa State University Press, 1976: ch. 6.
[4]WILLEMSEN EW. Understanding statistical reasoning. San Francisco: W.H. Freeman and Company, 1974: 84-7.
[5]SIEGEL S, CASTELLAN NJ. Nonparametric statistics. New York: McGraww-Hill, 1988.