De meting van de effectiviteit van arbeidsmarktprogramma's

Joost Bollens
HIVA
K.U.Leuven

7-2007

WSE Report

Steunpunt Werk en Sociale Economie
Parkstraat 45 bus 5303 – 3000 Leuven
T:32(0)16 32 32 39 F:32(0)16 32 32 40
steunpuntwse@econ.kuleuven.be
www.steunpuntwse.be

De meting van de effectiviteit van arbeidsmarktprogramma's

Joost Bollens

Een onderzoek in opdracht van de Vlaamse minister van Werk, Onderwijs en Vorming, in het kader van het VIONA-onderzoeksprogramma

Met ondersteuning van het departement Werk en Sociale Economie en het ESF
ESF: de Europese bijdrage tot de ontwikkeling van de werkgelegenheid door inzetbaarheid, ondernemerschap, aanpasbaarheid en gelijke kansen te bevorderen en door te investeren in menselijke hulpbronnen

Bollens, Joost

De meting van de effectiviteit van arbeidsmarktprogramma's
Joost Bollens – Leuven: HIVA - Katholieke Universiteit Leuven. Steunpunt Werk en Sociale Economie, 2007, 26 p.

ISBN-97 890-8873-007-8

Steunpunt Werk en Sociale Economie
Parkstraat 45 bus 5303 – B-3000 Leuven
T:32(0)16 32 32 39 - F:32(0)16 32 32 40
steunpuntwse@econ.kuleuven.be
www.steunpuntwse.be

Inhoudsopgave

Inleiding

Aan arbeidsmarktprogramma's, zoals het geven van opleiding aan werklozen of het geven van loonkostsubsidies bij het aanwerven van werklozen, worden aanzienlijke budgetten besteed. Dit laatste maakt dat het belangrijk is om na te gaan in welke mate ze hun doel bereiken: is hun effect inderdaad hetgene dat beoogd werd?

Bij het meten van die effectiviteit duiken al vlug een aantal vragen en problemen op, waarop in de literatuur verschillende antwoorden zijn gegeven. Dit artikel geeft een gestructureerd overzicht van dit stuk van de evaluatieliteratuur. Gezien het methodologisch karakter van de vraagstukken, werd er gekozen om toch in zekere mate gebruik te maken van een meer formele taal. Dit dwingt de gebruiker er toe om meer expliciet na te denken over alle gemaakte veronderstellingen, maar kan bovendien tot een beter inzicht leiden omdat op deze manier duidelijk wordt dat de basisstructuur van de diverse behandelde methodes eigenlijk telkens zeer gelijkaardig is, en het verschil wordt gemaakt door soms subtiele details.

De originaliteit van dit artikel ligt zeker niet bij de gemaakte selectie van methodes, noch bij de meer formele voorstelling ervan: in de literatuur kunnen verscheidene goede overzichten worden gevonden, waar dankbaar gebruik van werd gemaakt, en die op hun beurt alle in meerdere of mindere mate schatplichtig zijn aan het werk van Nobelprijswinnaar James Heckman.

De originaliteit ligt veeleer in het feit dat we geprobeerd hebben om de meer formele resultaten telkens uitgebreid in woorden te duiden, en op zoek te gaan naar de achterliggende intuïtie. De doelstelling die daarbij voor ogen werd gehouden is om aan beleidsvoorbereiders en beleidsmakers, die niet dagelijks met statistiek bezig zijn, een gedegen overzicht te bieden van wat er momenteel aan evaluatiemethoden op de markt is, met inbegrip van de mérites maar vooral ook beperkingen van de diverse methodes.

Om de lectuur niet al te theoretisch te maken, worden op diverse plaatsen recente onderzoeken besproken die de toepassing van een bepaalde methode illustreren.

In wat volgt wordt veelvuldig gebruik gemaakt van voorwaardelijke verwachtingen. In een appendix achteraan wordt getracht om dit begrip uit de statistische wiskunde in woorden uit te leggen.

1. Het evaluatieprobleem

Het gebruikmaken van of deelnemen aan een arbeidsmarktmaatregel gericht op het activeren van werkzoekenden zoals vb. beroepsopleiding, training van het zoekgedrag of tijdelijke werkervaring, wordt in de evaluatieliteratuur, naar analogie met de medische terminologie, meestal een 'treatment', een behandeling genoemd.

Stel voor de eenvoud dat een maatregel de positie van de deelnemers op één bepaalde dimensie, zoals het vinden van werk, of het vinden van meer duurzaam of beter betaald werk, wil verbeteren (men kan dit zonder verlies van algemeenheid uitbreiden naar een vector met verschillende dimensies).

Voor ieder individu i zijn er dan twee mogelijke uitkomsten, Y^1 en Y^0, waarbij Y^1 aangeeft wat de situatie is in het geval van een behandeling (of, met andere woorden, waarbij Y^1 aangeeft wat de arbeidsmarktsituatie van individu i zou zijn na deelname aan een bepaalde maatregel), en waarbij Y^0 de situatie weergeeft als i de behandeling niet ondergaat. Het effect van de behandeling voor een individu i is dan gelijk aan het verschil tussen zijn potentiële uitkomsten:

$$\Delta_i = Y_i^1 - Y_i^0 \qquad (1)$$

Als D dan vervolgens een binaire indicatorvariabele is die gelijk is aan 1 indien een individu de behandeling onderging, en gelijk is aan 0 in afwezigheid van de behandeling, observeert men voor ieder individu i de volgende uitkomst:

$$Y_i = D_i Y_i^1 + (1 - D_i)\, Y_i^0 \qquad (2)$$

Dit geeft het fundamentele evaluatieprobleem weer: voor wie werd behandeld, observeert men Y^1, voor wie de behandeling niet onderging, observeert men Y^0. Het is m.a.w. onmogelijk dat een individu tegelijkertijd wel en niet werd behandeld, zodanig dat men voor één en hetzelfde individu nooit tegelijkertijd Y^0 en Y^1 zal observeren, en het verschil in vergelijking (1) niet (direct) kan worden geschat. Het niet-geobserveerde deel in vergelijking (1) wordt de counterfactual genoemd.

Individuele effecten van een behandeling zoals in vergelijking (1) kunnen dus niet worden geschat. In de praktijk moet men zich dan ook richten op de populatiegemiddelde effecten van de behandeling i.p.v. op de individuele effecten. Hierbij zijn dan weer verschillende keuzes mogelijk. In een eerste benadering wil men te weten komen wat het effect van de behandeling zou zijn voor een willekeurig uit de populatie getrokken individu. Men noemt dit het (populatie) gemiddeld behandelingseffect, en het wordt afgekort als ATE ('average treatment effect'). ATE wordt gegeven als het verschil tussen de verwachte uitkomsten na deelname en na niet-deelname:[1]

$$\Delta_{ATE} = E(\Delta) = E(Y^1) - E(Y^0) \qquad (3)$$

Daarnaast wordt ook dikwijls gewag gemaakt van het gemiddelde behandelingseffect voor de behandelden, afgekort door ATT ('average effect of treatment on the treated'), wat het effect van behandeling weergeeft voor zij die effectief deelnamen.

$$\Delta_{ATT} = E(\Delta \mid D=1) = E(Y^1 \mid D=1) - E(Y^0 \mid D=1) \qquad (4)$$

Een beperking van ATT is dat, zoals het begrip trouwens zelf aangeeft, het geschat resultaat niet zo maar kan worden veralgemeend naar personen uit het beoogde doelpubliek die vooralsnog niet deelnamen aan de maatregel. Zeker voor programma's waarbij de deelname vrijwillig is, moet men zich bewust zijn dat het eerder onwaarschijnlijk is dat de geschatte effecten voor een deelnemersgroep die zichzelf selecteerde, kunnen worden veralgemeend naar personen die zich hadden kunnen selecteren en dat toch niet gedaan hebben.

Zoals Heckman stelt (Heckman 1997), is ATE wellicht niet zo heel relevant vanuit beleidsstandpunt, omdat het ook betrekking heeft op dat deel van de bevolking waarvoor het programma nooit was voorzien. Deze kritiek kan wellicht deels worden ondervangen door te specifiëren dat de populatie waarop het ATE betrekking heeft, de volledige doelgroep is die werd beoogd met de maatregel. Op dat moment krijgt het ATE een duidelijke betekenis, vb. als men zou overwegen om een maatregel met vrijwillige deelname verplicht te maken. ATT heeft dan weer een duidelijke rol te spelen als men een kosten-batenanalyse wil maken van een bestaande maatregel (Heckman et al. 1999).

In de evaluatiepraktijk wordt vooral ATT geschat. Als we terugkijken naar vergelijking (4), is duidelijk dat vooral de tweede term van de vergelijking het evaluatieprobleem weer scherp op de voorgrond stelt, $E(Y^0 \mid D=1)$ geeft immers weer wat voor de deelnemersgroep (i.e.: D=1) de gemiddelde

1 In een appendix wordt de betekenis van de symbolen E(Y) en E(Y | X) uitgelegd.

uitkomst (i.e.: E(Y)) zou zijn geweest, wanneer zij niet hadden deelgenomen (i.e. Y^0). Dit is duidelijk een counterfactual die nooit kan geobserveerd worden. Als evenwel zou gelden dat:

$$E(Y^0 \mid D=1) = E(Y^0 \mid D=0) \qquad (5)$$

is het probleem opgelost, want dan kan men de gemiddelde uitkomst van een controlegroep (van niet-deelnemers, D=0) hanteren als schatting voor de tweede term uit vergelijking (4). De vraag is dan ook wanneer voldaan is aan de gelijkheid in vergelijking (5). Binnen de zogenaamde experimentele benaderingen wordt er alles aan gedaan om dit resultaat te verkrijgen. In de volgende sectie wordt dit meer van nabij bekeken.

2. Experimentele benadering

2.1 Aanpak

Deelnemers aan een arbeidsmarktmaatregel kiezen dikwijls zelf bewust voor deze deelname ('zelfselectie'), en/of worden uitgekozen door de programmaverantwoordelijken of door medewerkers van de PES ('Public Employment Service'). Omgekeerd geldt dan ook dat niet-deelnemers bewust kiezen om niet deel te nemen, en/of bewust niet worden uitgekozen door derden. Deze selectieeffecten impliceren dat de groep van deelnemers (D=1) en de groep van niet-deelnemers (D=0) mogelijk erg verschillen wat betreft de verdeling van diverse kenmerken X (zoals opleidingsniveau, leeftijd, motivatie...). Dit vormt meteen de grootste bedreiging voor de gelijkheid in vergelijking (5): als de groep D=1 erg verschilt van de groep D=0, is het aannemelijk dat hun gemiddelde arbeidsmarktuitkomsten E(Y), zelfs in afwezigheid van deelname (Y^0), ook zullen verschillen, en dan is niet langer voldaan aan voorwaarde (5).

Binnen de experimentele benaderingen wordt dit probleem opgelost door enkel te werken met personen uit de groep waarvoor geldt dat D=1, d.w.z. door zich enkel te richten op die personen die zelf bewust kozen voor deelname en/of die door anderen (programmaverantwoordelijken, PES,..) werden uitgekozen. Via een systeem van toevalstoewijzing ('random assignment', vandaar ook 'randomisering') wordt vervolgens de helft van de groep D=1 toegelaten tot de maatregel, en de andere helft van de groep D=1 wordt uitgesloten van deelname. Stel dat R=1 voor personen die in de maatregel worden gerandomiseerd (de experimenteel behandelde groep), en dat R=0 voor personen die worden uitgerandomiseerd (de experimentele controlegroep). Het is dan mogelijk om de resultaten Y^0 van deze controlegroep R=0 te beschouwen als een goede benadering voor de resultaten die konden worden verwacht van de behandelde groep R=1, wanneer die niet behandeld zou zijn. Heel het opzet van toevalstoewijzing zorgt er immers voor dat het selectieprobleem, waarvan sprake bij het begin van deze sectie, zich niet stelt. Het deelnameproces, en de selectie die daar uit volgt, wordt immers toevallig gemaakt. Als experimentele en controlegroep voldoende groot zijn, zal bij een correct uitgevoerde toevalstoewijzing de verdeling van kenmerken X (die ook de uitkomst mee bepalen) vergelijkbaar zijn tussen beide groepen.

Heckman et al. 1999 stellen terecht dat er ook hier een aantal onderliggende veronderstellingen zijn. Laat '*' wijzen op de aanwezigheid van een experiment. Het geheel (Y^{1*}, Y^{0*},D^*) wijst dan op een experimentele situatie, terwijl men (Y^1, Y^0, D) heeft wanneer het programma normaal functioneert. De essentiële veronderstelling opdat men via toevalstoewijzing het evaluatieprobleem kan oplossen bij het schatten van ATT is de volgende, aldus deze auteurs:

$$E(Y^{1*} - Y^{0*} \mid X, D^*=1) = E(Y^1 - Y^0 \mid X, D=1) \qquad (6)$$

waarbij X staat voor de kenmerken van de betrokken personen. Deze gelijkheid stelt dat (het verschil van) de gemiddelden van de behandelde en controlegroep onder toevalstoewijzing gelijk is aan (het verschil van) de (niet-geobserveerde) populatiewaarden. Met andere woorden, het mag niet zo zijn dat het opzetten van een experiment ertoe leidt dat men een andere impact (Y^{1^*}-Y^{0^*}) bekomt dan degene die men zou bekomen hebben mocht het experiment niet hebben plaatsgevonden. Daarnaast mag het plaatsvinden van het experiment ook niet leiden tot een wijziging van het participatieproces aan het programma, d.w.z. het proces dat leidt tot de opdeling tussen $D^*=1$ en $D^*=0$, mag niet afwijken van het proces dat in normale (niet-experimentele) omstandigheden leidt tot de opdeling tussen D=1 en D=0.

