Podosophia (2018) 26:82–87
https://doi.org/10.1007/s12481-018-0195-8

Regressie-analyse

Eveline Wouters · Sil Aarts

Published online: 14 May 2018

Samenvatting Bij een vergelijking van groepen op numerieke variabelen, is het mogelijk om bij de analyse van de data rekening te houden met de (relatieve) invloed van een of meer andere variabelen. Hiervoor is de meervoudige regressieanalyse ontwikkeld. Bij een meervoudige regressieanalyse wordt ook rekening gehouden met de invloed die andere variabelen hebben op de uitkomstmaat.

Trefwoorden regressieanalyse

Inleiding

Het artikel in nummer 1 van Podosophia van dit jaar beschreef de *t*-toetsen en ANOVA [1]. Bij dit soort analyses worden groepen met elkaar vergeleken op bepaalde, numerieke variabelen, zonder dat daarbij rekening wordt gehouden met de invloed van andere variabelen. In dit artikel gaan we verder in op analysemethodes die de 'relatieve' invloed van andere variabelen ook bij de berekening betrekken. Bij die methodes wordt dus rekening gehouden met de invloed die andere variabelen op de uitkomst hebben.

In deze rubriek dragen de auteurs een steentje bij aan het vergroten van de kennis over wetenschappelijk onderzoek en de toepasbaarheid ervan in de podotherapeutische praktijk.

E. Wouters, PhD MD
Tranzo, Department of Tranzo, School of Social and Behavioral Sciences, Tilburg University, Tilburg, Nederland

E. Wouters, PhD MD (✉) · S. Aarts, PhD
Health Innovations & Technology, Fontys Paramedische Hogeschool, Eindhoven, Nederland
e.wouters@fontys.nl

Lineaire regressie

Lineaire regressieanalyse is een methode waarmee wordt nagegaan in hoeverre onafhankelijke variabelen (ook wel de 'voorspellers') een bepaalde uitkomst (de afhankelijke variabele, de uitkomstmaat waar je iets van wilt weten) kunnen voorspellen. De eenvoudigste vorm van een lineaire regressie is één voorspeller en één uitkomstmaat, bijvoorbeeld het verband tussen leeftijd (onafhankelijke variabele) en de hoogte van de bloeddruk (afhankelijke variabele, uitkomstmaat). In feite is dit niet anders dan wat bij de *t*-toets of een ANOVA wordt gedaan.

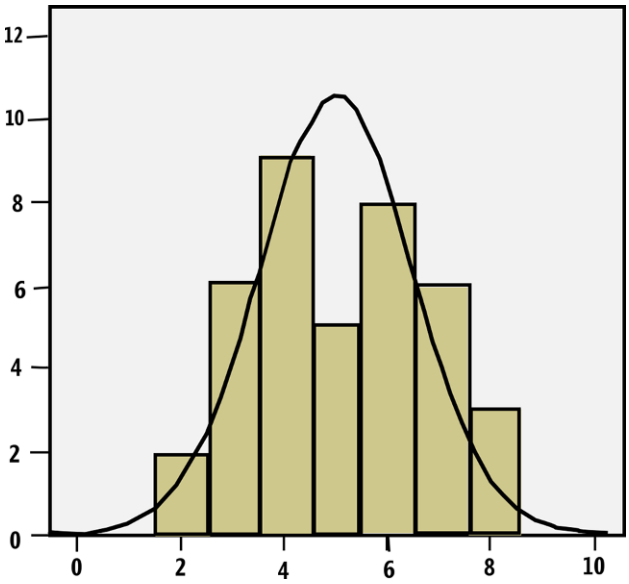
Meervoudige lineaire regressie

Interessanter wordt het als er sprake is van meervoudige regressie: in hoeverre voorspellen meerdere variabelen, *rekening houdend met andere variabelen,* een bepaalde uitkomst(maat)? Bijvoorbeeld: Wat is het verband tussen leeftijd en bloeddruk, rekening houdend met zoutconsumptie en BMI (dat is: gewicht gerelateerd aan lengte). Op die manier wordt als het ware de overlap tussen variabelen 'uitgefilterd', immers, ook (een hoge) BMI en (veel) zoutgebruik kunnen van invloed zijn op de bloeddruk. Deze beïnvloedende factoren worden ook wel *confounders* genoemd.

