METEN EN WETEN

STATISTISCHE ANALYSES VAN CAUSALE RELATIES

PROF. DR. KEES VAN MONTFORT

NYENRODE
BUSINESS UNIVERSITEIT

NYENRODE
BUSINESS UNIVERSITEIT

Straatweg 25
P.O. Box 130 × 3620 AC Breukelen
The Netherlands
www.nyenrode.nl

10 MAART 2006

INAUGURELE REDE

Meten en Weten

Statistische analyses van causale relaties

Prof. dr. Kees van Montfort
Breukelen, Maart 10, 2006

Rede in verkorte vorm uitgesproken bij
het aanvaarden van het ambt van hoogleraar in
de Kwantitatieve Bedrijfskundige Onderzoekstechnieken aan de
Nyenrode Business Universiteit op vrijdag 10 maart 2006.

ISBN number: 9073314844

Contents

1. Inleiding

Mijnheer de Decaan,

Hooggeleerde Heren,

Zeer gewaardeerde toehoorders,

De titel van deze oratie is 'meten *en* weten'. De gangbare zegswijze luidt echter 'meten *is* weten'. Hiermee wordt dan bedoeld dat het op aanzienlijke schaal en op systematische wijze verzamelen van gegevens over een bepaald verschijnsel leidt tot wetenschappelijke kennis en inzichten over dat verschijnsel. Het is echter vrij algemeen bekend dat de stelling 'meten is weten' niet zonder meer klopt. Zo leveren verkeerde metingen geen zinvolle bijdrage aan de ontwikkeling van wetenschappelijke kennis. Indien een meting bijvoorbeeld niet valide, onbetrouwbaar of niet gebaseerd op een adequate steekproef is, ontstaat geen waarheidsgetrouw beeld van het gemeten verschijnsel. Bij enquêtes doet zich bijvoorbeeld de complicatie voor dat mensen die geen interesse hebben voor het te onderzoeken verschijnsel, minder geneigd zijn aan de enquête mee te werken. Dat kan leiden tot vertekende uitkomsten. Zo geldt voor de continue Gezondheidsenquête van het Centraal Bureau voor de Statistiek dat mensen met gezondheidsproblemen relatief vaak aan de enquête lijken mee te doen. Daardoor geeft de enquête een overschatting van een aantal ligdagen in het ziekenhuis. Dat blijkt indien de enquête-gegevens worden vergeleken met reële gegevens van ziekenhuizen (Bakker en Van Rooijen, 2000; Te Riele, 2002).
Soms leidt meten zelfs nadrukkelijk tot niet-weten. Sommige zaken zijn immers niet goed te kwantificeren. Doet men toch een poging daartoe, dan zal men door de uitkomsten van het onderzoek op het verkeerde been worden gezet. In dit verband valt bijvoorbeeld te wijzen op de risico's van de nadruk die tegenwoordig bij de overheid en in de non-profit sector wordt gelegd op outputmeting als instrument om de kwaliteit van de

dienstverlening van organisaties vast te stellen. De kwaliteit van onderwijsinstellingen wordt tegenwoordig hoofdzakelijk bepaald aan de hand van de studierendementen. Een hoog studierendement zou echter wel eens niet zozeer een aanwijzing voor goed onderwijs kunnen zijn maar veeleer een teken dat de betreffende onderwijsinstelling tentamens en scripties soepel nakijkt. Hoge scores van welzijns-instellingen of politiekorpsen op bepaalde prestatie-indicatoren zouden wel eens niet zozeer kunnen duiden op een adequate vervulling van de maatschappelijke opdracht maar meer op een gerichtheid op gemakkelijke gevallen (De Bruijn, 2002; Tonkens, 2003, WRR, 2004).

Het voorgaande neemt uiteraard niet weg dat 'meten' en 'weten' onlosmakelijk met elkaar zijn verbonden. Voor die aspecten van de werkelijkheid die door bedrijfskundigen, economen, econometristen, e.d. worden bestudeerd, geldt dat men slechts dan tot waarheids-getrouwe kennis kan komen indien correcte metingen worden uitgevoerd. Correcte metingen vormen met andere woorden een noodzakelijke voorwaarde voor het vergaren van kennis. Het betreft echter geen voldoende voorwaarde. Willen goede metingen resulteren in kennis(vermeerdering), dan moet tevens aan minstens twee aanvullende voorwaarden worden voldaan.
De eerste aanvullende voorwaarde is dat bij het analyseren en interpreteren van gegevens correcte redeneringen moeten worden toegepast. Zijn de gegevens via een correcte meetmethode verkregen maar worden ze vervolgens onderworpen aan een drogredenering, dan zullen de conclusies zonder waarde zijn (Dunn, 2004: 421-426). Stel dat volgens een zorgvuldig opgezette en uitgevoerde enquête het merendeel van de ondernemers van oordeel is dat de belastingen hier in Nederland aanmerkelijk hoger zijn dan in omringende landen. Uit deze informatie kan vanzelfsprekend niet de conclusie worden getrokken dat het belastingklimaat voor ondernemers in Nederland ongunstiger is dan elders (Van Eemeren, Grootendorst en Kruiger, 1986: 113 en 118). In dit verband zijn de vermanende woorden van de inmiddels al weer vijftien jaar geleden overleden *Vrij Nederland*-columniste Renate Rubinstein leerzaam: "Honderdduizend juichende lezeressen, lezer, kijkers, luisteraars zijn geen enkel bewijs voor de juistheid van een denkbeeld en het is pure demagogie om het als argument te gebruiken" (*Vrij Nederland*, 6 maart 1983).