Een voldoende (maar niet noodzakelijke) voorwaarde opdat dit het geval zou zijn wordt geboden door de twee volgende voorwaarden:

$$E(Y^{1^*} \mid X, D^*=1) = E(Y^1 \mid X, D=1) \quad (7)$$

en

$$E(Y^{0^*} \mid X, D^*=1) = E(Y^0 \mid X, D=1) \quad (8)$$

Als (7) waar is, mag het resultaat van de experimentele groep beschouwd worden als het resultaat van deelname aan het programma:

$$E(Y^1 \mid X, D=1, R=1) = E(Y^1 \mid X, D=1) \quad (9)$$

Analoog kan op basis van (8) via de controlegroep de counterfactual worden geïdentificeerd:

$$E(Y^0 \mid X, D=1, R=0) = E(Y^0 \mid X, D=1) \quad (10)$$

Een en ander betekent dat in een experimentele benadering de impact van het programma eenvoudig kan worden bepaald door twee gemiddelden van elkaar af te trekken, en het resultaat kan (voor een gegeven X) worden samengevat met één enkel getal dat gemakkelijk te interpreteren is:

$$\Delta_{ATT}^{EX} = E(\Delta \mid X, D=1) = E(Y^1 - Y^0 \mid X, D=1) \quad (11)$$

Deze eenvoud vormt één van de aantrekkelijke kanten van de experimentele benadering, zeker als men dit vergelijkt met de min of meer bewerkelijke statistische procedures die nodig zijn bij de meeste niet-experimentele methodes (zie verder).

2.2 Valkuilen en problemen

Het ideaaltypische sociale experiment is eenvoudig qua opzet, vergt geen ingrijpende statistische bewerkingen, en geeft resultaten die eenvoudig uit te leggen zijn. Deze benadering wordt dan ook door sommigen beschouwd als de 'gouden standaard' van het evaluatieonderzoek. Nochtans is ook een experimenteel opzet niet per definitie vrij van problemen. In wat volgt, worden een aantal elementen bekeken die de validiteit van een experiment kunnen aantasten of zelfs ondergraven.

- *afhakers ('drop -out')*: in het algemeen is het mogelijk dat een zeker aandeel van de personen die werden ingerandomiseerd (D=1 en R=1) de bestudeerde behandeling niet of niet volledig ondergaan. In dat geval is het niet langer mogelijk om $E(\Delta \mid X, D=1)$ te identificeren door het vergelijken van gemiddelden (11 geldt niet langer). In plaats van de impact van het programma, meet men dan het gemiddeld effect van het krijgen van een aanbod tot behandeling ('intent to treat'). Volgens Heckman et al. 1999 is dit overigens een beleidsrelevante parameter, aangezien

men ook in ieder bestaand programma de facto afhakers zal hebben. Op te merken valt verder dat als de kans op afhaken samenhangt met individuele kenmerken, of samenhangt met de door de deelnemers gemaakte inschatting van de baten van een verdere deelname, het toevalskarakter van de toewijzingsregel wordt ondergraven, en er een selectievertekening ontstaat.[2]

- *vertekening door substitutie ('substitution bias')*: een vergelijkbaar probleem kan zich ook voordoen bij de controlegroep. Wanneer met name personen uit de controlegroep elders toch een behandeling ondergaan die kan worden beschouwd als een substituut voor de behandeling waarvoor ze werden uitgerandomiseerd, is niet langer voldaan aan voorwaarde (8). In zekere zin geraakt de controlegroep dan 'besmet' door eenheden die toch werden behandeld, vandaar dat men dit soms ook omschrijft als een 'contamination bias'. Door Heckman et. al. 1999 wordt opgemerkt dat men zich juist in de context van een experimenteel opzet kan verwachten aan dit probleem, aangezien de toevalstoewijzing met zich meebrengt dat de behandeling wordt ontzegd aan een groep van personen waarvan mogelijk een gedeelte nochtans bewust had gekozen voor de behandeling. De kans is niet denkbeeldig dat deze laatste groep na uitrandomisering op zoek zal gaan naar een alternatief.

- *vertekening door de randomisering ('randomization bias')*: Heckman en Smith (Heckman and Smith 1995) behandelen problemen die voortvloeien uit het proces van toevalstoewijzing, en die bijgevolg, in tegenstelling tot de meeste andere in deze subsectie behandelde problemen, specifiek verbonden zijn aan een experimenteel opzet. Vertekening door de randomisering houdt in dat het doorgaan van het experiment zelf ertoe leidt dat andere resultaten worden bekomen en/of andere personen in het programma terecht komen dan wat het geval zou zijn onder normale omstandigheden, zodanig dat niet langer wordt voldaan aan de voorwaarden (6), (7) of (8).

Bij een klassiek dubbel blind experiment, zoals vb. gebruikelijk bij het testen van een nieuw geneesmiddel, weet een persoon die deelneemt aan het experiment niet of zij het product, waarvan de werking wordt onderzocht, krijgt toegediend, dan wel of ze een placebo krijgt. De persoon weet met andere woorden niet of zij tot de experimentele of tot de controlegroep behoort. Het dubbel blinde karakter van het experiment heeft betrekking op het feit dat ook degenen die de producten toedienen niet weten welke de placebo's zijn en welke niet. Dit soort van condities is moeilijk, zo niet onmogelijk te realiseren in het geval van een sociaal experiment. Beschouw eerst degenen die de behandeling geven (vb. opleidingsinstelling) en/of bepalen wie welke behandeling krijgt. Het is moeilijk te vermijden dat niet minstens sommigen uit deze groep op de hoogte zijn van het feit dat een experiment plaatsvindt. Alleen al het proces van toevalstoewijzing zal bij de evaluatie van reeds lopende programma's een en ander duidelijk maken. Immers, als men in een bestaand programma zijn intake en selectieproces constant houdt, zal men na randomisering slechts de helft van het cliënteel moeten behandelen dan wat in normale omstandigheden het geval is. De betrokkenheid van programmaverantwoordelijken en -uitvoerders bij dit proces zorgt er alvast voor dat deze kant niet echt blind is. En aangezien net deze personen het experiment kunnen percipiëren als een evaluatie van hun eigen werking, kan niet worden uitgesloten dat de kennis over het feit dat het experiment loopt, invloed zal hebben op het gedrag, men zal dan vb. extra zijn best doen om goede resultaten te boeken met de behandelden, of zal men proberen om meer selectief te zijn bij het toelaten van klanten, waardoor voorwaarde (7) in het gedrang komt. Men kan zich overigens de vraag stellen of het bij wijze van spreken op halve capaciteit draaien sowieso ook geen invloed zal hebben op het experimenteel resultaat. Als men, om dit probleem te vermijden, de intake verdubbelt om na randomisering toch op volledige capaciteit te kunnen blijven werken, ontstaat dan weer het gevaar op een beïnvloeding van het selectieproces, men gaat dan

2 Afhaken, dat enkel voorkomt binnen de experimentele groep, is te onderscheiden van attritie: bij attritie verdwijnen leden van de experimentele of controlegroep op een bepaald moment uit het vizier van de onderzoeker. Dit geeft vooral problemen bij het opvolgen van effecten over een langere periode.

onder experimentele condities immers op een andere manier selecteren dan onder normale omstandigheden.

Ook aan de kant van de behandelden kunnen er verschillen zijn tussen normale omstandigheden en de experimentele condities. Zeker binnen de groep van de uitgerandomiseerden (D=1,R=0) is blindheid niet altijd mogelijk. Het probleem moet wellicht niet overroepen worden, maar er kan niet worden uitgesloten dat het gevoel dat men onfair of ongelijk behandeld is, gevolgen heeft voor de arbeidsmarktresultaten Y^{0^*}.

Als al deze potentiële problemen van een (sociaal) experiment samen worden beschouwd, moet worden besloten dat in werkelijkheid de vermeende eenvoud van opzet en uitvoering van deze werkwijze zeker niet altijd kan worden gerealiseerd. De experimentele benadering blijft ongetwijfeld een aantal aantrekkelijke eigenschappen hebben. Tegelijkertijd zou moeten duidelijk zijn dat bij het opzetten van een experiment veel aandacht moet worden besteed aan een design dat een aantal van de vermelde problemen vermijdt, of dat toelaat om de invloed van deze problemen te minimaliseren en bij te sturen. Dit gegeven gaat ten koste van de eenvoud, en is typisch ook vrij duur. Het bijsturen van problemen van selectievertekeningen, die omwille van diverse reden ook bij een sociaal experiment kunnen opduiken, impliceert bijna onvermijdelijk dat de experimentele benadering zal moeten worden aangevuld met niet-experimentele technieken.

Blijft tot slot nog de ethische en juridische kant van de zaak. Omwille van het principe van gelijke behandeling van personen in een vergelijkbare situatie, wordt er dikwijls bezwaar gemaakt tegen het opzetten van een sociaal experiment. Wellicht moet hier toch de nodige nuancering aan de dag gelegd worden. Voor veel van de gangbare actieve arbeidsmarktmaatregelen kunnen we eigenlijk niet met zekerheid zeggen of ze wel zo veel baat opleveren voor de diverse doelgroepen die er momenteel aan participeren. Sommige combinaties van maatregelen en doelgroepkenmerken zijn wellicht zelfs schadelijk voor de deelnemers (in de zin dat hun arbeidsmarktverloop gunstiger zou zijn geweest, mochten ze niet hebben deelgenomen). Gegeven deze fundamentele onzekerheden, kan men ethisch gezien weinig bezwaar hebben tegen het op toevallige wijze uitsluiten van geïnteresseerde deelnemers: misschien zal ex post, na afloop van het experiment blijken dat zij bij deze gang van zaken verloren hebben, maar ex ante kan dit niet worden bepaald, een experiment zal immers juist typisch worden opgezet als er geen zekerheid bestaat over het effect van een maatregel.

3. Niet-experimentele benaderingen

3.1 Probleemstelling

Wanneer het niet mogelijk is om een experimentele studie op te zetten, of wanneer dit om welke reden dan ook ongewenst is, kan men een beroep doen op de zogenaamde niet-experimentele methoden. Onder de noemer niet-experimentele methodes valt een groot gebied van diverse benaderingen die verschillen in gesofisticeerdheid, en in de eisen die ze stellen aan de kwaliteit van de benodigde gegevens.

Wat betreft die kwaliteit van de benodigde gegevens, kan men verschillende dimensies onderscheiden. Zo is er de vraag op hoeveel meetmomenten de data betrekking hebben. Bij gewone doorsnedegegevens ('cross-sectie') wordt typisch een momentopname gemaakt van een (steekproef uit een) populatie, en is het aantal meetmomenten dan ook gelijk aan één. Bij longitudinale gegevens ('panel') wordt typisch eenzelfde (steekproef uit een) populatie beschouwd op twee of meer verschillende momenten in de tijd (vb. voorafgaand en volgend op de deelname aan een maatregel). Tussenvormen zijn de herhaalde doorsnede, waarbij doorsnedegegevens beschikbaar

zijn over twee of meer momenten in de tijd, en waarbij het verschil met panelgegevens zit in het feit dat deze doorsneden niet betrekking hebben op dezelfde steekproef (al kan er mogelijk wel overlap zijn). Een andere tussenvorm zijn retrospectieve doorsnedegegevens, waarbij, vb. aan de hand van een survey, aan de respondenten wordt gevraagd om niet alleen hun huidige situatie te beschrijven, maar ook hun situatie op diverse punten in het verleden. Omwille van geheugenproblemen etc. zijn dergelijke retrospectieve gegevens doorgaans minder betrouwbaar dan panelgegevens.

Een andere dimensie van de kwaliteit van de beschikbare gegevens heeft betrekking op de vraag hoe ruim en hoe gedetailleerd ze zijn, en in welke mate ze beantwoorden aan dat wat nodig is om de gestelde onderzoeksvraag te beantwoorden. Zo zijn administratieve gegevens (type datawarehouse KSZ) dikwijls aantrekkelijk omdat ze de volledige populatie bestrijken, en longitudinaal onderzoek mogelijk maken, maar ze hebben ook duidelijk een aantal nadelen. Aangezien deze gegevens niet omwille van onderzoeksdoeleinden worden ingezameld, komen de gebruikte concepten en definities niet noodzakelijk overeen met datgene wat vanuit onderzoeksstandpunt wenselijk zou zijn. Bovendien zal men er zelden of nooit informatie vinden m.b.t. persoonskenmerken die wellicht vanuit administratief standpunt niet relevant zijn, zoals vb. de motivatie van de persoon of de spreekvaardigheden en het uiterlijk van de persoon. Het probleem stelt zich overigens niet alleen m.b.t. deze inderdaad wellicht meer subjectieve inschattingen. Ook een meer objectief gegeven, zoals het scholingsniveau, is bij veel administratieve databanken niet voorhanden.

Bepaalde verschillen tussen personen, waarover evenwel informatie beschikbaar is, noemt men geobserveerde verschillen. Als in een gegevensverzameling m.b.t. deelnemers aan een bepaalde arbeidsmarktmaatregel het scholingsniveau van de deelnemers is opgenomen, is het scholingsniveau een geobserveerd verschil. Als er geen informatie beschikbaar is over het scholingsniveau, is dit een niet-geobserveerd verschil.