Voorwaarden lineaire regressieanalyse

Net als bij de *t*-test gelden bij de lineaire regressieanalyse(s) voorwaarden om deze te kunnen uitvoeren. Bekende voorwaarden zijn:

- de afhankelijke variabele is altijd een numerieke variabele;
- de variabelen zijn normaal verdeeld.



Figuur 1 Voorbeeld van een normale verdeling. Op de x-as (*horizontale as*) staat het medicatiegebruik (aantal doses) per week van een bepaalde populatie en op de y-as (*verticale as*) staat de frequentie van dat medicatiegebruik

Een normale verdeling betekent dat de waarden geconcentreerd zijn rondom een bepaald gemiddelde (bijv. fig. 1). Met andere woorden, er zijn geen uitschieters naar boven of onder. Als dat laatste wel het geval is (uitschieters), dan is het gemiddelde geen goede representant.

Voorbeeld lineaire regressieanalyse

In een artikel over de invloed van pes planovalgus (platvoet) op kwaliteit van leven, worden een aantal uitkomstmaten voorspeld aan de hand van een aantal onafhankelijke variabelen [2]. We bekijken één uitkomstmaat uit dit onderzoek, de variabele 'pijn' in aangepaste en vereenvoudigde vorm, en vertaald, in tab. 1.

In tab. 1 staan in de eerste kolom de 'onafhankelijke' variabelen, ofwel de variabelen waarvan vermoed wordt dat ze mogelijk invloed hebben op de 'afhankelijke' variabele (de variabele waarin we geïnteresseerd zijn, de uitkomst, hier: de pijnscore). In de andere kolommen staan de volgende tekens:

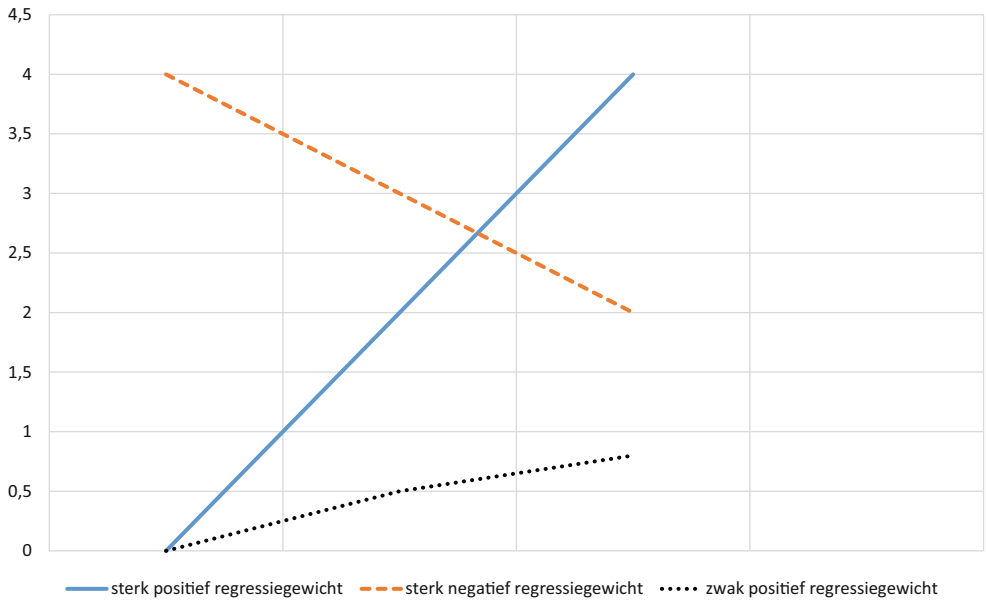
- B en Bèta (Grieks) zijn de zogenaamde regressiecoëfficiënten: ze zijn een maat voor het gewicht dat elke variabele in de schaal legt als voorspeller voor de afhankelijke variabele. De letter B wordt daarbij gebruikt als aanduiding voor dit 'aandeel' in de steekproef die in het onderzoek is gebruikt. Bèta wordt gebruikt indien wordt uitgegaan van de gehele populatie. Een minteken geeft aan dat de relatie omgekeerd is: een hogere waarde van de onafhankelijke variabele, gaat gepaard met een lagere waarde van de afhankelijke variabele. Dit staat toegelicht in fig. 2. In het voorbeeld in tab. 1 leidt een hogere score op de Charlson-index dus tot een lagere, met de FHSQ gemeten pijnscore.
- De standaard error is een maat voor de nauwkeurigheid van de relatie tussen de afhankelijke en onafhankelijke variabele(n).
- De *t*-waarde (vergelijkbaar met *t*-toets) kijkt of de H_0 (de nulhypothese) al dan niet verworpen moet worden. De H_0 gaat ervan uit dat er geen statistisch significante relatie is tussen afhankelijke en onafhankelijke variabele. De alternatieve hypothese (H_1) gaat ervan uit dat dat wel het geval is. In het geval van bijvoorbeeld geslacht wordt getoetst of de variabele geslacht, gecontroleerd voor de andere aanwezige onafhankelijke variabele, een statistisch significante relatie heeft met pijnscore.
- De *p*-waarde geeft aan of de (van tevoren opgestelde) kans dat de gevonden associatie tussen afhankelijke en onafhankelijke variabele berust op toeval, overschreden wordt of juist niet. Ofwel, of het gevonden verband statistisch significant is. Wat de grootte van de overschrijdingskans is, wordt door de onderzoekers van tevoren vastgesteld. Meestal[1] wordt die kans gesteld op 5 % (= 0,05), wat in het voorbeeld betekent dat alleen geslacht en Charlson-score (een maat voor comorbiditeit) statistisch significant geassocieerd zijn met ervaren pijn. De variabele 'platvoet' zelf niet (*p*-waarde is net iets groter dan 5 %, namelijk 0,053).

