

Causaliteit, zelfselectie en geavanceerde beschrijving

Jaren geleden raakte ik verzeild op een workshop over het verband tussen kinderen krijgen en opleidingsniveau. Het gezelschap bleek vooral uit economen en econometristen te bestaan, maar er waren ook enkele sociologen en demografen. In de voordrachten van de sociologen en demografen werd opleidingsniveau opgevoerd als een van de oorzaken van al dan niet uitstellen van kinderen krijgen. Fout, helemaal fout, zeiden de economen en econometristen. Opleidingsniveau was endogeen. Vrouwen die graag kinderen wilden, kozen voor de lagere opleidingen, terwijl vrouwen die niet zo op gezinsvorming zaten te wachten naar de universiteit gingen. Opleidingsniveau van de ouders, dat kon nog net als oorzaak worden beschouwd. Maar opleidingsniveau van de vrouwen zelf, *no way*. Tja, daar sta je dan, met je verhaal over de onverenigbaarheid van de rollen van student en moeder, die ervoor zorgt dat vrouwen pas na hun opleiding aan kinderen beginnen. Die vrouwen hadden dat uitstel allang in hun hoofd en hebben zichzelf geselecteerd voor een bepaald opleidingsniveau. Ik vond dat destijds een hele rare voorstelling van zaken. Ik geloofde niet dat vrouwen van 16, 17, 18 jaar, meisjes nog, op die manier te werk gingen bij het kiezen van een opleiding (of zelfs 12 jaar, want de middelbare school sorteert al voor). Inmiddels ben ik wel een beetje gewend aan de redenering en vind ik 'm ook wel interessant. Maar ik geloof nog steeds dat beslissingen over opleidingsniveau voorafgaan aan beslissingen over kinderen krijgen, terwijl de omkering niet meer kan zijn dan een heel klein deel van het verhaal. Ik gebruik expres het woord 'geloof', want onderzoek naar het beslissingsproces ken ik niet.

Omgekeerde causaliteit en zelfselectie: elke sociaal-wetenschapper die wel eens regressiemodellen gebruikt, heeft ermee te maken. Een correct gebruik van regressiemodellen vereist immers dat er een oorzakelijk verband is tussen de onafhankelijke variabelen en de afhankelijke variabele, terwijl dat in de praktijk vaak niet zo duidelijk is. Uit mijn eigen onderzoekspraktijk kan ik moeiteloos voorbeelden noemen. Worden tweeverdieners vaker huiseigenaar omdat je met twee inkomens een betere toegang hebt tot de koopsector, of vereist een in de toekomst gewenst eigen huis dat beide partners werken om de hypotheek te kunnen betalen? Bereiken bewoners van de Randstad hogere posities op de beroepsladder omdat de Randstad ze daartoe meer mogelijkheden biedt, of wonen die mensen juist daar om die hogere posities te kunnen bereiken? Zo kan ik nog wel een tijdje doorgaan. De hamvraag is natuurlijk: wat doe je, als je een variabele als onafhankelijke in een regressiemodel wilt opnemen, maar je te maken hebt met mogelijke omgekeerde causaliteit of zelfselectie?

De meest rigoureuze en voorzichtige oplossing is: bij twijfel niet doen. Zodra er een mogelijkheid bestaat dat de causaliteit ook maar enigszins (mede) de andere richting heeft, de variabele niet opnemen. Beperk je liefst tot onafhankelijke variabelen die betrekking hebben op de

ouders, de vroege jeugd, toegeschreven kenmerken, of die uit experimenten zijn verkregen. Zelfs paneldata brengen geen oplossing, want mensen kunnen anticiperen. De 'niet doen'-strategie zou ik bepleiten als de omgekeerde causaliteit minstens even plausibel is als de bedoelde, of als er een goede alternatieve onafhankelijke variabele is waarvoor het bezwaar niet geldt. Maar als we deze strategie al te snel en vaak toepassen kunnen we heel veel onderzoek niet doen en moeten we wel erg veel kinderen met het badwater weggooien. Zie het voorbeeld van opleidingsniveau en kinderen krijgen.