Het uitgangsidee van de meeste niet-experimentele benaderingen is dat het mogelijk is om een groep van individuen samen te stellen die maximaal vergelijkbaar is met de groep van individuen die deelnamen aan de maatregel. Op één dimensie is er echter wel een verschil tussen beide groepen, de leden van de eerste groep, de zogenaamde vergelijkingsgroep, hebben niet deelgenomen aan de maatregel. (Om het onderscheid tussen een experimentele en een niet-experimentele benadering duidelijk te benadrukken, volgen we hier de conventie uit de literatuur, waarbij in het geval van een experiment wordt gesproken van een *controlegroep*, en in andere gevallen van een *vergelijkingsgroep*).

Als de deelnamestatus inderdaad het enige verschil is tussen beide groepen, dan volgt dat het verschil tussen de arbeidsmarktprestaties van de deelnemersgroep en die van de vergelijkingsgroep enkel en alleen een gevolg kan zijn van de deelname aan de maatregel. De arbeidsmarktprestaties van de vergelijkingsgroep kunnen in dat geval beschouwd worden als schatter voor de counterfactual.

In de werkelijkheid is het uiteraard onmogelijk om twee groepen samen te stellen die volledig identiek zijn, op één dimensie na. Nu kan de eis van de volledige vergelijkbaarheid wel gemilderd worden, in de zin dat de vergelijkbaarheid enkel betrekking heeft op alle factoren die de bestudeerde uitkomst beïnvloeden. Stel dat de te bestuderen uitkomst Y zoals voorheen de kans is op al dan niet aan het werk zijn na deelname, dan heeft de vergelijkbaarheid betrekking op alle factoren die deze kans beïnvloeden. Zoals geweten wordt deze kans bij werklozen onder meer beïnvloed door het opleidingsniveau, het geslacht, de leeftijd, de etniciteit, de werkloosheidsduur, maar wellicht ook door factoren zoals de gezondheid, het uiterlijk, het taalgebruik, de motivatie en het arbeidsethos van de werkloze, etc.

Het is dan ook eerder onwaarschijnlijk dat men ooit in staat zal zijn om een vergelijkingsgroep samen te stellen die, afgezien van de deelnamestatus, volledig vergelijkbaar is met de deelnemersgroep m.b.t. alle factoren die van invloed kunnen zijn op de bestudeerde uitkomst. Een en ander impliceert dat een niet-experimentele benadering niet kan volstaan met het louter vergelijken van de gemiddelde uitkomsten van deelnemers- en vergelijkingsgroep (een praktijk die in het geval van een sociaal experiment mogelijk wel gerechtvaardigd is).

Verschillen tussen deelnemers- en vergelijkingsgroep kunnen immers aanleiding geven tot selectie-effecten. Stel dat het effect wordt onderzocht van de deelname aan een beroepsopleiding op de kansen om werk te vinden. Naast een deelnemersgroep wordt ook een vergelijkingsgroep samengesteld. Als nu de gemiddelde motivatie om werk te vinden groter is bij de deelnemersgroep dan bij de vergelijkingsgroep, -- en het feit dat ze kiezen om deel te nemen aan een opleiding is misschien wel een indicatie voor deze hogere motivatie -- dan stelt zich wel een probleem bij de interpretatie van de resultaten: als zou blijken dat de deelnemers aan de opleiding nadien een betere gemiddelde uitkomst scoren (een hoger aandeel vindt werk) in vergelijking met de vergelijkingsgroep, dan kan men niet met zekerheid dit verschil in uitkomst toewijzen aan de deelname aan de opleiding. Mogelijk is het verschil geheel of gedeeltelijk ook op rekening van het motivatieverschil te schrijven. Of nog, zelfs al had niemand deelgenomen aan de opleiding, zelfs dan zou mogelijk over de beschouwde tijdsperiode de gemiddelde uitkomst van de 'deelnemers'-groep beter geweest zijn dat de gemiddelde uitkomst van de vergelijkingsgroep, louter en alleen omwille van het gemiddeld motivatieverschil.

Er treedt m.a.w. een selectie-effect op door het feit dat bepaalde kenmerken niet alleen de kans op deelname aan een maatregel beïnvloeden, maar ook de arbeidsmarktuitkomsten. Niet-experimentele benaderingen moeten een antwoord formuleren op dit probleem. Globaal worden er in de literatuur twee grote groepen van niet-experimentele benaderingen onderscheiden: benaderingen die kunnen worden gehanteerd wanneer er alleen sprake is van selectie op observeerbare verschillen enerzijds, en benaderingen die ook kunnen worden toegepast wanneer er sprake is van selectie op niet-geobserveerde verschillen anderzijds.

Vergelijking (4) geeft het reeds eerder vermelde gemiddeld effect van de behandeling op de behandelden:

$$\Delta_{ATT} = E(\Delta \mid D=1) = E(Y^1 \mid D=1) - E(Y^0 \mid D=1) \qquad (4)$$

Voorheen werd vastgesteld dat in een experimentele benadering het evaluatieprobleem wordt opgelost door het tweede lid van deze vergelijking, met name $E(Y^0 \mid D=1)$, gelijk te stellen aan $E(Y^0 \mid D=0)$.[3] Bij een niet-experimentele benadering daarentegen, zal dit in het algemeen niet mogelijk zijn:

$$E(Y^0 \mid D=1) \neq E(Y^0 \mid D=0) \qquad (12)$$

Als men dit negeert, ontstaat een probleem van selectievertekening, en zal de geschatte impact onzuiver zijn.

3 In de toen gebruikte terminologie werd $E(Y^0 \mid D=1)$ gelijk gesteld aan $E(Y^0 \mid D=1, R=0)$, wat kon, omdat er een controlegroep was. In de niet-experimentele situatie is er enkel een vergelijkingsgroep, en vallen we terug op $E(Y^0 \mid D=0)$.

Om de discussie verder te stroomlijnen, worden m.b.t. de uitkomsten Y^0 en Y^1 de twee volgende uitkomstenvergelijkingen voorgesteld (Blundell and Costa Dias 2002):[4]

$$Y^1_{it} = g^1_t(X_i) + U^1_{it} \qquad \text{en } Y^0_{it} = g^0_t(X_i) + U^0_{it} \tag{13}$$

waarbij i staat voor een individu, t voor een periode. De functie g(.) beschrijft het verband tussen de potentiële uitkomsten en de geobserveerde kenmerken X_i van individu i (een typisch voorbeeld van geobserveerde kenmerken in X zijn het geslacht, de leeftijd, het opleidingsniveau, enz. Als de analist geen gegevens heeft over één deze kenmerken, zit het uiteraard niet onder X, en komt het in de restterm terecht). De U^0 en U^1 zijn (niet-geobserveerde) resttermen, die staan voor enerzijds die niet-geobserveerde kenmerken van individu i die samenhangen met de uitkomst (typische voorbeelden: motivatie, uiterlijk, ...), en anderzijds voor alle andere niet-geobserveerde elementen die de uitkomst beïnvloeden, maar buiten het individu i liggen (toeval, conjunctuur,...). Daarbij wordt verondersteld dat het gemiddelde van de resttermen nul is, en dat de resttermen niet correleren met de regressoren X.

Heckman en Robb (Heckman and Robb, Jr. 1985) gaan er van uit dat de selectiebeslissing (deelnemen aan de maatregel of niet) kan worden beschreven door een indexfunctie IN_i,

$$IN_i = f(Z_i) + V_i \tag{14}$$

waarbij de (latente) index IN_i een hogere waarde aanneemt naarmate de kans op deelname toeneemt, en wordt gerelateerd aan functie f(.) van geobserveerde kenmerken Z_i en aan niet-geobserveerde variabelen V_i. Typisch zullen veel van de kenmerken in X ook aanwezig zijn in Z, maar dat hoeft niet (de werkloosheidsduur zal vb. dikwijls meespelen bij de bepaling of iemand toegang heeft tot een bepaald programma, en zit dan bij de Z kenmerken. Als de werkloosheidsduur ook de uitkomst meebepaalt, zit die ook bij de X kenmerken).

Er is dan sprake van deelname aan het programma, d.w.z. $D_i = 1$, als $IN_i > 0$. Als $IN_i \leq 0$, is er geen deelname, dan is $D_i = 0$.

Stel verder dat de behandeling plaatsvindt in periode k, dan kan het individueel behandelingseffect als volgt worden beschreven, ervan uitgaande dat t > k:

$$\Delta_{it}(X_i) = Y^1_{it} - Y^0_{it} = [g^1_t(X_i) - g^0_t(X_i)] + [U^1_{it} - U^0_{it}] \tag{15}$$

Ook hier kunnen individuele effecten uiteraard niet geschat worden (aangezien een individu nooit tegelijkertijd wel en niet kan hebben deelgenomen). We definiëren dan ook weer een gemiddeld groepseffect, dat voor een periode na behandeling, t > k, als volgt wordt gegeven:

$$\Delta_{ATT} = E(\Delta_{it} \mid X=X_i, D_i=1) \tag{16}$$

Als het selectieproces niet toevallig is, wat te verwachten valt in een niet-experimentele situatie, zal er een stochastische afhankelijkheid[5] ontstaan tussen (U^0,U^1) en de niet-geobserveerde V_i uit (14), of zal er een stochastische afhankelijkheid ontstaan tussen (U^0,U^1) en de geobserveerde Z_i uit (14). Beide situaties zullen aanleiding geven tot een correlatie tussen deelnamestatus D_i en de

4 In het vervolg van deze sectie wordt wederom nauw aangesloten bij de artikels van Caliendo en Hujer 2006, Blundell en Costa Dias 2002 , Smith 2000 en Heckman e.a. 1999.

5 Stochastische of statistische afhankelijkheid wijst op een afhankelijkheid tussen toevalsvariabelen waarbij de voorwaardelijke verdeling van de ene variabele wijzigt bij een wijziging van een van de andere variabelen. Een evident voorbeeld is de correlatie (een maat voor de *lineaire* samenhang), maar variabelen kunnen nog op tal van andere wijzen afhankelijk zijn van elkaar.

resttermen (U^0,U^1). De eerste situatie noemt men een selectie op basis van niet-geobserveerde kenmerken, de tweede situatie is dan een selectie op basis van geobserveerde kenmerken.

Deze mechanismen kunnen eenvoudig geïllustreerd worden in het geval waar wordt verondersteld dat de impact van het programma identiek is voor alle deelnemers. Als wordt verondersteld dat de laatste term uit (15) wegvalt, hetgeen het geval is als $U^0 = U^1$, is er sprake van een homogene impact voor iedereen met dezelfde X:

$$\Delta_t(X_i) = g^1_t(X_i) - g^0_t(X_i) \qquad (17)$$

Deze vergelijking stelt dat voor alle individuen met dezelfde X, de impact van het verhuizen van toestand '0' naar toestand '1' identiek is, hetgeen geldt zolang hun $U^0 = U^1$. Een nog verdere vereenvoudiging reduceert de impact tot een constante Δ_t. In dat geval winnen (of verliezen) alle deelnemers eenzelfde hoeveelheid als ze van '0' naar '1' gaan, ongeacht hun X. Aangezien dan voor iedereen geldt dat $Y^1 - Y^0$ = constante = Δ_t, en $U^0_i=U^1_i=U_i$, kan de situatie geresumeerd worden als volgt:

$$Y_{it} = g^0_t(X_i) + \Delta_t D_{it} + U_i \qquad (18)$$

wat, als men kiest om $g_0(X)$ voor te stellen met een lineair verband $X\beta$, geschreven kan worden als het bekende dummyvariabele regressiemodel:

$$Y_{it} = X_i\beta + \Delta_t D_{it} + U_i \qquad (19)$$

Deze wel erg vereenvoudigde kijk op de werkelijkheid (die in toegepast werk overigens algemeen wordt gebruikt) stelt dat de arbeidsmarktprestaties van een willekeurig individu i in een willekeurige periode t (nl. Y_{it}) bepaald worden door ten eerste het samenspel van de geobserveerde kenmerken van i (met name de lineaire predictor $X_i\beta$), ten tweede door de niet-geobserveerde elementen U_i, en ten derde, door de deelnamestatus: als $D_{it} = 1$, wordt bij de twee voorgaande termen een constante grootheid Δ_t bijgeteld, als $D_{it} = 0$, is uiteraard ook $\Delta_t D_{it}$ gelijk aan nul.

Zoals gezegd, kan het mechanisme van selectie op basis van niet-geobserveerde kenmerken eenvoudig worden geïllustreerd aan de hand van vergelijking (19). Stel dat de motivatie van kandidaten een erg belangrijke rol speelt bij de zelfselectie/de selectie door programmaverantwoordelijken. We mogen dan aannemen dat de behandelden gemiddeld gezien hoger zullen scoren op motivatie dan personen uit een vergelijkingsgroep van niet-behandelden. Dat betekent m.a.w. dat er een correlatie ontstaat tussen motivatie enerzijds, en de deelnamestatus D_{it} anderzijds. Stel verder dat de onderzoeker die het programma evalueert, geen informatie heeft over de motivatie van individuele personen. Motivatie zit in de selectievergelijking (14) dan bij de niet-geobserveerde component V_i (en hangt op die wijze samen met D_{it}). Is dit een probleem voor de meting van de impact van het programma middels vergelijking (19)? Neen, als de motivatie niet voorkomt in deze vergelijking. Als de motivatie echter wel de arbeidsmarktuitkomsten Y meebepaalt -een veronderstelling die niet al te ver gezocht lijkt- zit de motivatie ook aan de rechterzijde van vergelijking (19). En aangezien de motivatie niet-geobserveerd wordt door de onderzoeker, ons startpunt, impliceert één en ander dat haar effect zal worden weerspiegeld door U_i, het geheel van niet-geobserveerde componenten die mee de uitkomst bepalen.