Tabel 1 De relatie tussen verschillende onafhankelijke variabelen (o.a. geslacht en platvoet) en de afhankelijke variabele pijnscore [2]

Variabelen	Pijnscore gemeten met de *Foot Health Status Questionnaire* (FHSQ)				
	B	*Standard error*	*Bèta*	*t*	*p-waarde*
geslacht	–9,225	3,743	–0,249	–7,016	< 0,001
leeftijd	–0,007	0,060	–0,004	–2,134	0,913
Charlson-score[a]	–1,284	0,602	–0,080	–2,134	0,003
platvoet	–2,931	1,510	–0,070	–1,942	0,053

[a]De Charlson-score is een maat voor de aanwezigheid van comorbiditeit

[1] Dit percentage is afhankelijk van de omvang van de populatie in het onderzoek: als de populatie heel groot is, wordt de overschrijdingskans kleiner genomen.



Figuur 2 Drie regressielijnen met respectievelijk sterk-positieve (*de blauwe lijn*) en sterk-negatieve regressiegewichten (oranje *gestreepte lijn*) en een zwak positief regressiegewicht (*zwarte stippellijn*)

Logistische regressieanalyse

Bij logistische regressieanalyse is de uitkomstmaat (de afhankelijke variabele) dichotoom: er zijn slechts twee uitkomsten mogelijk. Een voorbeeld van een dichotome variabele is geslacht (man *vs.* vrouw) of een aandoening, zoals een diabetisch ulcus (aanwezigheid *vs.* afwezigheid ervan). De voorspellende waarden kunnen allerlei soorten variabelen zijn, zowel dichotome variabelen (bijv. geslacht), als ordinale (bijv. opleidingsniveau) als continue (bijv. leeftijd).