De tweede aanvullende voorwaarde is dat een eventueel oorzakelijk verband, een eventuele *causaliteit*, tussen aspecten van het bestudeerde verschijnsel op een juiste wijze wordt vastgesteld. In vrijwel alle theorieën op het domein van de bedrijfswetenschappen en de economische wetenschappen wordt uitgegaan van causale relaties. Dergelijke relaties zijn niet altijd even gemakkelijk aan te tonen. En omgekeerd kan er ook gemakkelijk een causale relatie lijken te bestaan terwijl dat in werkelijkheid niet het geval is.
Een voorbeeld hiervan betreft de relatie tussen het aantal afstudeerders van het parttime doctoraalprogramma (tegenwoordig parttime MSc-programma) aan de Nyenrode Business Universiteit enerzijds en de winstgevendheid van Nederlandse bedrijven anderzijds. Een toename (of afname) van het aantal doctoraalbullen aan de Nyenrode Business Universiteit bleek in het recente verleden gepaard te gaan met een toenemende (of afnemende) winstgevendheid van het Nederlandse bedrijfsleven. Is hier sprake van een causale relatie? Hoewel de onderwijs-programma's van de Nyenrode Business Universiteit onmiskenbaar van een hoog niveau zijn, lijkt in dit geval geen sprake te zijn van een oorzaak-gevolg relatie. Dat wordt duidelijk als we een aantal causaliteits-voorwaarden uit de literatuur in ogenschouw nemen.

De literatuur reikt een aantal causaliteitsvoorwaarden aan die steun bieden bij het analyseren en interpreteren van gegevens om een eventuele causaliteit te kunnen vaststellen. Swanborn (1977 en 1991) formuleert in zijn standaardwerken drie causaliteitsvoorwaarden: statistische samenhang, juiste tijdsvolgorde en geen schijnverband. Met statistische samenhang wordt bedoeld dat causaal-gerelateerde verschijnselen statistisch gecorreleerd moeten zijn. Met de juiste tijdsvolgorde wordt bedoeld dat als verschijnsel A verschijnsel B causaal beïnvloedt, verschijnsel B in chronologische zin niet vooraf mag gaan aan verschijnsel A. Een in empirisch onderzoek vaak veronachtzaamde voorwaarde is mijn inziens de derde. Deze voorwaarde houdt in dat er een theoretische relatie moet kunnen worden gelegd tussen de oorzaak en het gevolg. Met andere woorden, een causale relatie moet theoretisch verklaarbaar en waarschijnlijk zijn.
Deze derde voorwaarde houdt ook in het uitsluiten van verborgen variabelen. In het voorbeeld van het aantal doctoraalbullen van de Nyenrode Business Universiteit en de winstgevendheid in het

Nederlandse bedrijfsleven is er een verborgen variabele, namelijk het budget dat bedrijven beschikbaar stellen voor het bij- en nascholen van hun werknemers. Deze verborgen variabele zorgt voor de vermeende causaliteit. In werkelijkheid is er sprake van een schijnverband.

Deze oratie handelt over het vaststellen van een eventuele causaliteit tussen bepaalde aspecten van een verschijnsel op basis van een statistische analyse van databestanden. Meer in het bijzonder richt ik mij op de volgende vraag: “Stel we hebben meetgegevens over bepaalde variabelen. Op welke wijze kunnen dan met behulp van moderne statistische methoden en technieken eventuele oorzaak-gevolg relaties tussen twee of meer variabelen worden aangetoond?”

2. Begripsbepaling: een model voor causaliteit

Om bovenstaande vraag te kunnen beantwoorden wordt nu eerst het begrip “causaliteit” nader verkend. Daarbij maak ik gebruik van het Rubin-Holland (RH)-model, dat door Rubin (1974) is bedacht en door Holland (1986) verder is uitgewerkt. Rubin en Hollander hebben inmiddels een Nobelprijs ontvangen voor hun wetenschappelijke werk. Het RH-model heeft betrekking op causale verbanden in de context van gemeten waarnemingen. Deze gemeten waarnemingen kunnen verkregen worden middels een *van tevoren* geconstrueerd experiment. Een belangrijke beperking van experimenten is dat we de verschillende gevolgen van een oorzaak kunnen onderscheiden, maar niet de oorzaak van een gevolg. In een experiment staat van tevoren vast wat de oorzaak is en wat het gevolg. In werkelijkheid is dat echter niet altijd duidelijk. Doorgaans maken onderzoekers de indeling in oorzaak en gevolg op basis van de volgorde waarin de gebeurtenissen optreden: de oorzaak gaat vooraf aan het gevolg. In de bedrijfswetenschappen en de economie is de tijdsvolgorde van gebeurtenissen echter niet altijd relevant voor de indeling in oorzaak en gevolg. Een voorbeeld hiervan wordt gegeven door Van Den Berg (1990). Werklozen blijken intensiever naar werk te gaan zoeken wanneer het moment nadert waarop de werkloosheidsuitkering omlaag gaat. In chronologische zin wordt eerst intensiever naar werk gezocht en daarna gaat de uitkering eventueel omlaag. Het intensievere zoekgedrag kan echter moeilijk als oorzaak van de daling van de uitkering beschouwd worden. Dit voorbeeld illustreert dat oorzaak en gevolg soms in een omgekeerde tijdsvolgorde optreden. De indeling in oorzaak en gevolg kan in een dergelijk geval gerechtvaardigd worden door theorie en niet door waarneming.