Resumerend geeft dit dan dat de motivatie correleert met D_{it}, en dat de motivatie ook terugkomt bij U_i. Onvermijdelijk zal er dan ook bij vergelijking (19) sprake zijn van een correlatie tussen de regressor D_{it} en de restterm U_i. Een bekend resultaat, maar intuïtief zal het ook duidelijk zijn, is dat in zo een geval de coëfficiënt die het effect van deelname D_{it} op de uitkomst schat, met name Δ_t, onzuiver of vertekend zal worden geschat: deze coëfficiënt zal niet alleen het effect van D_{it}, maar

ook het effect van de motivatie op de uitkomst capteren. Dit is een typisch geval van selectievertekening. In de literatuur spreekt men ook wel eens over 'sample selection bias'. Het woord 'sample', d.i. steekproef, wijst op het feit dat het selectieprobleem fundamenteel voortvloeit uit het feit dat men met een verkeerde steekproef werkt, in het beschouwde probleem zou de selectievertekening van meet af aan kunnen vermeden zijn, mocht de vergelijkingsgroep bestaan hebben uit een steekproef van personen met een vergelijkbare motivatie als de steekproef van de deelnemers, quod non.

Het mechanisme van selectie op basis van geobserveerde verschillen is analoog. Als wederom de motivatie een rol speelt bij de selectie, maar nu voor iedereen een gekende waarde is, zit de motivatie in vergelijking (14) in de vector Z_i, en ontstaat een correlatie tussen de motivatie en D_{it}. Het grote verschil met de situatie van selectie op basis van niet-geobserveerde verschillen is dat de motivatie in vergelijking (19) kan worden opgenomen in de vector van geobserveerde verschillen X_i, waardoor er niet langer een correlatie optreedt tussen de regressor D_{it} en de restterm U_i, en de coëfficiënt Δ_t zuiver kan worden geschat (gesteld natuurlijk dat er geen andere niet-geobserveerde kenmerken zijn die zowel de deelnamekans als de uitkomst beïnvloeden).

Zoals gezegd, is de veronderstelling van homogene behandelingseffecten nogal simplistisch. Een uitbreiding naar heterogene behandelingseffecten is gemakkelijk te maken met de bouwstenen die reeds voorhanden zijn. Vertrekkende van de startvergelijking (2):

$$Y_{it} = D_{it}Y_{it}^{1} + (1 - D_{it})\, Y_{it}^{0} \qquad (2)$$

krijgen we, gebruik maken van (13):

$$Y_{it} = g^{0}{}_{t}(X_i) + U^{0}_{it} + \Delta_{it}(X_i)\, D_{it} \qquad (20)$$

Vergelijking (20) stelt dat in het algemeen de arbeidsmarktuitkomst gelijk is aan $Y^{0}_{it} = g^{0}{}_{t}(X_i) + U^{0}_{it}$, (dat is wat men bekomt als men niet deelneemt, cfr. (13)) en stelt vervolgens dat voor degenen die wel deelnemen en dus een D_{it} hebben die gelijk is aan 1, hier nog een bedrag $\Delta_{it}(X_i)$ moet worden bijgeteld. De notatie $\Delta(X_i)$ verwijst naar het feit dat het bedrag Δ afhangt van de kenmerken X, de subscripten it bij de Δ verwijzen bovendien naar het feit dat het bedrag $\Delta_{it}(X_i)$ voor ieder individu i op elk moment t uniek is. Dit laatste is duidelijk te veel van het goede: een model waarin ieder individu zijn persoonlijke impactquote heeft, is niet schatbaar.

Dit laatste wordt opgelost door het persoonsspecifieke element dat niet gerelateerd is aan de geobserveerde kenmerken X naar de niet-geobserveerde restterm te verplaatsen. Daartoe wordt $\Delta_{it}(X_i)$ opgesplitst in een geobserveerde en een niet geobserveerde component, waarbij deze laatste uiteraard de mate is waarin de U^1 (de U bij deelname) verschilt van de U^0 (de U bij niet-deelname):

$$\Delta_{it}(X_i) = \Delta_t(X_i) + (U^{1}_{it} - U^{0}_{it}) \qquad (21)$$

Als (21) dan vervolgens wordt ingeplugd in (20), krijgen we:

$$Y_{it} = g^{0}{}_{t}(X_i) + \Delta_t(X_i)\, D_{it} + [U^{0}_{it} + D_{it}\,(U^{1}_{it} - U^{0}_{it})] \qquad (22)$$

waarbij de notatie $\Delta_t(X_i)$ aangeeft dat de geschatte impact afhankelijk is van de waarde van X_i. Deze parameter kan worden gezien als de verwachte waarde ('het gemiddelde') van de $\Delta_{it}(X_i)$:

$$\Delta_t(X_i) = E[\Delta_{it}(X_i)] = g^{1}{}_{t}(X_i) - g^{0}{}_{t}(X_i) \qquad (23)$$

Vergelijking (22) is een al veel algemenere vergelijking dan (19). Nu wordt het vb. mogelijk om vast te stellen dat een bepaald programma weliswaar een gemiddeld negatieve impact heeft, maar dat desalniettemin deelgroepen met een bepaalde X er toch duidelijk baat bij hebben.

Tegelijkertijd is natuurlijk duidelijk dat het meten van heterogene behandelingseffecten ook hogere eisen stelt aan de benodigde data. Zo wijzen (Blundell and Costa Dias 2002) op het 'common support'-probleem:[6] als men vb. iets wil zeggen over het effect van het programma op een bepaalde doelgroep, is het wenselijk dat zowel in de behandelde groep als in de vergelijkingsgroep leden met doelgroepkenmerken aanwezig zijn.

3.2 Selectie op basis van geobserveerde verschillen

3.2.1 Matching

Bij het schatten van een counterfactual moet op basis van een geobserveerde populatie iets worden afgeleid over een andere, hypothetische of niet-geobserveerde populatie. Dit is een oefening die men uiteraard niet zonder meer mag ondernemen. Afhankelijk van de context zal men aan strengere of minder strenge eisen moeten voldoen opdat deze oefening kan gemaakt worden. Deze voorwaarden noemt men de identificerende veronderstellingen. Of nog, wanneer men een bepaalde schattingsmethode hanteert, gaat men er (impliciet of expliciet) vanuit dat voldaan is aan de bij die methode horende veronderstellingen. Als effectief aan die voorwaarde(n) is voldaan, is er sprake van identificatie, het resultaat van de methode is in dat geval identiek aan de counterfactual.

Bij de methode van matching wordt op één of andere manier voor iedere deelnemer een niet-deelnemer gezocht waarbij wordt gelet op het feit dat er een match, een overeenkomst is tussen beide wat betreft de geobserveerde kenmerken X. Als voldaan is aan de volgende voorwaarde, is het enige verschil dat overblijft tussen deelnemers- en vergelijkingsgroep de deelnamestatus. De identificerende veronderstelling is dat, voorwaardelijk op de verzameling van covariaten X, de uitkomst Y onafhankelijk is van deelname D, of nog[7]:.

$$Y^0, Y^1 \perp D \mid X \qquad (24)$$

Dit wil dus zeggen dat, als we rekening kunnen houden met de rol van alle relevante covariaten X, (waarbij relevant betekent dat de covariaat de uitkomst én de deelname mee determineert), we mogen stellen dat de verdeling van uitkomsten zonder behandeling gelijk is aan wat de verdeling van uitkomsten met behandeling zou geweest zijn bij afwezigheid van behandeling, en vice versa.[8] Of nog, als aan (24) is voldaan, volgt dat:

$$F(Y^0 \mid X, D=0) = F(Y^0 \mid X, D=1) \qquad (25)$$

$$\text{en } F(Y^1 \mid X, D=0) = F(Y^1 \mid X, D=1) \qquad (25')$$

---

6 De "support" betreft in de statistiek de verzameling van waarden waarvoor de dichtheidsfunctie niet-nul is, m.a.w. heeft dit betrekking op de waarden van een variabele waarvan de kans dat ze voorkomen, groter is dan nul.

7 Het symbool $\perp$ verwijst naar het feit dat het één loodrecht ("orthogonaal") op het ander staat, en dat er dus met andere woorden sprake is van onafhankelijkheid. Voorwaarde (24) staat in de literatuur bekend als CIA ("conditional independence assumption").

8 Kunnen rekening houden met alle covariaten die de deelname beïnvloeden, impliceert dat er sprake moet zijn van selectie op basis van geobserveerde verschillen, met niet-geobserveerde kenmerken kan men immers geen rekening houden, althans niet direct.

Uitdrukking (25) geeft weer dat de verdeling van de niet-programmauitkomsten Y^0 identiek zijn voor de vergelijkingsgroep en de deelnemersgroep, bij gegeven X. Of nog, neem twee personen met dezelfde X, dan zullen ze ook dezelfde Y^0 hebben, zelfs al zit de ene in de vergelijkingsgroep (D=0) en de andere in de deelnemersgroep (D=1). Uitdrukking (25') zegt hetzelfde m.b.t. Y^1.

De gemiddelden van deze verdelingen lossen dan het selectieprobleem op:

$$E(Y^0 \mid X,D=0) = E(Y^0 \mid X,D=1) = E(Y^0 \mid X) \quad (26)$$

$$\text{en } E(Y^1 \mid X,D=0) = E(Y^1 \mid X,D=1) = E(Y^1 \mid X) \quad (27)$$

Vergelijking (26) kan dan worden gebruikt om te schatten wat de deelnemers zouden hebben gepresteerd, mochten ze niet hebben deelgenomen, vergelijking (27) kan dan weer worden gehanteerd bij het schatten van wat niet-deelnemers zouden presteren als ze wel hadden deelgenomen. Dit laatste voegt een counterfactual toe aan de ATE en de ATT die vroeger al werden gespecifieerd, namelijk het gemiddeld effect van de behandeling op de niet-behandelden ('average effect of treatment on the untreated'):

$$\Delta_{ATU} = E(\Delta \mid D=0) = E(Y^1 \mid D=0) - E(Y^0 \mid D=0) \quad (28)$$

Wanneer de behandelingseffecten volledig homogeen zijn, en dus op geen enkele manier afhangen van individuele kenmerken (al dan niet geobserveerd), geldt uiteraard dat $\Delta_{ATU} = \Delta_{ATT} = \Delta_{ATE}$.

Een belangrijke randvoorwaarde opdat de methode van matching op de volledige groep van behandelden van toepassing zou zijn, is dat er voldoende overlapping moet zijn tussen het 'soort van mensen' die behandeld worden (lees: de configuraties van kenmerken X die voorkomen binnen de groep van behandelden), en 'het soort van mensen' die beschikbaar zijn voor de vergelijkingsgroep. Dit is het reeds eerder vermelde 'common support'-probleem. Als bij wijze van voorbeeld alle laaggeschoolde werkloze jongeren een behandeling krijgen, zijn er geen lager geschoolde werkloze jongeren beschikbaar om mee te matchen. Er wordt dan ook opgelegd dat:

$$0 < Pr(D=1 \mid X) < 1, \text{ voor alle X} \quad (29)$$

Het mag niet zo zijn dat de kans op deelname, Pr(D=1), voor een gegeven configuratie van kenmerken X gelijk is aan 1, want dan zal men voor deze personen geen vergelijkingsgroep van niet-deelnemers vinden. Dat deze kans bovendien groter moet zijn dan nul, moet enkel worden opgelegd als men ATU wil schatten.[9]

In de praktijk kan het natuurlijk voorvallen dat de support van X bij de behandelden niet volledig die van de niet-behandelden overlapt. Er zijn dan deelnemers waarvoor men geen vergelijkbare niet-deelnemer vindt. In dat geval moet de matching beperkt worden tot die groep waarvoor er wel een gemeenschappelijke support is.

In principe is het mogelijk om ATT te schatten op een punt X=x, waar x een bepaalde realisatie is van X. Bij uitbreiding kan dan ook gesteld worden dat matching onder zekere voorwaarden ruimte laat voor individuele effectheterogeniteit (cfr. 22).

Aangezien voor een goed resultaat informatie nodig is over alle relevante covariaten die de uitkomst en de deelname mee determineren, zal er bij matching steeds behoefte zijn aan rijke gege-

9 Als alleen ATT en niet ATU moet worden geïdentificeerd, kan men bovendien ook in veronderstelling 24 de Y^1 laten vallen, en volstaat $Y0 \perp D \mid X$

vens. Dit kan op zijn beurt leiden tot een zekere 'curse of dimensionality', hoe groter het aantal variabelen in de vector X, hoe groter de kans dat er configuraties van kenmerken zullen aanwezig zijn in de behandelde groep die men niet terugvindt bij de niet-behandelden. Sinds 1983 is er evenwel een oplossing voorhanden voor dit probleem. (Rosenbaum and Rubin 1983) tonen immers aan dat conditioneren op de vector X equivalent is aan het conditioneren op de scalair P(X), waarbij dit laatste de kans op deelname, gegeven de kenmerken X, weergeeft. De kans P(X) kan eenvoudig worden geschat door een waarschijnlijkheidsmodel te schatten waarin de deelnamestatus wordt gerelateerd aan de kenmerken X. Na schatting kan men dan voor alle deelnemers en niet-deelnemers de kans op deelname voorspellen. Voor iedere deelnemer wordt dan een niet-deelnemer gezocht met een vergelijkbare voorspelde deelnamekans, ook propensity score genoemd. De benadering heet dan ook propensity score matching (PSM).

