ANGELIQUE CRAMER

DE GEREEDSCHAPSKIST

VERKLAREN OF VOORSPELLEN?

'Leuker kunnen wij het als methodologen niet voor u maken, hopelijk wel makkelijker.' De rubriek De Gereedschapskist wordt verzorgd door redactieraadslid Angélique Cramer, universitair docent aan de Programmagroep Psychologische Methodenleer van de Universiteit van Amsterdam. Deze derde aflevering gaat over verklaren en voorspellen.

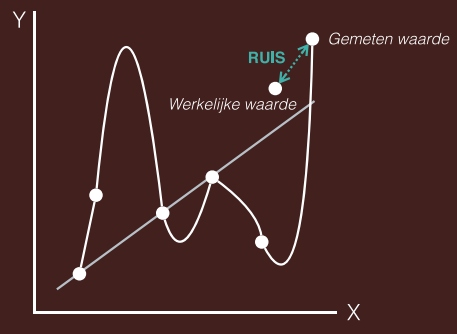
Veel vragen in psychologisch onderzoek die we willen beantwoorden, behoren tot één van twee categorieën. De eerste categorie betreft het soort vragen die een bepaald psychologisch verschijnsel willen verklaren: onderzoekers trachten de causale antecedenten van een verschijnsel te ontrafelen, bijvoorbeeld waarom sommige mensen depressief worden. De tweede categorie betreft het soort vragen die een bepaald psychologisch verschijnsel willen voorspellen: het accuraat kunnen voorzien van een verschijnsel dat nog niet heeft plaatsgevonden. Bijvoorbeeld: welke mensen hebben het grootste risico op het ontwikkelen van een depressieve episode? Is het model dat het beste verklaart ook het model dat het beste voorspelt? Nee, zeggen Yarkoni en Westfall (2016). Hoe komt dat?

Stel: u heeft een mogelijke verklaring voor verschijnsel Y (bijvoorbeeld: depressie), namelijk X (zeg: mate van neuroticisme). U verzamelt data over X en Y bij zeven mensen. Vervolgens gaat u na welke regressievergelijking het beste op de data past. In de figuur staan de resultaten; elke stip stelt één van de zeven deelnemers aan het onderzoek voor. De zwarte regressielijn past het beste op de data. Sterker, deze lijn past perfect: de lijn gaat immers door alle stippen heen. Zo is de zwarte lijn de perfecte verklaring voor uw resultaten: de mate van depressie (Y) wordt volledig verklaard door de mate van neuroticisme (X). Nu lijkt de zwarte lijn tevens een perfecte voorspelling te genereren: ik kan met grote zekerheid voorspellen wat de waarde van een willekeurig persoon uit de onderzoeksgroep is op variabele Y als ik weet wat diens waarde is op variabele X.

Maar: kan ik met de perfecte zwarte regressielijn uit de figuur accuraat voorspellen wat de waarde is van variabele Y voor een persoon buiten onze dataset als ik diens waarde op variabele X weet?

Vaak niet. Een belangrijke reden is dat psychologische data veel ruis bevatten: bijvoorbeeld vanwege een neiging tot sociaal wenselijk antwoorden is het reëel te veronderstellen dat een gemeten waarde vaak niet overeenkomt met de werkelijke waarde. De perfecte zwarte regressielijn past dus weliswaar op alle datapunten; maar aangenomen dat die punten deels ruis bevatten, is het zeer hypothetisch – en dat vergeten we vaak in onderzoek – dat de regressielijn ook iets vertelt over de relatie tussen X en Y in een andere steekproef. Als voorspelling in de gehele populatie – dus niet alleen in de gemeten steekproef – het doel is, dan wordt aangeraden genoegen te nemen met een model (bijvoorbeeld de grijze lijn in de figuur). Die geeft weliswaar in uw steekproef een minder goede verklaring voor de gevonden resultaten (de lijn gaat immers niet door alle datapunten heen), maar verhoogt de kans accurate voorspellingen te doen buiten de steekproef. De invloed van ruis is namelijk minder.

Iets om mee te nemen in uw overwegingen, als u een prachtig passend verklarend model tegenkomt met de pretentie tevens voorspellend te zijn.



Yarkoni, T, & Westfall, J. (2016). Choosing prediction over explanation in psychology: Lessons from machine learning. Figshare. https://dx.doi.org/10.6084/m9.figshare.2441878.v1.

FOTO: SHUTTERSTOCK