Causaliteit veronderstelt in het RH-model het bestaan van een of meer causale effecten. Als we over causale effecten spreken, moeten we eerst aangeven wat onder een causaal effect wordt verstaan. We zullen het

causale effect op een variabele ten gevolge van een actie ten aanzien van een oorzaak-variabele definiëren als de waarde van de eerstbedoelde variabele na de actie ten aanzien van de oorzaak-variabele minus de waarde van de eerstbedoelde variabele zonder de actie ten aanzien van de oorzaak-variabele. Dit kunnen we verduidelijken aan de hand van een voorbeeld met betrekking tot de gevolgen van een reclamecampagne voor de verkoopcijfers van het betreffende product. Het causale effect is het verkoopcijfer na de reclamecampagne minus de waarde van het verkoopcijfer zonder de reclamecampagne. Bij dit voorbeeld stuiten we echter op een probleem. Het is immers onmogelijk om de verkoopcijfers zowel met als zonder de reclamecampagne te meten. Slechts één van de twee verkoopcijfers kan gemeten worden, niet beide cijfers.
Voor het in praktijk kunnen toepassen van een causaal effect moeten we derhalve een aanname maken. We zullen aannemen dat *experimentele eenheden in de tijd identiek* zijn. In bovenstaand voorbeeld betekent dat, dat het gewenste verkoopcijfer zonder reclamecampagne gelijk is aan het verkoopcijfer voordat de reclamecampagne werd uitgevoerd. Dit impliceert dat er gedurende de reclamecampagne geen andere observeerbare en niet-observeerbare omstandigheden zijn veranderd. Veranderingen van omstandigheden kunnen de verkoopcijfers immers eveneens beïnvloeden.
De aanname van in de tijd identieke experimentele eenheden, oftewel de veronderstelling dat de beschouwde eenheden in de loop van de tijd geen verandering ondergaan indien geen actie ten aanzien van de oorzaak-variabele plaatsvindt, gaat echter in de praktijk niet altijd op. Bovengenoemde definitie van een causaal effect kan dan niet aangehouden worden. Wel kunnen we iets doen indien we bereid zijn met minder dan het causale effect voor alle afzonderlijke eenheden (individuen, organisaties, producten, e.d.) genoegen te nemen. We kunnen bijvoorbeeld er toe overgaan slechts het *gemiddelde* causale effect voor een groep eenheden te meten. Het gemiddelde causale effect is bijvoorbeeld in het zojuist besproken voorbeeld het gemiddelde verkoopcijfer voor de producten waarvoor een reclamecampagne is gevoerd (de stimulus groep), minus het gemiddelde verkoopcijfer voor de producten die niet door een reclamecampagne zijn ondersteund (de controlegroep). Dit gemiddelde causale effect heeft echter ook weer een nadeel. Het is voor geen van de producten uit het assortiment relevant. Er is slechts sprake van een gemiddeld effect.

Het gemiddelde waargenomen effect en het gemiddelde causale effect kunnen wij overigens niet zonder meer aan elkaar gelijk stellen. Strikt genomen is het gemiddelde waargenomen effect de som van het gemiddelde causale effect voor de aan het experiment deelnemende eenheden en het verschil in uitgangspositie tussen de aan het experiment deelnemende eenheden en de niet-deelnemende eenheden, zij het dat de gemeten uitgangspositie betrekking heeft op de situatie na de verandering in de oorzaak-variabele. Bij een reclamecampagne kan de verkoop van bepaalde producten gevoeliger zijn voor reclamecampagnes dan de verkoop van andere producten. Als de uitgangsposities van de deelnemende en de niet-deelnemende eenheden gelijk zijn, meten wij het gemiddelde causale effect ten aanzien van de deelnemende eenheden. Als we at random een persoon, organisatie of product als deelnemer dan wel niet-deelnemer bestempelen, lijkt het reëel om aan te nemen dat de stimulus- en de controlegroep dezelfde uitgangspositie hebben (Heckman, 1979).
Een voorbeeld van een onderzoek waarin de uitgangsposities van de deelnemers en niet-deelnemers niet aan elkaar gelijk zijn, kreeg begin oktober van het vorige jaar aandacht in diverse kranten en op verschillende radiozenders. Het Noorse onderzoek (Endresen en Olweus, 2005) concludeerde dat jongens tussen 11 en 15 jaar, door de beoefening van vechtsporten, zoals boksen, karate en judo, agressiever worden. Bij nadere bestudering van het onderzoek door mij bleek er niet gekeken te zijn naar de uitgangsposities van de jongens, zodat de conclusie onmogelijk getrokken kon worden. Andere onderzoeken, die wel zorgvuldig zijn uitgevoerd en zowel naar de uitgangs- als de eindpositie kijken, geven aan dat vechtsporten juist tot een vermindering van agressiviteit leiden. Dit was overigens een behoorlijke geruststelling voor mij aangezien ik drie zoons in de desbetreffende leeftijdscategorie op judo heb zitten.

Zoals causaliteit hierboven beschreven is, beperkt het vaststellen van causale relaties zich tot het meten van oorzaak en gevolg. Dat kan een beperking zijn, omdat veel econometrisch onderzoek betrekking heeft op onderwerpen waarvoor geldt dat er slechts weinig theorie voorhanden is of de beschikbare theorie gewantrouwd wordt. In een dergelijk situatie is het econometrische onderzoek juist gericht op het ontdekken van

oorzaken van een gemeten gevolg. De theoretisch statistische rechtvaardiging hiervoor biedt de probabilistische theorie van causaliteit (Suppes, 1970; Glymour, Scheines, Spirtes en Kelly, 1987). In deze theorie wordt gesteld dat X een oorzaak is van Y als er geen Z gevonden kan worden, die indien constant gehouden, X en Y onafhankelijk maakt.
De praktische uitwerking van deze theorie leidt echter tot problemen. Er zijn voorbeelden te verzinnen, waarbij niet aan de causaliteitsdefinitie wordt voldaan, terwijl er overduidelijk wel sprake is van causaliteit. Dit kan onder andere geïllustreerd worden met een voorbeeld van Freedman (1994). Stel, dat scholing (=X) alleen een hoger inkomen (=Y) geeft door een kleinere kans op werkloosheid (=Z). In dat geval zou scholing, uitgaande van de probabilistische theorie van causaliteit, geen oorzaak zijn van het hogere inkomen.