Bij het schatten van ATT (waarbij enkel de counterfactual $E(Y^0 \mid D=1)$ nodig is) wordt het volgende verondersteld:

$$Y^0 \perp D \mid P(X) \tag{30}$$

$$Pr(D=1 \mid X) < 1, \text{ voor alle } X \tag{31}$$

zodat (als het gemiddelde is gedefinieerd) het evaluatieprobleem aldus opgelost wordt:

$$E(Y^0 \mid P(X),D=1) = E(Y^0 \mid P(X),D=0) = E(Y^0 \mid P(X)) \tag{32}$$

Smith (Smith 2000) vermeldt een drietal aandachtspunten bij het gebruik van PSM. Ten eerste, voor de concrete uitvoering van de matching bestaan er diverse benaderingen (zie (Smith 2007, Smith and Todd 2005)). De keuze van de methode kan tot verschillende resultaten leiden bij kleine steekproeven. Ten tweede moet men de nodige aandacht besteden aan welke kenmerken X worden gebruikt bij het schatten van P(X), aangezien de schattingsresultaten (32) erg gevoelig kunnen zijn voor deze keuze. Ten derde moet men geschatte standaardfouten corrigeren, omdat niet alleen het schatten van de propensity scores, maar ook het matching proces zelf variatie toevoegen die de gebruikelijke steekproefvariatie overschrijden. Bij de correctie kan men gebruik maken van bootstrappingprocedures.

Een erg interessante uitbreiding van het evaluatiekader op basis van matching is te vinden bij Lechner (Lechner 2002). In realiteit lopen dikwijls verschillende arbeidsmarktprogramma's naast elkaar, en de voorgestelde uitbreiding evalueert die 'multiple programmes' binnen één kader. Essentieel krijgt men dan i.p.v. Y^0 en Y^1 een vector $(Y^0,Y^1,...,Y^M)$ met (M+1) elkaar wederzijds uitsluitende toestanden, want een individu kan op een bepaald moment slechts in één toestand zijn. Voor iedere toestand zijn er dan M counterfactuals. Als de deelname aan een specifieke behandeling m wordt weergegeven door de indicator S = (0,1,...,M), kunnen paarsgewijze gemiddelde behandelingseffecten worden gedefinieerd van de behandelingen m en n voor deelnemers aan behandeling m:

$$\Delta_{ATT}^{ml} = E(Y^m - Y^n \mid S=m) = E(Y^m \mid S=m) - E(Y^n \mid S=m) \tag{33}$$

Als deelnemers aan behandelingen m en n verschillen in de verdeling van hun X, en als die X invloed heeft op de effecten, zijn de verschillende ATT's niet symmetrisch, d.w.z. $\Delta_{ATT}^{ml} \neq -\Delta_{ATT}^{lm}$.

De identificerende veronderstelling is:

$$Y^0,Y^1,...,Y^M \perp S \mid X \tag{34}$$

De onderzoeker moet dus alle kenmerken observeren die gemeenschappelijk de verschillende uitkomsten en de verschillende selectieprocessen voor alle behandelingen bepalen. Aanvullend is er ook een voorwaarde m.b.t. de gemeenschappelijke support:

$$0 < Pr(S=m \mid X=x) \text{ , voor alle } m= 0,...,M \text{ en voor alle } x \in X \qquad (35)$$

Ten einde de paarsgewijze vergelijking mogelijk te maken, moeten er m.a.w. voor alle waarden van x die voorkomen bij de behandelden met een behandeling m, vergelijkingsobservaties beschikbaar zijn. Ook hier is een aangepaste versie van PSM toepasbaar.

Deze benadering is a fortiori aantrekkelijk in een activerende arbeidsmarkt waar een groot deel van de werklozenpopulatie één of andere behandeling krijgt, zodanig dat het moeilijk wordt om nog vergelijkingsgroepen samen te stellen van onbehandelden (of waar het feit dat iemand met een gegeven werkloosheidsduur nog geen behandeling onderging juist suggereert dat deze persoon op een of andere dimensie afwijkt van de gemiddelde werkloze, en dus juist niet gewenst is in een vergelijkingsgroep). Zie ook Aho (Aho 2005), die oplossingen zoekt voor het feit dat er in Finland niet alleen sprake is van bijna universele participatie, maar dat bovendien vele deelnemers reeds in het verleden aan andere programma's deelnamen.

Een andere interessante uitbreiding van het matchingkader, die expliciet rekening houdt met het feit dat individuen inderdaad over de tijd heen kunnen deelnemen aan verschillende programma's, is eveneens gemaakt door Lechner (Lechner 2004) met het voorstellen van sequentiële matching schatters.

Controle en sanctionering van (de zoekinspanning van) werklozen

Vanaf juli 2004 werd in België het plan tot activering van het zoekgedrag ingevoerd. Dit hield in dat de RVA systematisch startte met een opvolging van de zoekinspanningen van de langdurig werklozen. Van langdurige werkloosheid is bij min-25-jarigen sprake vanaf een werkloosheidsduur van 7 maanden, en bij de plus-25-jarigen vanaf 13 maanden. Alle werklozen die deze grens overschrijden krijgen een aankondigingsbrief waarin staat vermeld wat hun verplichtingen zijn, en waarin ook wordt aangekondigd dat ze circa 8 maanden later, als ze dan nog in de werkloosheid zitten, zullen worden uitgenodigd voor een gesprek waarin zal worden nagegaan of ze inderdaad voldoende inspanningen hebben gedaan om werk te vinden.

Na 8 maanden volgt dan een eerste gesprek. Als de RVA op basis van dit gesprek oordeelt dat voldoende zoekinspanningen zijn aangetoond (vb. via sollicitatiebrieven etc.), hoeft de werkloze gedurende de volgende 16 maanden niet terug te komen. Als men een negatief oordeel krijgt, wordt gezamenlijk een actieplan opgesteld met een opsomming van te ondernemen stappen. Als men niet komt opdagen op dit eerste gesprek, wordt de werkloosheidsuitkering opgeschort tot men wel komt opdagen.

Degenen met een actieplan worden 4 maanden later uitgenodigd voor een tweede gesprek. Als het resultaat positief is, wordt men gedurende 12 maanden met rust gelaten. Als het resultaat negatief is, volgt een tijdelijke sanctie (schorsing van de betaling van de uitkering) en een nieuw actieplan. Ook hier zal niet komen opdagen leiden tot een sanctie. Een derde gesprek vindt dan vervolgens nog eens 4 maanden later plaats. Als het resultaat positief is, wordt men gedurende 12 maanden met rust gelaten. Als het resultaat negatief is, volgt een schorsing voor onbepaalde duur.

In sectie 3.3.4 bespreken we een onderzoek dat het effect probeert te schatten van de bovenstaande procedure. Hier wordt ingefocust op een onderzoek naar de gevolgen van de sanctionering. In de periode voor de invoering van het plan tot activering van het zoekgedrag werden er ook tijdelijke en definitieve sancties gegeven. Eén van de manieren waarop men gesanctioneerd kon worden, was via het zogenaamde art. 80 van de RVA-regelgeving rond werkloosheidsuitkeringen, dat in voege was sinds juli 1992. Het hield in dat men de werkloosheidsuitkering kon verliezen als de werkloosheidsduur 1,5 keer (vanaf 1996: 2 keer) hoger was dan het gemiddelde, waarbij werd rekening gehouden met de regio, de leeftijd en het geslacht. Gezins-

hoofden en alleenstaanden vielen evenwel buiten deze regeling, zodanig dat in de praktijk vooral vrouwelijke werklozen werden geviseerd.

Met de invoering van het plan tot activering van het zoekgedrag werd dit art.80 afgeschaft. Daarmee verdween een bron van discriminatie, maar ontstond tegelijkertijd bij sommigen de vrees dat het totaal aantal schorsingen zou toenemen, omdat nu ook gezinshoofden en alleenstaanden in aanmerking kwamen. Daarbij werd bovendien gevreesd dat deze laatsten zich na schorsing onmiddellijk tot het OCMW zouden richten voor het krijgen van een uitkering, zodanig dat bij wijze van spreken de factuur door de OCMW's zou worden betaald.

Begin 2005 werd een onderzoeksopdracht uitgeschreven met de vraag deze problematiek te onderzoeken. Aangezien op dat moment het plan tot activering van het zoekgedrag pas in een opstartfase zat (en het aantal schorsingen nog veeleer gering was), werd besloten om een nulmeting te doen: om te kunnen beoordelen of het plan tot een wijziging leidt op het vlak van de schorsingspraktijk (omvang, samenstelling, mate van doorstroom naar het OCMW,...) was het alleszins nodig te weten wat de situatie was voorafgaand aan de invoering van het actieplan.

In het sanctiebestand van de RVA werden over de loop van 5 kwartalen (2002.IV t.e.m. 2003.IV) 25 219 personen gevonden die vanuit een werkloosheidsstatuut een sanctie opliepen. Een en ander werd gekoppeld aan het datawarehouse van de KSZ, zodanig dat kon worden nagegaan in welke arbeidsmarktstatuten de gesanctioneerden zich bevonden in het kwartaal/de kwartalen volgend op het kwartaal van schorsing (Heylen and Bollens 2006).

Een vergelijkingsgroep van werklozen die geen schorsing opliepen, werd geselecteerd uit het werklozenbestand van de RVA op basis van Propensity Score Matching. Het geschatte waarschijnlijkheidsmodel hield daarbij rekening met de kenmerken geslacht, leeftijd, provincie, gezinssituatie (alleenstaand of niet, kinderen ten laste of niet, etc.), en de werkloosheidduur op het moment van de schorsing.

Vooreerst werd nagegaan of voor de diverse personen in de behandelde groep en in de vergelijkingsgroep betalingen door een OCMW konden worden teruggevonden. In het kwartaal van de sanctie zelf, was er voor 9,8% van de geschorsten sprake van een betaling door een OCMW, bij de vergelijkingsgroep bedroeg het overeenkomstige percentage 1,4%. In het kwartaal volgend op het kwartaal van schorsing blijft deze verhouding met 10,4% versus 1,5% vergelijkbaar. In het tweede kwartaal na het kwartaal van schorsing zakt het aandeel binnen de geschorste groep naar 5,7%, het aandeel binnen de vergelijkingsgroep handhaaft zich, zoals te verwachten, op 1,4%. Het aantal personen met een vergoeding van een OCMW stijgt duidelijk naarmate de duur van de schorsing langer is.

Daarnaast is het ook interessant hoe de doorstroom is naar respectievelijk werk en een toestand buiten de arbeidsmarkt. In het kwartaal van de sanctie hebben 45,5% van de gesanctioneerden (minstens gedurende enige tijd) gewerkt, in de twee volgende kwartalen is dit het geval voor 43,6% en 44,4%. De cijfers voor de vergelijkingsgroep zijn met 42,5%, 41,9% en 42,0% van een vergelijkbare orde.

Een groter verschil is er echter wat betreft personen die niet kunnen worden teruggevonden in een andere databank van het datawarehouse, en die dus vermoedelijk de arbeidsmarkt verlaten. Voor de gesanctioneerden is dit in het kwartaal na het kwartaal van de sanctie het geval voor 22,9% (versus 6,9% bij de vergelijkingsgroep), in het tweede kwartaal na het kwartaal van sanctie zijn de overeenkomstige cijfers 20,9% versus 7,3%.

3.2.2 Lineaire Regressie

In het geval van selectie op geobserveerde variabelen, biedt ook een gewone regressieaanpak een oplossing voor het evaluatieprobleem. Als de uitkomst wordt geregresseerd op de X en een deelname-indicator D, bekomt men eveneens een schatting van de impact die conditioneert op X, zoals bij matching. Wat is dan eigenlijk het verschil met een matching-benadering? Een eerste verschil is dat de regressie parametrisch is, en er o.a. functionele vorm veronderstellingen worden gemaakt

('een lineair verband'), terwijl de matchingbenadering niet-parametrisch is, en daar dus niet dergelijke veronderstellingen moeten worden gemaakt (zoals ook een experimentele benadering geen parametrische veronderstellingen nodig heeft).

Als de potentiële uitkomsten als volgt zijn te schrijven (voor de eenvoud met weglating van subscripten i en t): $Y^1 = X\beta_1 + U^1$ en $Y^0 = X\beta_0 + U^0$, dan is ATT bij regressie gelijk aan:

$$\Delta_{ATT}^{REG} = E(Y^1-Y^0 \mid X,D=1) =$$

$$X(\beta_1- \beta_0) - E(U^1-U^0 \mid X, D=1) \quad (36)$$

Selectievertekening kan in het algemeen optreden als $F(Y^0 \mid X,D=0)$ wordt gebruikt als benadering voor $F(Y^0 \mid X,D=1)$. Bij het schatten van $E(Y^1-Y^0 \mid X,D=1)$ wordt de omvang van deze selectievertekening gegeven door:

$$B(X) = E(Y^0 \mid X,D=1) - E(Y^0 \mid X,D=0) \quad (37)$$

Bij matching bekwamen we:

$$E(Y^0 \mid X,D=0) = E(Y^0 \mid X,D=1) = E(Y^0 \mid X) \quad (26)$$

Voor ieder 'stratum' in X wordt de vertekening B(X) = 0. Dit betekent niet dat er geen selectievertekening is, het hoeft met name niet zo te zijn dat $E(U^0 \mid X,D=1) = 0$. Wat dan wel gebeurt, is dat matching de vertekening balanceert binnen een stratum:

$$E(U^0 \mid X,D=1) = E(U^0 \mid X,D=0) = E(U^0 \mid X) \quad (38)$$

Matching zal met andere woorden zorgen dat het gemiddelde van restterm voor de behandelde groep binnen een bepaald stratum, gelijk is aan het overeenkomstig gemiddelde voor de vergelijkingsgroep. In een regressiecontext daarentegen, mag er geen afhankelijkheid aanwezig zijn tussen (U^0,U^1) en X (zie ook vroeger), zodanig dat hier een veel strengere eis moet worden opgelegd:

$$E(U^1 \mid X,D=1) = E(U^0 \mid X,D=0) = 0 \quad (39)$$

In de praktijk zal dit veelal betekenen dat zelfs bij het gebruik van dezelfde X er bij het gebruik van matching minder vertekening zal optreden dan bij het gebruik van een regressiebenadering.