Enkele afkortingen

- *Odds*. De verhouding tussen de kans op het optreden van een ziekte, uitkomst of gebeurtenis en de kans op het *niet* optreden ervan. Bijvoorbeeld: de kans dat in een populatie met obesitas diabetes voorkomt, in vergelijking met de kans dat diabetes niet voorkomt in deze groep.
- *Odds ratio (OR)*. De odds op het optreden van bijvoorbeeld een ziekte binnen een bepaalde groep, in vergelijking met de odds in een andere groep, bijvoorbeeld: de odds voor het optreden van diabetes bij mensen met obesitas vergeleken met de odds binnen een groep met normale BMI.
- *Relatief risico (RR)*. De kans dat een ziekte in een bepaalde groep optreedt, in vergelijking met de kans dat deze optreedt in een andere groep, bijvoorbeeld: de kans op longkanker in een populatie die rookt, in vergelijking met de kans op longkanker in een populatie die niet rookt of nooit gerookt heeft.

In tab. 3 wordt aan de hand van een getallenvoorbeeld het verschil tussen kans, odds, RR en OR weergegeven.

Voorbeeld logistische regressieanalyse

In het artikel van Almobarak worden de risicofactoren voor het ontwikkelen van een diabetisch voetulcus onderzocht bij mensen met diabetes mellitus type 2 [3]. Het gaat er in dit artikel om te ontdekken welke variabelen een relatie vertonen met het ontwikkelen van een ulcus. Een aantal variabelen wordt daarbij verondersteld van belang te zijn. Men vroeg zich in dit onderzoek af in hoeverre onder andere het geslacht, de leeftijd, de BMI, de duur van het bestaan van de diabetes (langer of korter dan 10 jaar), het cholesterolgehalte in het bloed, het HbA1C (maat voor bloedsuikerspiegel over langere tijd) en het al dan niet bestaan van neuropathie, mede bepalend zijn voor het ontstaan van een diabetisch voetulcus. In tab. 2 wordt, omwille van de duidelijkheid, slechts een deel van de resultaten van de logistische regressieanalyse weergegeven.

De uitkomstmaat in het voorbeeld is een dichotome maat: in dit geval het al dan niet aanwezig zijn van een diabetisch voetulcus. Er zijn diverse verondersteld voorspellende variabelen. In tab. 2 staan enkele andere tekens in de kolommen dan in tab. 1 over lineaire regressie:

- OR staat voor odds ratio [4]. Dit is de verhouding tussen twee zogenaamde wedverhoudingen of odds (NB: niet hetzelfde als kans!). De wedverhouding is de verhouding tussen de waarschijnlijkheid dat een gebeurtenis voorvalt (of zal voorvallen) en de waarschijnlijkheid dat ze niet voorvalt (of zal voorvallen). Bijvoorbeeld: de waarschijnlijkheid dat bij de aanwezigheid van neuropathie een diabetes voetulcus zal voorvallen, gedeeld door de waarschijnlijkheid dat die voetulcus *niet* zal voorvallen (in de tabel: OR = 1,858). Als de OR precies 1 is, dan betekent dit

Tabel 2 Logistische regressieanalyse voor de voorspelling van diabetische voetulcera met diverse voorspellende variabelen. Bron: [3] (deel van *tabel 2*, pag. 5 overgenomen en aangepast)

Variabelen	OR	95 %-CI		*p*-waarde
geslacht	1,234	0,521	2,920	0,633
leeftijd	0,992	0,959	1,026	0,642
BMI	0,968	0,891	1,053	0,448
duur DM >10 jaar	4,158	1,507	11,476	0,006
cholesterol	1,002	0,995	1,009	0,560
HbA1C[a]	0,943	0,750	1,185	0,613
neuropathie	1,858	0,767	4,502	0,170

BMI body mass index (gewicht gedeeld door lengte in het kwadraat)
[a]HbA1C (geglyceerd hemoglobine) is een maat voor de gemiddelde bloedsuikerwaarde over een langere periode (maanden).

Tabel 3 Een getallenvoorbeeld met R, O, RR en OR. Bron: [3]

	Uitkomst (bijv. ziekte) aanwezig	Uitkomst (bijv. ziekte) niet aanwezig	Totaal aantal mensen
met behandeling	10	50	60
zonder behandeling	40	30	70
	50	80	130

Risico op uitkomst (ziekte) met behandelingen (R+): 10/60 = 0,167
Risico op uitkomst (ziekte) zonder behandeling (R–): 40/70 = 0,571
RR: 0,167/0,571 = 0,292
Odds op uitkomst (ziekte) met behandeling: 10/50 = 0,200
Odds op uitkomst (ziekte) zonder behandeling: 40/30 = 1,333
OR: 0,200/1,333 = 0,150

dat mét neuropathie het hebben van een diabetisch ulcus even waarschijnlijk is als zonder neuropathie. In dit voorbeeld is het getal echter ruim 1,8, al is deze OR (zie verderop) niet statistisch significant (*p*-waarde is 0,170).