In tijdreeks-analyses gebruikt men vaak het begrip Granger causaliteit (zie Granger, 1969). In het geval van twee tijdreeksen X_t en Y_t (t=1,2,3,....,T) van de variabelen X en Y, beïnvloedt X Y, als $X_1,...,X_{t-1}$ een statistisch significante bijdrage hebben in de regressie van Y_t op $X_1,...,X_{t-1}$, $X_t,Y_1,...,$ Yt_{-1}. De variabele Z is hier de reeks van vertraagde waarden van Y, te weten $Y_1,...,Y_{t-1}$. Als X volgens deze definitie Y beïnvloedt, is het ook goed mogelijk dat Y X veroorzaakt. In de definitie verwisselen we dan gewoon X en Y. We kunnen dan de conclusie krijgen dat X Y beïnvloedt en Y X. In feite is er dan sprake van simultaniteit of terugkoppeling tussen deze variabelen.
Ridder (1996) heeft de tijdreeksen (op jaarbasis) van het aantal broedparen ooievaars en het aantal geboorten van baby's in de Duitse deelstaat Baden-Wurtemberg op Granger causaliteit onderzocht. Hij krijgt de volgende regressievergelijking:

$$\text{GEBOORTE}_t = 0.40\ (0.15)\ \text{OOIEVAARS}_t + 0.50\ (0.094)\ \text{GEBOORTE}_{t-1} + 22.07\ (3.87).$$

Tussen haakjes staan de standaard deviaties van de schatters.
De bijbehorende Durban-Watson statistic heeft de waarde 2.11 en de R2 is 0.96. Het zal duidelijk zijn dat in dit geval aan de definitie van Granger causaliteit is voldaan en dat volgens deze definitie het aantal parende ooievaars het aantal geboorten beïnvloedt. Maar er is uiteraard geen sprake van een causale relatie. De verklaring voor de statistische

samenhang is immers dat ooievaars vooral te vinden zijn in stedelijke gebieden. Dergelijke gebieden kennen een relatief hoog geboortecijfer (Berkhout, Heyma en van Leeuwen, 2005: 27).

Bij Granger causaliteit worden in feite correlatie en causaliteit aan elkaar gelijk gesteld. Uiteraard kunnen we echter niet volstaan met het laten spreken van de gemeten gegevens, omdat de analyse kan leiden tot schijnverbanden. Bij het verifiëren van de aangetroffen correlaties op verschillende databestanden kunnen schijnverbanden onverminderd overeind blijven staan.

3. Experimentele randomisatie

Wat is dan wel een correcte methode om een eventueel causaal verband vast te stellen met behulp van statistische analyses? De zuiverste manier om een eventuele causaliteit te bepalen in een onderzoek is het gebruik van experimenten met gerandomiseerde toewijzing van personen aan de selectiegroep en de controlegroep (*gerandomiseerde experimenten*). In de literatuur zijn weinig voorbeelden van dit soort experimenten bekend. Een van de weinige voorbeelden is beschreven door Lalonde (1986). Lalonde analyseert de resultaten van het effect van een scholingsprogramma van de National Supported Work Demonstration (NSW) in de Verenigde Staten van Amerika. Het scholingsprogramma was een experimenteel programma dat tijdelijk banen aanbood aan onder andere bijstandsmoeders. Het was de bedoeling van het programma om personen met een zwakke positie op de arbeidsmarkt 9 tot 18 maanden werkervaring in de publieke of private sector te geven. Na afloop van het programma moesten de deelnemers zelf een andere baan zoeken. De controlegroep werd samengesteld door een aantal door het lot aangewezen potentiële deelnemers, dat wil zeggen personen die aan alle criteria voldeden en wilden meedoen, geen toegang te geven tot het scholingsprogramma. De toewijzing tot de geselecteerde groep deelnemers en de controlegroep was daardoor gerandomiseerd. Dit maakte het mogelijk om het gemiddelde causale effect van het scholingsprogramma op het inkomen een jaar na afloop van deelname te meten aan de hand van het waargenomen verschil in gemiddeld inkomen tussen de bijstandsmoeders uit het scholingsprogramma en de niet-deelnemers in de controlegroep. Voor de deelname waren het gemiddelde inkomen van de bijstandsmoeders in de deelnemersgroep en dat van de personen in de controlegroep nagenoeg gelijk, wat aangeeft dat de randomisatie goed was uitgevoerd. Tijdens het programma was het inkomen van de deelnemers veel hoger dan dat van de controlegroep. Dit werd veroorzaakt door het gegeven dat alle deelnemers inmiddels een baan had gekregen, terwijl een groot deel van de niet-deelnemers nog steeds werkloos was. Het inkomen van de controlegroep steeg wel.

Dit kon verklaard worden door het feit dat alle vrouwen op het moment van indeling werkloos waren en ook de vrouwen in de controlegroep zelf hun positie verbeterden. Een jaar na afloop van het programma verdienden de voormalige deelnemers aan het scholingsprogramma gemiddeld 22 % meer dan de niet-deelnemers, hetgeen zowel in statistische zin als in praktische zin significant is.
Lalonde vergeleek bovendien de schatting van het inkomensverschil, die met behulp van gerandomiseerde toewijzing was verkregen, met de schatting van het inkomensverschil, die verkregen werd door de deelnemers aan het scholingsprogramma te vergelijken met een controlegroep die niet door randomisatie was verkregen. Deze tweede controlegroep had een jaar voor de start van het scholingsprogramma hetzelfde gemiddeld inkomen als de deelnemers aan het scholingsprogramma. Ook waren de gemiddelde leeftijd, het aantal genoten schooljaren en de raciale achtergronden in de deelnemersgroep en de tweede controlegroep nagenoeg gelijk. De schatting van het causale effect, gemeten aan de hand van het inkomensverschil tussen de deelnemersgroep en de tweede controlegroep, bleek echter 3,5 keer zo groot te zijn als het effect dat gevonden werd bij het gerandomiseerde experiment.