Een tweede belangrijk verschil tussen de beide benaderingen is dat men bij matching automatisch op gemeenschappelijke supportproblemen zal botsen als ze zich stellen, en men er dan ook bewust mee zal omgaan. In een regressiebenadering zal de opgelegde lineaire structuur er voor zorgen dat zelfs in afwezigheid van geschikte vergelijkingseenheden toch schattingen worden gemaakt, zonder dan men zich altijd bewust is van het support probleem. Stel dat er in de behandelde groep laaggeschoolde jongeren zitten, en in de vergelijkingsgroep niet. In de vergelijkingsgroep zijn wel (hogergeschoolde) jongeren, en (oudere) laaggeschoolden. In een matchingbenadering (waarbij leeftijd en scholingsniveau deel uitmaken van X) zal men dan onmiddellijk vaststellen dat er geen match kan worden gevonden. In een regressiebenadering zullen resultaten over de (weliswaar hogergeschoolde) jongeren en over de (weliswaar oudere) laaggeschoolden geëxtrapoleerd worden naar laaggeschoolde jongeren. Of dit een voordeel of een nadeel is, hangt af van de context.

Een derde verschil betreft de mogelijkheid om rekening te houden met effect-heterogeniteit. De matching benadering legt geen restricties op het individuele causaal effect en bijgevolg laat het een willekeurige effect-heterogeniteit toe (Lechner 2002).

3.3 Selectie op basis van niet-geobserveerde verschillen

3.3.1 Instrumentvariabele

Een instrumentvariabele (IV) is een variabele die weliswaar de deelname mee determineert, maar die anderzijds geen invloed heeft op de uitkomsten. Als zo een instrument kan gevonden worden, kan het evaluatieprobleem worden opgelost: de opname van een dergelijke variabele in een uitkomstenvergelijking stelt geen probleem (want het instrument is per definitie niet gecorreleerd met de restterm U_{it}), en zal toch toelaten om causale effecten te identificeren, aangezien het instrument, wederom per definitie, correleert met de deelnamestatus.

De vraag is natuurlijk of het mogelijk is om een dergelijke, toch wat speciale variabele te vinden. Er is ooit voorgesteld om de afstand tussen de woonplaats en het opleidingscentrum te selecteren als instrument voor de deelnamestatus aan opleiding door werkzoekenden. Aan de hand van de volgende drie voorwaarden kan worden nagegaan of zo een voorstel voldoet aan de vereisten die moeten worden opgelegd aan een instrument. Gegeven de deelnamebeslissingsregel (14),

$$IN_i = f(Z_i) + V_i \qquad (14)$$

moet een geschikt instrument Z* voldoen aan:

(a) Z* moet samenhangen met de deelname aan het programma. Dit betekent dat Z* in (een lineaire versie van) de beslissingsregel (14) een coëfficiënt moet hebben die verschilt van nul.

(b) Z* wordt niet volledig gedetermineerd door X.

(c) Z* correleert niet met (U^1,V) en (U^0,V), gegeven X.

Voor een binair instrument (ééntje met slechts twee waarden, Z*=0 of Z*=1), wordt de IV-schatter gegeven door:

$$\Delta^{IV} = \frac{[E(Y \mid X,Z^*=1) - E(Y \mid X,Z^*=0)]}{[Pr(D=1 \mid X,Z^*=1) - Pr(D=1 \mid X,Z^*=0)]} \qquad (40)$$

De teller geeft het verschil tussen de gemiddelde uitkomsten van 'ingeinstrumenteerden' en 'uitgeinstrumenteerden'. De correlatie tussen het instrument en de deelnamestatus is niet perfect, dit wordt dan gecorrigeerd in de noemer.

Z* mag de uitkomsten enkel en alleen beïnvloeden langs deelnamestatus D. Bijgevolg stelt voorwaarde (c) dat Z* geen effect mag hebben op de uitkomsten langs de niet-geobserveerde component. Deze laatste voorwaarde impliceert dat het behandelingseffect homogeen is (m.b.t. niet-geobserveerde verschillen). Immers, zelfs als Z* niet correleert met U_{it}, dan zal Z* per definitie wel correleren met de restterm in het geval van een heterogeen behandelingseffect, dat, zie (22), gelijk is aan:

$[U^0_{it} + D_{it} (U^1_{it} - U^0_{it})]$

Immers, Z* correleert per definitie met D_i.

Toegepast op het voorbeeld van de afstand tussen woonplaats en opleidingscentrum, is voorwaarde (a) wellicht niet altijd onrealistisch, gesteld dat de kans op deelname daalt naarmate de afstand toeneemt. Voorwaarde (c) legt de weinig realistische hypothese van homogene effecten op. Als men toch heterogene effecten wil binnen dit kader, moeten een aantal andere erg onrealistische veronderstellingen worden gemaakt. Zoals Heckman e.a 1999 stellen, lijkt het logisch om te veronderstellen dat binnen de groep van potentiële deelnemers die ver wonen van het opleidingscentrum, degenen die inschatten dat ze veel te winnen hebben bij deelname, een grotere kans op deelname hebben dan degenen die minder baat bij deelname zien. M.a.w., de niet geobserveerde, door de potentiële deelnemer ingeschatte winst van deelname, met name U^1_{it}-U^0_{it} , is dan groter voor de eerste dan voor de tweede groep. Potentiële deelnemers die dichterbij wonen, zullen al deelnemen bij een lagere ingeschatte winst. Binnen de groep van deelnemers ontstaat er dan een correlatie tussen de afstand en de niet-geobserveerde component m.b.t. de winst van deelname, waardoor afstand niet langer geschikt is als instrument. Opdat het toch een aanvaardbaar instrument zou zijn moet men opleggen dat, ofwel, potentiële deelnemers niet op voorhand kunnen inschatten of ze veel of weinig baat hebben bij de deelname, ofwel, als ze dat wel kunnen, deze informatie alvast geen invloed heeft op de deelnamebeslissing. Geen van beide veronderstellingen is erg realistisch.

Een instrument is 'sterk' als men, via een wijziging in het instrument Z_i, alle individuen i er toe kan bewegen om hun D_i naar een bepaalde waarde te wijzigen. Stel dat zowel D als Z binaire variabelen zijn (vb. D= al dan niet deelname aan opleiding, Z = beleidskeuze a of b, vb. geen of wel opleidingscheques uitkeren, waarbij we (toegegeven, wat kunstmatig) veronderstellen dat op toevallige wijze sommigen hiervoor in aanmerking komen en anderen niet). Z_i is dan een sterk instrument als het in staat is alle individuen hun D te laten wijzigen van 0 naar 1 (of omgekeerd) als Z wordt gewijzigd van 0 naar 1. In de praktijk zal zo een sterk instrument zelden voorkomen, en zal het instrument met name slechts betrekking hebben op een deelpopulatie: voor het gegeven voorbeeld, als de beleidskeuze wijzigt van a naar b, gaan sommigen die onder a niet deelnamen aan de opleiding, wel beslissen tot deelname. In de literatuur noemt men degenen die hun gedrag laten beïnvloeden door het instrument 'compliers', de volgzamen. Deze term komt van oorsprong uit de medische literatuur, waar 'compliers' in een experiment degenen zijn die het in een experiment toegediende geneesmiddel ook effectief innemen, en dat anders niet zouden innemen.

Als er sprake is van een homogeen behandelingseffect, kan een IV-benadering ATE of ATT identificeren, immers, als men het homogeen effect kent voor een deelgroep, kent men het voor iedereen. In een wereld met heterogene behandelingseffecten, is dit niet langer waar. Dit is net te verklaren door het feit dat slechts een deelverzameling van de populatie wordt beïnvloed door het instrument: na de invoering van de opleidingscheque, zullen sommigen die niet voor het gebruik van die cheque in aanmerking komen toch opleiding volgen, en zal binnen de groep van personen die wel in aanmerking komt, zeker niet iedereen er gebruik van maken. Anderzijds is er een groep van personen die opleiding gaan volgen, en dat niet zouden gedaan hebben in afwezigheid van de cheque. Het opleidingscheque-instrument zegt niets over de effecten van opleiding van personen van wie het opleidingsgedrag niet werd beïnvloed door de invoering van de cheque. Het zegt echter wel iets over de 'compliers'.

In deze context wordt gesproken over een lokaal gemiddeld behandelingseffect (LATE, Local average treatment effect) (Imbens and Angrist J. 1994). Hiervan is sprake als de relatie tussen D en Y slechts kan worden gereconstrueerd voor de deelpopulatie die reageert op wijzigingen in het instrument. LATE geeft het gemiddeld effect op de uitkomst Y van een wijziging in D, voor de deel-

populatie van personen waarvan de waarde van D zou wijzigen bij een exogene wijziging van het instrument Z, m.a.w. voor de 'compliers'. Een en ander heeft een betekenis voor instrumenten die een indicator zijn voor verschillende alternatieve beleidskeuzes, of voor verschillende intensiteiten van een bepaalde beleidskeuze. Voor instrumenten die verwijzen naar persoonlijke of buurtkenmerken (vb. de afstand woonplaats -opleidingscentrum), is het veel minder duidelijk wat de betekenis van LATE zou kunnen zijn (Heckman e.a. 1999). In het afstandsvoorbeeld zal LATE het effect van de deelname geven voor die groep van personen die werden aangezet tot deelname door een wijziging in de pendelafstand.

Verder valt nog op te merken dat voor LATE naast de voorgaande voorwaarden nog een bijkomende voorwaarde moet worden opgelegd, met name monotoniciteit. Ook al heeft het instrument geen effect op sommige personen, voor degenen die wel beïnvloed worden, moet het effect wel voor iedereen in dezelfde richting gaan. Als het instrument wijzigt van a naar b, mogen er m.a.w. binnen de groep van beïnvloede personen ofwel alleen maar 'compliers', ofwel alleen maar 'defiers' zijn. De invoering van de cheque mag alleen betekenen dat de kans op opleiding toeneemt.

3.3.2 Natuurlijke experimenten

Heel de benadering van het zogenaamde natuurlijke experiment is nauw verbonden aan de voorgaande benadering. De uitgangsidee is dat de natuur zelf soms op een volstrekt toevallige wijze sommigen een behandeling geeft en anderen uitsluit van die behandeling. Gezien dit toevallig karakter van de toewijzing wordt m.a.w. het natuurlijk equivalent van een toevalstoewijzing uit een sociaal experiment verkregen, en ontstaat een natuurlijk experiment. Het toevallig karakter zorgt ervoor dat deze natuurlijke gebeurtenissen als instrumentvariabele kunnen worden gehanteerd. Voorbeelden van toevallige gebeurtenissen die voortvloeien uit biologische of klimatologische mechanismen en die reeds werden gebruikt als instrument zijn de geboorte van tweelingen, eeneiige tweelingen, de geboortedatum, het geslacht, en bepaalde gebeurtenissen op het vlak van het weer (Rosenzweig and Wolpin 2000).

Militaire dienst en de latere verdiensten

Een bekend voorbeeld dat mee aan de wieg stond van het enthousiasme voor de benadering van het natuurlijk experiment, is een bekende paper van Angrist (Angrist J. 1990): "Lifetime earnings and the Vietnam era draft lottery: evidence from social security administrative records". In deze studie wordt het feit geëxploiteerd dat het al dan niet onder de wapens geroepen worden in het VS ten tijde van Vietnam gebaseerd was op een loterij, zodanig dat er sprake was van een toevalstoewijzing naar de 'behandeling' (i.e. Militaire dienst). De auteur gebruikt dan ook deze loterij als instrument om dan vervolgens het effect van militaire dienst te schatten op wat men later verdient. De steekproef bestond uit mannen geboren van 1950 tot 1953. De loterij van 1970 had betrekking op de cohorte uit 1950, de loterij uit 1971 had betrekking op de cohorte van 1951, enz. Iedere man kreeg een loterijnummer van 1 tot 365, op basis van toevalstrekkingen uit geboortedatums. Al wie uitkwam op een nummer beneden een bepaalde grens (in 1972 lag die grens vb. op 95), kwam in aanmerking voor de militaire dienst. Voor deze laatste groep volgden dan een aantal testen om uit te maken welke deelverzameling van de in aanmerking komende mannen ook effectief moesten dienen (wat bleek uit te komen op 15%). Angrist regresseerde vervolgens de jaarinkomsten na twintig jaar (respectievelijk voor 1981 tot 1984) op een constante en een indicatorvariabele die aangeeft of men al dan niet veteraan is. Daarbij wordt het feit of men op basis van de loterij in aanmerking kwam voor militaire dienst gehanteerd als instrument voor de veteranenstatus. Laat Z*=1 voor degenen die in aanmerking kwamen (degenen die een laag nummer trokken) en laat Z*=0 voor degenen die niet in aanmerking kwamen op basis van hun nummer. Aangezien dit een binair instrument is, geldt de formule (40) uit de vorige sectie:

$$\Delta^{IV} = [E(Y \mid Z^*=1) - E(Y \mid Z^*=0)] / [Pr(D=1 \mid Z^*=1) - Pr(D=1 \mid Z^*=0)]$$

E(Y) is het gemiddelde inkomen voor de respectievelijke groepen, en D=1 is uiteraard de veteranenstatus. Op te merken valt dat $Pr(D=1 \mid Z^*=1)$ zoals reeds gezegd kleiner is dan 100% omdat van de in aanmerking komenden na testen een groot aantal wegviel, en ook dat $Pr(D=1 \mid Z^*=0)$ groter is dan nul omdat de niet in aanmerking komenden nog altijd op vrijwillige basis konden in dienst treden (naar verluidt waren vb. de Vietnam-veteranen overwegend vrijwilligers).