Een tweede mogelijkheid is het toepassen van technische middelen. Er zijn er aardig wat. Zelfselectie kunnen we te lijf met Heckman-tweestapsmodellen. Dan wordt de kans om te behoren tot een bepaalde selectieve categorie gemodelleerd met een selectievergelijking, waarin een of meer variabelen zijn opgenomen die niet in de substantiële vergelijking zitten. Met die kans wordt rekening gehouden in de substantiële vergelijking waarin de onderzoeker eigenlijk is geïnteresseerd. De Instrumentele-Variabelenbenadering is een soortgelijke techniek. Ook hierin worden verschillende modellen gebruikt, waarbij sommige variabelen wel in het ene en niet in het andere model worden gebruikt. Voor longitudinale data kunnen simultane modellen worden geschat met aML (*multiprocess multilevel modeling*). Daarmee kunnen simultaan effecten worden geschat van externe covariaten op het optreden van verschillende gebeurtenissen, en van het optreden van die verschillende gebeurtenissen op elkaar.

Technische middelen kunnen soms heel behulpzaam zijn, maar ze hebben wel belangrijke nadelen. Ze lossen de problemen naar mijn overtuiging nooit helemaal op. Het blijven technieken en je kunt inhoudelijke problemen niet goed te lijf gaan met technische oplossingen (vrij naar Matthijs Kalmijn in zijn toespraak bij het afscheid van Wil Arts van de Universiteit van Tilburg). De bedoelde technieken zijn vaak moeilijk te begrijpen. En de modellen zijn vaak instabiel en/of moeilijk te identificeren, waardoor de resultaten niet altijd te vertrouwen zijn.

Er is ook een derde mogelijkheid. We kunnen de causaliteitsproblemen signaleren en er in de interpretatie van de resultaten aandacht aan besteden, geholpen door een zorgvuldige opbouw van de modellen (met en zonder de dubieuze variabele, met verschillende alternatieven, met goedgekozen subcategorieën respondenten). We kunnen daarbij wat bescheidener zijn in de pretenties die we hebben met de modellen. Kan het model echt worden beschouwd als een causaal model? Of is het eerder te beschouwen als een geavanceerde beschrijving van samenhangen (een vrije vertaling van wat econoom en demograaf Arnstein Aassve eens noemde *sophisticated description*)?

Een vierde mogelijkheid is de fraaiste, maar zeker niet de makkelijkste. We kunnen theorieën aanboren die juist de omgekeerde causaliteit poneren, en zoeken naar een zo scherp mogelijke toetsing van hypothesen afgeleid van die theorieën. Dit kan de vorm krijgen dat we op Popperiaanse wijze op zoek gaan naar falsificaties van hypothesen die in beide richtingen redeneren. Misschien is het toch niet zo gek als iemand eens een mooi onderzoek uitvoert naar de besluitvorming van meisjes en hun ouders en leraren met betrekking tot de keuze van een opleiding en de rol daarin van plannen om kinderen te krijgen. Nog rond 1970 kreeg mijn oudere zuster bij de keuze van haar vakkenpakket van een leraar te horen dat het er niet zo toe deed, want: 'Och, jij gaat waarschijnlijk toch trouwen'. Wie weet is er wel iets te achterhalen van de

mate waarin ook vandaag de dag dit soort redeneringen nog een rol spelen bij de schoolkeuze van meisjes.

Sommige sociaal-wetenschappelijke tijdschriften zullen bij twijfelachtige causaliteit wellicht al gauw van auteurs eisen dat zij technische middelen toepassen. De redactie van *Mens & Maatschappij* eist dat maar zelden, al hebben wij niets tegen geavanceerde technische hulpmiddelen als deze goed beredeneerd en met kennis van zaken worden toegepast. In duidelijke gevallen van omgekeerde causaliteit vragen wij de desbetreffende onafhankelijke variabele te verwijderen. Maar als een mogelijke omgekeerde causaliteit maar een klein deel van het verhaal is, hebben wij liever dat u de lezers met een goede argumentatie helpt uw resultaten naar waarde te schatten, dan dat u ze alleen maar met een batterij vergelijkingen om de oren slaat.

Clara H. Mulder