Bovenstaand voorbeeld illustreert de meerwaarde van gerandomiseerde experimenten. Toch wordt er in de onderzoekspraktijk maar weinig gebruik van gemaakt. Dat heeft ongetwijfeld te maken met de praktische problemen bij het opzetten van gerandomiseerde experimenten (Heckman en Hotz, 1989). Zo is randomisatie alleen mogelijk bij programma's waar er minder plaatsen dan deelnemers zijn. Bij programma's waar deelnemers worden geselecteerd via een toelatingstest is het dan vervolgens moeilijk te verkopen dat een geschikt iemand niet mee mag doen. De uitgeslotenen, die deel gaan uitmaken van de controlegroep, kunnen zich anders gaan gedragen en daarmee een onzuiverheid bij de meting van het causale effect introduceren. Ook kan het feit dat er gerandomiseerd wordt, potentiële deelnemers afschrikken.
Randomisatie is bovendien moeilijk te realiseren bij bestaande programma's. Bij dergelijke programma's vindt de selectie plaats door professionals. Deze worden niet graag vervangen door een loterij en zullen daarom niet van harte meewerken aan randomisatie.

Randomisatie gaat uit van directe manipulatie van de deelname aan een programma. Soms wordt de deelname aan een programma indirect gemanipuleerd. De deelname is dan het gevolg van een interventie die de deelname weliswaar beïnvloedt, maar verder niets met het programma te maken heeft. We noemen dit ook wel een *natuurlijk gerandomiseerd experiment* (Angrist, 1990).
Een ander voorbeeld van een natuurlijk gerandomiseerd experiment wordt gegeven door Imbens en Van der Klaauw (1995). Tot ongeveer 10 jaar geleden gold er nog dienstplicht voor jongens vanaf 18 jaar. Vooral de laatste jaren van de dienstplicht dienden minder dan de helft van de in een jaar geboren mannen. De fractie van het aantal jongens van een bepaald geboortejaar dat in dienst ging, varieerde sterk. De in 1959 geboren mannen werden zelfs collectief vrijgesteld van dienstplicht. De variatie in de fractie werd vooral bepaald door de vraag naar dienstplichtigen. De indeling in dienstuitvoerders en vrijgestelden vond daardoor via een natuurlijke randomisatie plaats. Gebruikmakend van dit idee signaliseren Imbens en Van der Klaauw dat de dienstplicht het latere jaarinkomen met 8 procent doet dalen. Dit correspondeert met het verlies van twee jaar werkervaring, hetgeen meer is dan de 14 maanden die gemiddeld voor de dienstplicht stond. Dit is op zich een opmerkelijk resultaat, dat de vraag oproept of er inderdaad wel een natuurlijke randomisatie heeft plaatsgevonden.
Bij natuurlijke randomisatie is er sprake van een instrumentele variabele, die gebruikt kan worden bij het schatten van het causale effect. Een instrumentele variabele correspondeert met een interventie die wel de deelname beïnvloedt maar niet de uitkomst. In bovenstaand voorbeeld betreft de fractie jongens van een geboortejaar dat uiteindelijk diende, een instrumentele variabele die niet alleen invloed heeft op het al of niet dienen maar ook op het latere inkomen. Deze instrumentele variabele is dus niet geschikt voor het realiseren van natuurlijke randomisatie.

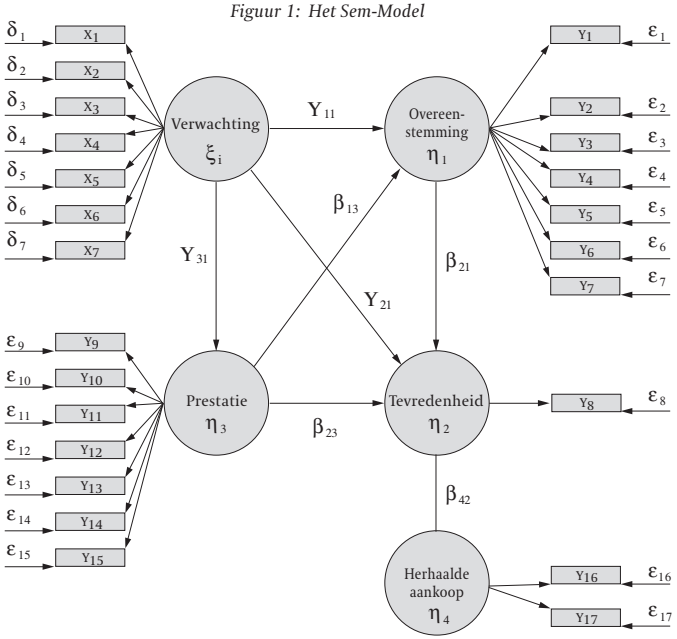
4. Structurele vergelijkingen-modellen

Behalve experimenten met gerandomiseerde toewijzing bieden ook structurele vergelijkingen-modellen goede aanknopingspunten om eventueel causale relaties vast te stellen door middel van statistische analyses. Op basis van bedrijfseconomische theorieën kunnen causale relaties vaak aannemelijk worden gemaakt met behulp van *structurele vergelijkingen- modellen*. Bedrijfseconomische theorie kan gebruikt worden voor het opstellen van een passend structurele vergelijkingen-model. Vervolgens kan met behulp van geschikte metingen (van empirische gegevens) worden vastgesteld of de veronderstelde causale relaties al dan niet waarschijnlijk zijn.
De bedrijfseconomische theorie kan bij het zoeken naar causale effecten bovendien helpen door identificatie en interpretatie van de gemeten effecten. Zonder interpretatie vernauwt onze kennis zich tot een aantal zuiver gemeten, maar verder ongerelateerde relaties. Theorieën die consistent zijn met die gemeten relaties, suggereren nieuwe inzichten en zetten aan tot nieuw onderzoek.