De schattingen impliceren dat militaire dienst het jaarinkomen voor blanken verminderde met 1 500 tot 3 000$, wat ongeveer een reductie van 15% is. Voor zwarten werd geen effect gevonden.

Op basis van dit 'experiment' kon dus worden besloten dat militaire dienst een duidelijk en autonoom negatief effect had op de latere verdiensten. Zonder het gebruik van het instrument, zou men bij het regresseren van inkomen op veteraanstatuut nooit zeker kunnen zijn over het feit dat dit effect autonoom uitgaat van de militaire dienst, dan zou altijd het vermoeden bestaan dat mensen bepaalde niet geobserveerde kenmerken hebben die maken dat ze kiezen voor militaire dienst en die ook maken dat hun verdienpotentieel lager is ('veteranen hadden al een lagere U om mee te beginnen').

Ondertussen zijn er ook heel wat natuurlijke experimenten gemeld waar de exogene variatie niet wordt veroorzaakt door de natuur, maar door andere oorzaken, zoals wijzigingen in de wet- of regelgeving. Een bekend voorbeeld is een verhoging van de minimumlonen in één staat van de VS, waarbij het effect van die wijziging wordt bestudeerd door de werkgelegenheid in de fastfood-restaurants in die staat en in een naburige staat waar de minimumlonen niet wijzigden, te bestuderen (Card and Krueger 1997). Heel dit opzet (naburige staten, beperken tot vergelijkbare bedrijven, etc.) is er natuurlijk op gericht om zoveel mogelijk factoren onder controle te houden. Dit kan overigens niet verhelen dat de onderzoeker bij natuurlijke experimenten typisch minder vat heeft op de experimentele condities dan wat het geval is bij een echt sociaal experiment. Om het onderscheid te maken tussen toevallige uitkomsten die voortvloeien uit natuurlijke mechanismen, en exogene variatie die voortvloeit uit andere mechanismen (wetswijziging e.d.), spreekt men in de literatuur ondertussen in het eerste geval over natuurlijke 'natuurlijke experimenten'. Het label 'natuurlijk experiment' geraakt in het taalgebruik ondertussen een beetje uitgehold, en slaat meer en meer op elke beleidsinterventie die twee groepen op een verschillende manier beïnvloedt, los van de vraag of die twee groepen initieel op elkaar leken of niet. Keane geeft als voorbeelden het gebruik van kleine bedrijven als controle voor grote bedrijven bij het schatten van het effect van wijzigingen in kredietwetten die enkel betrekking hebben op grote bedrijven, het gebruik van mannen als controle voor vrouwen bij het schatten van het effect van gewijzigde regels in de bijstand (wijziging die alleen betrekking had op vrouwen), het gebruik van mensen jonger dan 65 als controle voor mensen boven de 65 om het effect van de ziekteverzekering op de sterftecijfers te schatten. Hij besluit dat dit geen natuurlijke experimenten zijn, maar 'Un-Natural Non-Experiments' of, iets vriendelijker, 'Low-grade Quasi-Experiments' (Keane 2006).

Sociale experimenten worden overigens meestal bewust geïnitieerd, typisch met betrokkenheid van de onderzoeker. Natuurlijk experimenten worden normaal gezien niet bewust opgezet. Dit feit vormt alvast een belangrijke beperking. Het betekent dat de evaluatie van een bepaalde behandeling maar via deze methode kan gebeuren als toevallig een element van exogene variatie kan worden gevonden dat deelname beïnvloedt zonder de uitkomst te beïnvloeden. En ook dan zal de aard van de onderzoeksvragen die kunnen beantwoord worden vooral bepaald worden door de concrete omstandigheden, en minder door dat wat de onderzoeker eigenlijk zou willen weten.

In de praktijk kan het nog erger. Gedurende de laatste 20 jaren was deze benadering zo populair dat men soms het gevoel krijgt dat sommige onderzoeken niet werden uitgevoerd omdat de auteur het onderzoeksonderwerp intrinsiek belangrijk of interessant vond, maar wel omdat voor dat onderzoeksonderwerp een instrument was gevonden dat tot dan toe nog niet was opgedoken in de literatuur.

Men kan tegenwerpen dat naast de 'low grade quasi-experiments' toch ook de echte, natuurlijke natuurlijke experimenten staan, zoals het onderzoek naar het effect van de militaire dienst. Een dergelijk design heeft toch wel een grote validiteit?

Het blijkt dat ook hier methodologische bezwaren kunnen worden gemaakt. Een eerste opmerking is een variant op het bezwaar dat reeds werd geopperd bij het gebruiken van de afstand tot het opleidingscentrum als instrument. Stel dat het effect van militaire dienst op het latere inkomen heterogeen is (het tegendeel veronderstellen, zou nogal simplistisch zijn). Wel, zegt Heckman, dan is er een probleem ten gevolge van het feit dan men ook op vrijwillige basis kon beslissen om de militaire dienst te vervullen (Heckman 1997). Wanneer mensen ten dele anticiperen op de potentiële winst van militaire dienst (U^1-U^0), of als ze hun beslissing baseren op kenmerken die correleren met de niet geobserveerde component (U^1-U^0), dan is het veeleer waarschijnlijk dat personen met een hoog loterijnummer (die dus vrijgesteld waren), die toch de militaire dienst vervullen (en daar met andere woorden vrijwillig voor kozen) een hoge waarde hebben voor (U^1-U^0). In dat geval is het loterijnummer niet langer valide als instrument.

Een andere reden waarom het loterijnummer Z een slecht instrument is, is dat personen die een hoge Z hebben getrokken (en dus een grote kans maken om geen militaire dienst te moeten vervullen), aantrekkelijkere kandidaten zijn voor werkgevers die moeten beslissen in welke werknemers ze zullen investeren. Werknemers met een lage Z hebben daarentegen een grote kans dat ze zullen verdwijnen uit het bedrijf, en zullen dan ook minder kansen krijgen. Eigenlijk wordt dan het loterijnummer een X in de uitkomstenvergelijking.

Een aantal andere bedenkingen worden gemaakt door Rosenzweig and Wolpin 2000 in hun zeer interessant overzichtsartikel (zie ook Keane 2006). Zo houdt Angrist bij het berekenen van het loon geen rekening met het opleidingsniveau (gezien er gewerkt werd met sociale zekerheidsgegevens, was dit ook niet ter beschikking). Op het eerste zicht geen probleem, aangezien de randomisering het opleidingniveau exogeen maakt. Tenzij natuurlijk het randomiseringsproces zelf invloed heeft op het opleidingsniveau. Naar verluidt was er sprake van een zekere tijdsverloop tussen het moment dat men een nummer trok en het moment dat bekend werd gemaakt onder welke grens de nummers in aanmerking zouden worden genomen voor militaire dienst. Wie een laag nummer trok, was dus gedurende een zekere tijd in onzekerheid over het feit of hij al dan niet zou worden opgeroepen. Deze toegenomen kans op een mogelijke toekomstige onderbreking van de schoolloopbaan (of arbeidsmarktloopbaan voor degenen die al werkten), vermindert de opbrengst van investeringen in menselijk kapitaal op dat moment. Dit maakt dat binnen de groep van lage nummers, met inbegrip van zij die uiteindelijk niet zouden worden opgeroepen, men mogelijk tijdelijk de investering in menselijk kapitaal verminderde, wat voor een stuk lagere inkomsten later zou kunnen verklaren.

Verder dient opgemerkt dat het in de Angrist-studie gevonden effect een LATE is (zie ook sectie 3.3.1): het geeft het effect weer van de militaire dienst voor die subgroep die door de loterij werd verplicht om deel te nemen aan de militaire dienst, en die dus in afwezigheid van de loterij nooit op vrijwillige basis voor de militaire dienst zouden hebben gekozen (gesteld tenminste dat men bereid is om een aantal weinig realistische veronderstellingen te maken die de bovenstaande problemen vermijden). Keane 2006 geeft een inzichtelijk cijfervoorbeeld. Stel dat er twee types zijn, die allebei 100€ zullen verdienen als ze geen militaire dienst vervullen. Het type 1 zal echter na militaire dienst 20% winnen, het type 2 daarentegen zal na militaire dienst 20% verliezen. Stel verder dat in de populatie er 20% van het type 1 zijn, en 80% van het type 2. Het gemiddelde inkomensverlies t.g.v. militaire dienst is dan -12% (namelijk [(0,2 * (+20)) + (0,8*(-20))].

Veronderstel dat 20% van degenen met een laag nummer (nummer beneden de drempel) ook effectief verplicht worden om in dienst te gaan. Als dan de verschillende elementen van de schatter Δ^{IV} worden ingevuld, krijgen we, gegeven dat men ook op vrijwillige basis kon dienen,
$\Delta^{IV} = [E(Y|Z^*=1) - E(Y|Z^*=0)] / [Pr(D=1|Z^*=1) - Pr(D=1|Z^*=0)]$
= (100,8 - 104,0) / (0,36 - 0,20) = -20%.[10] Dit cijfer is niet alleen beduidend lager dan het gemiddelde inkomensverlies van -12%, het is bovendien gelijk aan wat de types 2 verliezen wanneer ze in dienst moeten. Niet toevallig, want IV geeft, zoals reeds gezegd, wanneer slechts een subpopulatie reageert op een wijziging van het instrument, in het geval van heterogene behandelingseffecten, een LATE. De types 1 tellen uiteindelijk zelfs niet mee in de berekening, omdat ze de beide verwachte inkomenstermen met eenzelfde bedrag doen toenemen. Dit is ook wat werd voorspeld, de types 1 worden bij hun keuze immers niet beïnvloed door het instrument. Als vrijwillige indiensttreding wordt uitgesloten, geeft de formule (97,6-100)/(0,20-0) = -12%, het gemiddelde effect.

Te noteren valt dat het LATE instrumentafhankelijk is. Bij de keuze van een andere Z krijgt men een ander LATE. Daarenboven is voor een gegeven instrument het LATE gedefinieerd voor een niet-gedefinieerde hypothetische populatie ('de personen die zeker veranderen van 0 naar 1 als het instrument wijzigt'), waarbij voor verschillende waarden van Z en bij het gebruik van andere instrumenten de LATE parameter wijzigt en ook de populatie waarvoor het wordt bepaalt, wijzigt. Heckman besluit dan ook dat gegeven deze onduidelijkheden, het helemaal niet duidelijk is op welke interessante beleidsvraag een geschat LATE het antwoord geeft (Heckman 1997).

Bij wijze van besluit kan gesteld worden dat het natuurlijk experiment zeker niet het wondermiddel is dat in één klap alle problemen van sociale experimenten en niet-experimentele methoden oplost. Zo is er alleen al het praktische bezwaar dat t.a.v. vele evaluatievragen niet zo maar een natuurlijk experimenteel design kan worden opgezet. Natuurlijke 'natuurlijke experimenten' kunnen op een elegante manier zorgen voor een randomisering, maar zelfs dan blijkt het lang niet zeker te zijn dat de randomisering het effect van alle relevante X kan exogeen maken, zodat een conditionering op sommige X, in de beste niet-experimentele traditie, nodig blijft. Daarnaast moet men steeds verdacht blijven op de complicaties die de heterogene behandelingseffecten met zich mee brengen.

Al deze opmerkingen gelden a fortiori voor de 'niet-natuurlijke' natuurlijke experimenten.

3.3.3 Selectiemodellen

De benadering van steekproef-selectiemodellen start bij een baanbrekend artikel van Heckman (Heckman 1979). Typerend aan deze modellen is dat zowel het deelnameproces als de uitkomst expliciet worden gemodelleerd, en dat bij de schatting wordt rekening gehouden met hun wederzijdse afhankelijkheid.

Hierbij moeten twee veronderstellingen worden gemaakt:

10 De kans dat men dienst doet, gegeven dat men een hoog nummer heeft, is uiteraard 20%. Dit zijn allemaal vrijwilligers, alle types 1 met een hoog nummer. Voor degenen met een laag nummer bestaat de groep van degenen die effectief dienst doen vooreerst uit de 20% die daartoe wordt verplicht en daarnaast zal 20% van de overige 80% met een laag nummer vrijwillig toetreden (met name die van het type 1). Samen geeft dit 36%. Binnen de groep met een hoog nummer kiest 80% zeker niet voor vrijwillige dienst, 20 % wel. Hun respectievelijke inkomens zijn dan 100 en 120. En dan geeft (0,8*100) + (0.2*120) het bedrag 104. Binnen de groep met een laag nummer, zou, als vrijwillige toetreding niet mogelijk zou zijn, uiteindelijk 80% geen dienst vervullen. De overige 20% bestaat op haar beurt uit 80% van het type 1, en 20% van het type 2. Het verwachte inkomen is dan ook (0.8*100) + 0,2*[(0.2*120) + (0.8*80)]= 97,6. Als men wel vrijwillige deelname toestaat, wordt dit: 0,8*[(0,2*120)+(0,8*100)] + 0,2*[(0.2*120) + (0.8*80)]=100,8.