- De *p*-waarde is, net als bij lineaire regressieanalyse, een maat voor de statistische significantie. Afhankelijk van de tevoren opgestelde grens waaronder toeval onwaarschijnlijk geacht wordt (veelal 5 %), kan aangegeven worden of een gevonden associatie berust op toeval. In het voorbeeld is alleen 'duur van diabetes langer dan 10 jaar' statistisch significant geassocieerd met het ontstaan van een diabetisch voetulcus: de *p*-waarde is immers < 0,05, namelijk: 0,006.
- 95 %-CI (*confidence interval*, in het Nederlands: betrouwbaarheidsinterval, BI). Dit is enerzijds een maat voor de precisie van de gevonden waarden (hoe kleiner het interval, des te preciezer, want des te dichter liggen de uitkomsten bij elkaar), anderzijds (net als de *p*-waarde) een maat voor het al dan niet op toeval berusten van een gevonden uitkomst. Bij een logistische regressie geldt dat als een CI de waarde 1 bevat, er *geen* sprake is van een statistisch significante associatie. In tab. 2 is te zien dat alle variabelen met *p*-waarden > 0,05 ook een CI-interval hebben waarin het getal 1 zit. De variabele 'neuropathie' bijvoorbeeld, heeft een CI tussen 0,767 en 4,502. Dat wil zeggen, de waarde '1' ligt hierin besloten. De *p*-waarde is 0,170, dus > 0,05, wat betekent 'niet statistisch significant'. Alleen 'duur diabetes >10 jaar' heeft zowel een *p*-waarde <0,05, als een CI waarin de 1 niet ligt besloten: de CI ligt tussen 1,507 en 11,576.

Take home messages

- Met regressieanalyse is het mogelijk om het relatieve aandeel van een variabele op een bepaalde uitkomst te berekenen, waarbij rekening wordt gehouden met de aanwezigheid van andere variabelen. Met andere woorden, er wordt gekeken of de aanwezigheid van bepaalde variabelen, bepaalde uitkomsten kunnen voorspellen.
- Bij lineaire regressie gaat het om het verband tussen numerieke variabelen.
- Bij logistische regressie is de uitkomstmaat dichotoom (wel-niet aanwezig).

Vooruitblik

Het volgende artikel zal gaan over diagnostische toetsen en begrippen zoals sensitiviteit en specificiteit. Sensitiviteit zegt iets over de gevoeligheid van een test. Als een test voor het aantonen van een aandoening sensitief is, zal deze slechts zelden een negatieve uitslag geven bij iemand die de ziekte toch blijkt te hebben. Als een test specifiek is, zal deze slechts zelden een positieve uitslag geven bij iemand die de aandoening toch niet blijkt te hebben.

Literatuur

1. Aarts S, Wouters E. De t-toets en de analysis of variance, ANOVA. Podosophia. 2018;26(1):28–33.
2. Pita-Fernandez S, Gonzalez-Martin C, Alonso-Tajes F, Seoane-Pillado T, Pertega-Diaz S, Perez-Garcia S, et al. Flat foot in a random population and its impact on quality of life and functionality. J Clin Diagn Res. 2017;11(4):LC22–LC7.
3. Almobarak AO, Awadalla H, Osman M, Ahmed MH. Prevalence of diabetic foot ulceration and associated risk factors: an old and still major public health problem in Khartoum, Sudan? Ann Transl Med. 2017;5(17):340.
4. Bouter LM, Dongen MCJM van, Zielhuis GA, Zeegers MPA. Leerboek epidemiologie. Houten: Bohn Stafleu van Loghum; 2016.

Eveline Wouters, hoogleraar/lector succesvolle technologische innovaties in de zorg

Sil Aarts, docent/onderzoeker