Ongeveer twintig jaar geleden ben ik voor het eerst in contact gekomen met structurele vergelijkingen-modellen bij een sollicitatiegesprek aan de Rijksuniversiteit Leiden. Ik had er nog nooit van gehoord, hoewel ik dat in dat gesprek uiteraard niet al te duidelijk kon laten blijken. Sindsdien heb ik mijn onderzoekstijd aan de Rijksuniversiteit Leiden en later de Vrije Universiteit Amsterdam naar schatting voor driekwart besteed aan het ontwikkelen van nieuwe theorie voor de verbetering van structurele vergelijkingen-modellen en aan het toepassen van deze structurele vergelijkingen-modellen op bedrijfseconomische vraag-stellingen.
Aan de hand van één van die toepassingen, namelijk een toepassing waarover is gerapporteerd in een artikel van Van Montfort, Masurel en Van Rijn (2000), kan de werking van een structurele vergelijkingen-model worden uitgelegd. In het betreffende onderzoek werd verondersteld dat de tevredenheid over een bepaalde dienstverlening gebaseerd is op de

feitelijke prestatie van die dienstverlening, alsmede op de mate waarin de verwachte en de feitelijke prestatie met elkaar overeen komen. Deze relatie was al aangetoond voor tastbare producten, maar nog niet voor producten in de dienstverlening.
Voor het aantonen van het veronderstelde verband werd een enquête uitgezet onder enkele honderden cliënten van een bank die ongeveer een jaar geleden een lening hadden afgesloten. Aan de personen werden over elk van de onderwerpen Verwachting, Prestatie en Overeenstemming zeven vragen gesteld. Verder werd er een vraag gesteld over Tevredenheid en volgden er twee vragen over Herhaalde Aankoop. Door middel van de enquête kregen wij enkele honderden antwoorden op deze 24 vragen, die als variabelen of indicatoren kunnen worden beschouwd. De indicatoren van Verwachting, Prestatie, Overeenstemming, Tevredenheid en Herhaalde Aankoop kunnen gebruikt worden voor het structurele vergelijkingen-model uit onderstaande figuur.

Figuur 1: Het Sem-Model



De zeven indicatoren voor Expectation vormen in bovenstaand plaatje een factor-model. Dit factor-model wordt ook wel een meetmodel genoemd, omdat de zeven indicatoren allemaal met een soort meetfout een indicatie (of meting) geven van Expectation. In het structurele vergelijkingen-model kunnen we zo in totaal vijf meetmodellen onderscheiden, die achtereenvolgens betrekking hebben op Verwachting, Prestatie, Overeenstemming, Tevredenheid en Herhaalde Aankoop. Het blijkt dat elk van deze vijf meetmodellen netjes een gezamenlijke latente factor oplevert. Deze latente factoren zijn niet direct observeerbaar.

Als we kijken naar de schattingen van de relaties tussen enerzijds Prestatie en Overeenstemming en anderzijds Tevredenheid, die overigens overeen komen met de pijltjes in Figuur 1, blijken deze relaties significant positief te zijn (zie Tabel 1). Dat is interessant. Met dit model zijn we blijkbaar in staat om zowel het directe verband tussen Prestatie en Tevredenheid aan te tonen, alsmede het indirecte verband via Verwachting en Overeenstemming. Bovendien blijkt Tevredenheid te leiden tot Herhaald Aankoopgedrag.

De juistheid van het model werd nog eens bevestigd door de gunstige waarden van de verschillende fitmaten, die aangeven of het model bij de metingen past.

Tabel 1: De schattingsresultaten voor het structurele vergelijkingen-model

Relaties	**Schattingen van regressie coëfficiënten**	**t-waarden**
Verwachting-Prestatie	0.87	7.1 *
Verwachting-Overeenstemming	- 0.46	- 2.0 *
Verwachting-Tevredenheid	- 0.07	- 0.41
Prestatie-Overeenstemming	0.18	0.68
Prestatie-Tevredenheid	0.96	4.1 *
Overeenstemm.-Tevredenheid	0.24	2.1 *
Tevredenheid-HerhaaldAank.	0.82	6.7 *
Chi-square statistic	238	
Aantal vrijheidsgraden	246	
Totale determinatie-coefficient voor structurele vergelijkingen	0.84	
GFI	0.93	
AGFI	0.92	

* significant op 0.05 niveau

Het structurele vergelijkingen-model uit Figuur 1 kan ook worden weergegeven in een algebraïsch model. Dit algebraïsche model is uiteraard gebruikt voor het schatten van de relaties tussen de verschillende bouwstenen van het model. In deze oratie zal ik het bijbehorende algebraïsche model, evenals de schattingsmethodieken van de relaties tussen de elementen van het model, achterwege laten (Van Montfort, Masurel en Van Rijn, 2000).

Het is ook nog interessant om te vermelden dat het mogelijk is om dynamische verbanden met structurele vergelijkingen-modellen te modelleren. Hierbij kan gedacht worden aan onder andere dynamische factor-modellen (Molenaar, de Gooijer en Schmitz, 1992) en state space modellen (Van Montfort en Bijleveld, 2004), die eveneens in een structurele vergelijkingen context kunnen worden geplaatst.

Zo'n twintig jaar geleden ging men er zondermeer van uit dat alle indicatoren een normale verdeling hadden. In de schattingstheorie met betrekking tot regressie-coëfficiënten, factor-ladingen, model-varianties en -covarianties, varianties van de schattingen, fitmaten en test statistics, en bij de keuze van het meest geschikte model werd steeds volledig uitgegaan van *normaal verdeelde indicatoren*. Ook werd steeds aangenomen dat de *steekproef*, bijvoorbeeld het aantal ondervraagde personen, *erg groot* is. Aan beide aannames kan echter doorgaans in praktijk niet voldaan worden. Vandaar dat er de afgelopen 20 jaar theorie is ontwikkeld, waarin deze aannames vaak niet meer noodzakelijk zijn.

Voor het schatten van regressie-coëfficiënten en factor-ladingen staan ons tegenwoordig een aantal mogelijkheden ter beschikking: we kunnen gebruik maken van hogere orde-momenten van de indicatoren (zie publicaties van A. Mooijaart en K. van Montfort); we kunnen niet-normaal verdeelde indicatoren transformeren (zie publicaties van Muthen); en we kunnen verdelingsvrije schattingsmethoden gebruiken (zie publicaties van Browne en Bentler). Bij al deze methoden krijgen we consistente schattingen van de onbekende modelparameters zonder de aanname van normaal-verdeelde variabelen te gebruiken.

Voor het verkrijgen van geschikte test statistics voor de onbekende model- parameters en de modelfit, terwijl aan de aanname van normaal

verdeelde indicatoren niet voldaan is, kan men tegenwoordig gebruik maken van *bootstrap-methoden*. Dit kan geïllustreerd worden met behulp van het volgende voorbeeld. Stel dat we willen weten hoe betrouwbaar de schattingen zijn uit het hierboven beschreven structurele vergelijkingen-model over de dienstverlening van een bank (Verwachting-Prestatie-Overeenstemming-Tevredenheid-Herhaalde Aankoop). Als we een andere steekproef hadden gebruikt, hadden we dan dezelfde schattingswaarden voor de onbekende modelparameters gekregen? Zonder aanname van normaal verdeelde indicatoren kunnen we niet, zoals twintig jaar geleden gebruikelijk was, gebruik maken van de inverse van de matrix van verwachte waarden van de tweede orde-afgeleiden.
De bootstrap-methode gaat er van uit dat we allereerst de onbekende model parameters schatten met dezelfde steekproef. Vervolgens trekken we uit de oorspronkelijke steekproef een nieuwe steekproef van dezelfde steekproefomvang. Met behulp van deze tweede steekproef schatten we opnieuw de onbekende modelparameters, zodat er nieuwe schattingen van de onbekende modelparameters worden verkregen. Nu hebben we voor elke parameter twee schattingen. Als we dit proces bijvoorbeeld 1000 maal herhalen, verkrijgen we ruim 1000 schattingen voor elke parameter. Zijn al deze schattingen ongeveer gelijk, dan hebben we blijkbaar heel betrouwbare schattingen van die parameters. Zijn deze schattingen echter heel verschillend, dan beschikken we over zeer instabiele en dus zeer slechte schattingen van die parameters. Met behulp van moderne computers kan zelfs een schattingsproces met ongeveer 1000 schattingsstappen heel snel verlopen.
Bovenstaand voorbeeld illustreert hoe wij de bootstrap-methode kunnen gebruiken om de stabiliteit van schattingen te bepalen. We kunnen de bootstrap-methode ook gebruiken om na te gaan of het gevonden structurele vergelijkingen-model wel goed is. De methode is met andere woorden nuttig voor de zogenaamde modelselectie. Daarbij gaat het om de vraag of de geobserveerde gegevens wel in overeenstemming zijn met het model dat we van tevoren hebben gekozen.
Doorgaans hanteert men bij het controleren van de juiste modelselectie de aanname dat de steekproef groot tot zeer groot is. Uiteraard gaat deze aanname in praktijk vaak niet op. Als deze aanname wel opgaat, dan geldt dat een bepaalde test statistic chi-kwadraat verdeeld is. Maar is de steekproef niet groot genoeg, dan is de test statistic niet chi-kwadraat verdeeld. Zonder de aanname van een grote steekproef kunnen we een

zogenaamde parametrische bootstrap-methode gebruiken voor het toetsen van de modelselectie. Op basis van de originele steekproef verkrijgen we schattingen van de onbekende modelparameters en bepalen we de grootte van een maat die aangeeft hoe ver de gegevens volgens het model verwijderd zijn van de geobserveerde gegevens. Dit is de zogenaamde fitmaat. We doen net alsof de gevonden schattingen van de parameters de modelparameters in de populatie zijn. Deze nieuw geconstrueerde populatie gebruiken we om bijvoorbeeld 1000 nieuwe steekproeven te trekken. Bij elke nieuw verkregen steekproef bepalen we vervolgens dezelfde fitmaat als bij de oorspronkelijke steekproef. Als we nu een bepaalde alfa-waarde kiezen, dan kunnen we op grond van de empirische verdeling van alle 1000 verkregen fitmaten bekijken of de waarde van de fitmaat met de oorspronkelijke steekproef in het verwerpingsgebied ligt of niet. Zo ja, dan verwerpen we het model. Zo nee, dan accepteren we het model.

De bootstrap-methoden zoals hierboven besproken ogen eenvoudig. Deze methoden kunnen echter behoorlijk complex worden als we met latente variabelen te maken hebben. Voor die latente variabelen, zoals bijvoorbeeld de latente factoren in het structurele vergelijkingen-model, kunnen we niet altijd eenduidig scores afleiden. Hierdoor kan een bootstrap-methode, bij latente variabelen, niet zonder meer worden toegepast. De komende tijd zal ik mij onder andere toeleggen op het toepasbaar maken van de bootstrap-methodiek op gedeelten van of wellicht het gehele structurele vergelijkingen-model.

5. Slotwoord

In deze oratie ben ik ingegaan op enkele definities van causaliteit, waarbij een model voor causaliteit is gepresenteerd. Aansluitend is aandacht geschonken aan een tweetal statistische methoden om eventuele causale relaties tussen variabelen vast te stellen: experimentele randomisatie en structurele vergelijkingen-modellen. In dat verband is tevens aandacht besteed aan het onderzoek waarmee ik mij de afgelopen jaren heb bezig gehouden en waarop ik mij in de toekomst ook nadrukkelijk zal richten.

Een half jaar geleden baarde onderzoeker John Ioannidis in het Journal of the American Medical Association opzien met zijn bevinding dat van de belangrijkste medische artikelen er na een paar jaar nog maar weinig recht overeind staan (Ioannidis, 2005). Hij stelde op basis van een statistische analyse dat de meeste medische onderzoeksresultaten onwaar zijn. Meestal is het onderzoek te klein van opzet, de significantie te laag en/of de kans op systematische vertekening te groot. De geclaimde causaliteit wordt dan onvoldoende onderbouwd.
Uiteraard is het ook de vraag of onderzoeker Ioannidis op zijn beurt de juiste conclusies heeft getrokken bij zijn statistische analyses. Een en ander geeft echter wel aan dat bij elk empirisch onderzoek en dus ook bij het bedrijfseconomische onderzoek, de weg *van meten naar weten* geen vanzelfsprekend traject is. Het is een hobbelig pad dat steeds met zorgvuldigheid zal moeten worden bewandeld.

Graag wil ik de leden van de benoemingscommissie van de leerstoel Kwantitatieve Bedrijfskundige Onderzoekstechnieken bedanken voor het vertrouwen dat zij in mij hebben uitgesproken. Met name wil ik hier noemen de hooggeleerde heren Ed Peelen en Henry Robben. Ik hoop de komende jaren een stevige bijdrage te leveren aan de verdere uitbouw van de Nyenrode Research Groep, ook wel NRG (spreek uit: energy) geheten.
Ook wil ik de hooggeleerde heren Jan de Leeuw en Ab Mooijaart bedanken voor de invloed die zij op mijn onderzoekswerk hebben gehad.

Wellicht realiseren zij zich dat zelf niet echt, maar zij hebben mij met hun aanpak wegwijs gemaakt in het verrichten van wetenschappelijk onderzoek.
Tot slot wil ik mijn ouders, mijn gezin en mijn naaste familie bedanken voor hun meelevendheid en steun.

Ik heb gezegd

Referenties

Angrist, J.D. (1990). Lifetime earnings and the Vietnam draft lottery: evidance from social security administrative records. *American Economic Review*, 80, 3, 313-336.

Bakker, B.F.M. & J. van Rooijen (2000). One figure for the supply and demand of services. Netherlands Official Statistics, Vol. 15, Spring, p. 40-46.

Van den Berg, G.J. (1990). Nonstationarity in job search theory. *Review of Economic Studies*, vol. 57, 255-277.

Berkhout, E., Heyma, A. en M. van Leeuwen (m.m.v. Berkhout, P., Brouwer, N. & C. Zijderveld) (2005). *Waarom dalen de inningspercentages van transacties en geldboetevonnissen?* Stichting voor Economisch Onderzoek der Universiteit van Amsterdam, Amsterdam (te raadplegen op http://www.seo.nl/assets/binaries/publicaties/rapporten/2005/0_801.pdf).

Bruijn, J.A. de (2002). Outputsturing in publieke organisaties. Over het gebruik van een product- en procesbenadering. *Management & Organisatie*, nr. 3, p. 5-21.

Dunn, W.N. (2004). *Public Policy Analysis. An Introduction* (third edition). Prentice Hall, New Jersey.

Eemeren, F.H. van, Grootendorst, R. & T. Kruiger (1986). *Argumentatieleer 2, Drogredenen*. Wolters-Noordhoff, Groningen.

Endresen, I.M. en Olweus, D. (2005). Participation in power sports and anti social involvement for pre adolescent and adolescent boys. *Journal of Child Psychology Psychiatry*, vol. 46, no. 5, 468-478.

Freedman, D. (1994). *From Association to causation via regression*. Research Report Statistics Department, no. 408, University of Californya, Berkeley.

Glymour, C., Scheines, R., Spirtes, P. en Kelly, K. (1987). *Discovering Causal Structure*. Academic Press, San Diego, California.

Granger, C.W.J. ((1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37, 3, 424-438.

Heckman, J.J. (1979). Sample selection bias as a specification error. *Econometrica*, vol. 47, 153-161.

Heckman, J.J. and Hotz, V.J. (1989). Choosing among alternative nonexperimental methods for estimating the impact of social programs: the case of manpower training. *Journal of the American Statistical Association*, vol. 84, 862-880.

Holland, P.W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, vol. 81, 945-970.

Imbens, I. en Van der Klaauw, W. (1995). Evaluating the cost of conscription in the Netherlands. *Journal of Business and Economic Statistics*, vol. 13, issue 2, 207-215.

Ioannidis, J.P.A. (2005). Contradicted and Initially Stronger Effects in Highly Cited Clinical Research. *Jounal of the American Medical Association*, vol. 294, no. 2, 20 pages.

LaLonde, R.J. (1986). Evaluating the economic evaluations of training programs with experimental data. *American Economic Review*, vol. 76, 4, 604-620.

Molenaar, P.C.M., de Gooijer, J.G., en B. Schmitz (1992). Dynamic factor analysis of nonstationary multivariate time series. *Psychometrika*, vol. 57, 333-349.

Riele, S. te (2002). Vertekening door non-respons, Hoe nauwkeurig zijn de uitkomsten van persoonsenquês? *Sociaal-economische maandstatistiek* (digitale uitgave CBS), april, 20-25 (te raadplegen op http://www.cbs.nl/NR/rdonlyres/C4F72666-8C9D-463D-89E1-768FD57B0555/0/2002m04v4p020art.pdf).

Ridder, G (1996). Oorzaak en gevolg. *Kwantitatieve Methoden*, nr. 51, 93-111.

Van Montfort, K. en Bijleveld, C. (2004). Dynamic Analysis of Multivariate Panel data with Nonlinear Transformations. *Journal of Mathematical Psychology*, vol. 48, issue 5, 322-333.

Van Montfort, K., Masurel, E. en Van Rijn, I (2000). Consumer satisfaction: an empirical analysis of consumer satisfaction in financial services. *The Service Industries Journal*, vol. 20, no. 3, 80-92.

Rubin, D.B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, no. 66, 688-701.

Suppes, P.A. (1970). *Probabilistic Theory of Causality*. North-Holland, Amsterdam.

Swanborn, P.G. (1977). *Aspecten van Sociologisch onderzoek*. Boom, Meppel.

Swamborn, P.G. (1991). *Basisboek sociaal onderzoek*. Boom, Meppel.

Tonkens, E. (2003). *Mondige burgers, getemde professionals. Marktwerking, vraagsturing en professionaliteit in de publieke sector*. NIZW, Utrecht.

WRR (2004). *Bewijzen van goede dienstverlening*. Amsterdam University Press, Amsterdam.