(a) er moet een bijkomende regressor zijn die alleen voorkomt in de vergelijking die de deelname beschrijft (de selectievergelijking), die een coëfficiënt heeft die verschilt van nul, en die onafhankelijk is van de foutenterm V van de deelnamevergelijking.

(b) de gemeenschappelijke dichtheid van de verdelingen van de foutentermen U_{it} en V_i moet gekend zijn en moet schatbaar zijn.

In essentie gaat deze benadering rechtstreeks controleren voor dat gedeelte van de restterm U_{it} dat correleert met de indicatorvariabele die de deelnamestatus weergeeft. Typisch wordt daarbij in twee stappen tewerk gegaan. In een eerste stap wordt het gedeelte van de restterm U_{it} dat correleert met D_i geschat. In een tweede stap wordt dit dan expliciet opgenomen bij de regressoren X van de uitkomstvergelijking, zodanig dat de resterende restterm in de uitkomstvergelijking gezuiverd is voor het gedeelte dat samenhangt met de deelnamevariabele.

In het basismodel (maar er bestaan talloze varianten) wordt de uitkomstvergelijking gegeven als een lineaire regressie $Y_{it} = X_i\beta_0 + \Delta_t D_{it} + U_i$ en de selectievergelijking door $D_i^* = Z_i\gamma + V_i$ waarbij D^* de latente index is uit vergelijking (14), en de rechterkant van (14) hier in lineaire termen werd geschreven.

Als dan wordt verondersteld dat U_i en V_i bivariaat normaal verdeeld zijn (een mogelijke manier waarop aan één van de vereisten van voorwaarde (b) kan worden voldaan), kunnen we schrijven:

$$\begin{aligned} E(Y_i \mid D=1) &= E(Y_i \mid D^* > 0) \\ &= E(Y_i \mid Vi > -Z_i\gamma) \\ &= \beta_0 + \Delta_t + E(U_i \mid Vi > -Z_i\gamma) \end{aligned} \quad (41)$$

Nu bestaat er een mooi resultaat in de statistiek dat zegt dat als U en V bivariaat normaal verdeeld zijn met gemiddelden μ_U en μ_V, met standaarddeviaties ς_U en ς_V en met correlatie ρ, geldt dat (Greene 2000):

$$E(U \mid V > a) = \mu_U + \rho\,\varsigma_U\,[\varphi(a_z)/[1-\Phi(a_z)]] \quad (42)$$

Waarbij a_z staat voor een gestandaardiseerde a, namelijk $a_z=(a-\mu_V)/\varsigma_V$, waarbij $\varphi(k)$ staat voor de standaardnormale dichtheid geëvalueerd op waarde k, en $\Phi(k)$ staat voor de standaardnormale cumulatieve verdeling, geëvalueerd op waarde k.

Gegeven dit resultaat, kan dan (41) herschreven worden als volgt:

$$E(Y_i \mid D=1) = \beta_0 + \Delta_t + \rho\,\varsigma_U\,[\varphi(Z_i\gamma/\varsigma_V)/[1-\Phi(Z_i\gamma/\varsigma_V)]] \quad (43)$$

Via een analoge weg kan ook een uitdrukking worden gevonden voor $E(Y_i \mid D=0)$. Het aantrekkelijke van dit resultaat is dat er een uitdrukking wordt bekomen die het gedeelte van U dat correleert met de deelnamestatus D weergeeft. Als deze uitdrukking kan worden berekend voor iedereen in de steekproef, krijgt men een bijkomende variabele die eenvoudig wordt toegevoegd aan de X van de uitkomstenvergelijking, en wordt er gecontroleerd voor steekproefselectie. In de praktijk wordt de latente index D^* natuurlijk niet waargenomen. Wat men wel waarneemt is D, die ofwel 0 ofwel 1 is. Het is dan vb. mogelijk om een probitmodel te schatten, dat uitgaande van D en Z de kans op deelname schat gegeven Z. Na schatting van zo een probitmodel $Pr(D=1) = Z_i\gamma$ is het dan vervolgens een louter rekenkundige oefening om de waarde van de laatste term uit (43) te berekenen voor iedere deelnemers, en de waarde van een gelijkaardige term voor de niet-deelnemers. Eens

die berekening gedaan, schat men dan een lineaire regressie voor de uitkomst, waarbij de berekende waarden één van de regressoren is.

Terugkijkend naar voorwaarde (a) zal vlug duidelijk zijn dat hier eigenlijk gevraagd wordt naar een instrument: een variabele die de deelname beïnvloedt en niet de uitkomst. Op zich kan de voorgestelde schattingswijze ook wel worden toegepast als er zo geen variabele wordt opgenomen. Aangezien in de praktijk echter de verzameling van (geobserveerde) variabelen die de deelname beïnvloeden heel erg gelijkaardig is aan de verzameling van variabelen die de uitkomst meebepalen, krijgt men dan de facto de situatie waarin de rechterkant van de deelnamevergelijking heel erg gelijkend is aan de rechterkant van de uitkomstenvergelijking. De identificatie van de modellen berust dan volledig op verschillen in de functionele vorm (in het bovenstaande voorbeeld een lineaire uitkomstenvergelijking versus een niet-lineaire probitvergelijking), of op verschillen in de resttermstructuur. Dat is wellicht wat mager, en kan aanleiding geven tot grote standaardfouten bij de schatting (hetgeen de grotere mate van onzekerheid weerspiegelt).

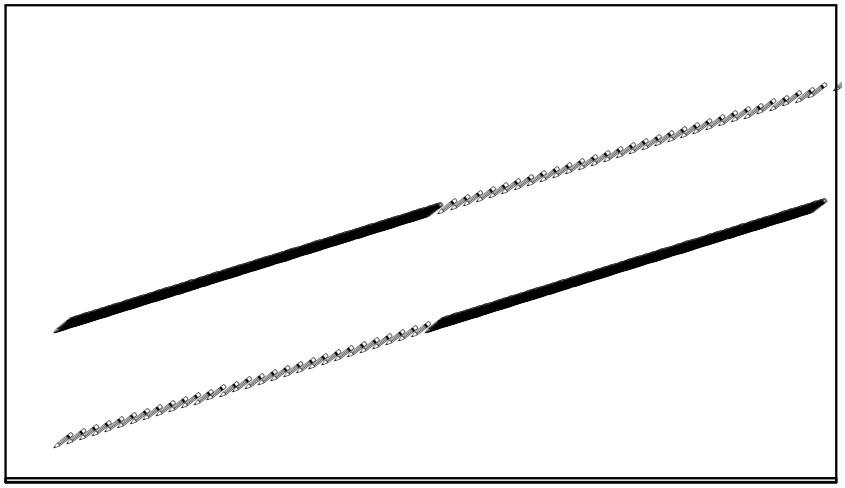
3.3.4 Regressie-discontinuïteitsmodel

Het 'regression discontinuity model' is van toepassing in het speciale geval waar er weliswaar selectie op basis van geobserveerde verschillen is, maar waar er anderzijds helemaal geen gemeenschappelijke support is. In dat geval zijn matching-benaderingen uiteraard niet langer toepasbaar.

Het model vereist dat het al dan niet verkrijgen van een behandeling afhangt van een geobserveerde variabele S, en wel op een gekende, deterministische manier. Zo voldoet vb. het volgende mechanisme: D=1 als $S < S_0$, zo niet is D=0. Als de uitkomst Y_0 afhangt van S, en als $\Delta \neq 0$, dan zal deze toewijzingsregel zorgen dat er een discontinuïteit optreedt in de relatie tussen enerzijds $Y = Y_0 + \Delta D$ en anderzijds S, en wel op het punt waar $S= S_0$.

Barnow e.a. 1980 geven een denkbeeldig voorbeeld (Barnow et al. 1980). Stel dat er een lineair verband bestaat tussen het gezinsinkomen en de schoolprestaties van de kinderen (...). Als dan een programma wordt opgezet dat gericht is op een verbetering van de schoolprestaties, en waarbij alle kinderen waarvan het gezinsinkomen beneden een bepaalde grens zit, deelnemen, en alle andere kinderen worden uitgesloten, ontstaat een discontinuïteit die de impact van het programma identificeert. Figuur 1 geeft een grafische weergave van dit voorbeeld.

Figuur 1. Regressie-discontinuïteitsmodel



De onderste rechte geeft het verband tussen gezinsinkomen (horizontale as) en schoolprestaties (verticale as) voor de niet-behandelden. Waar de stippellijn ophoudt, bereikt men de inkomensdrempel S_0. Alle kinderen uit gezinnen boven deze drempel worden niet behandeld en zitten dus op de onderste rechte. De stippellijn op de onderste rechte laat dan zien wat het resultaat van kinderen uit gezinnen met een inkomen beneden S_0 zou zijn geweest, mochten ook zij niet behandeld zijn. De bovenste rechte geeft het resultaat na behandeling. Dit design laat onmiddellijk toe om de impact te identificeren: Δ kan zonder vertekening worden geschat via de vergelijking $Y = \beta_0 + \Delta D + \beta_1 S + U$.

Natuurlijk moet daarbij worden opgelegd dat (a) de relatie tussen inkomen en resultaat lineair is; en (b) behandelingseffecten homogeen zijn. Bij heterogene behandelingseffecten kan alleen maar met zekerheid worden vastgesteld wat de grootte is van de impact voor gevallen die zich bevinden ter hoogte van de discontinuïteit.

Te noteren valt dat dit een specifieke toepassing is van het ('onnatuurlijk') natuurlijk experiment.

Controle en sanctionering van (de zoekinspanning) van werklozen, vervolg

Een bespreking van het opzet van het plan voor de activering van het zoekgedrag werd gegeven in sectie 3.2.1. Bij de opstart van dit plan werd omwille van capaciteitsproblemen met verschillende fases gewerkt, niet in het minst omdat bij de aanvang niet alleen al de instromers in de langdurige werkloosheid tot de doelgroep behoorden, maar bovendien ook de reeds bestaande voorraad van mensen die al langdurig werkloos waren, nog de revue moesten passeren.

In het eerste jaar werden de doelgroep beperkt tot personen jonger dan 30 jaar, gedurende het tweede jaar (vanaf juli 2005) werd dit uitgebreid tot de langdurig werklozen jonger dan 40, gedurende het derde jaar (vanaf juli 2006) werd de doelgroep uitgebreid tot 50 jaar. Langdurig werklozen ouder dan 50 vallen niet onder het plan.

Bij het overschrijden van de grens van de langdurige werkloosheid krijgen de werklozen een aankondigingsbrief waarin hun verplichtingen worden uitgelegd, en wordt aangekondigd dat ze circa 8 maanden later zullen worden opgeroepen. Voor de voorraad van reeds langdurig werklozen werd een iets andere procedure gevolgd: zo werd in het eerste jaar de voorraad van de langdurige werklozen jonger dan 30 jaar geleidelijk gedurende de loop van het jaar opgeroepen in een volgorde overeenkomstig hun leeftijd, startend bij de jongsten.

De drie fasen bij de invoering van het plan leiden tot een discontinuïteit in de relatie tussen leeftijd en het krijgen van een behandeling. Deze discontinuïteit wordt door Cockx en Dejemeppe gebruikt om het effect van het plan te identificeren (Cockx and Dejemeppe 2007).

Op basis van een literatuurstudie stellen ze vast dat er weinig evidentie is te vinden voor de stelling dat het 'monitoren' van werklozen effectief zou zijn (onder monitoring wordt het opvolgen van de zoekinspanningen van de werkloze begrepen, waarbij typisch ook sancties worden gegeven in het geval wordt geoordeeld dat er te weinig inspanningen gebeuren. In wat volgt, spreken we over controle (van de zoekinspanning). 'Monitoring' of controle staat dan tegenover 'counseling', hetgeen we in wat volgt begeleiding noemen). Een van de aangehaalde redenen voor de geringe effectiviteit is dat de PES meestal enkel de meer formele elementen van het zoekgedrag kan beoordelen, of ook omdat bij een erg strenge controle een zeker aandeel van de werklozen hun uitkering zullen opgeven en vervolgens nog minder intensief zullen zoeken.

Maar ook als controle als dusdanig geen invloed zou hebben op de overgang van werkloosheid naar werk, dan nog is het mogelijk dat er wel een effect uitgaat van de dreiging dat men zal gecontroleerd worden. Als werklozen er niet van houden gecontroleerd te worden, aldus deze hypothese, dan zullen ze meer zoekinspanningen doen en trachten werk te vinden vooraleer ze worden gecontroleerd (zelfs al gaat dat dan misschien ten koste van de kwaliteit van het aangenomen werk).

De auteurs gaan dan ook na of er een dreigingseffect uitgaat van de brief die aankondigt dat men zich 8 maanden later zal moeten verantwoorden. Bij de identificatie van het effect wordt de discontinuïteit tussen leeftijd en behandeling uitgebuit: de geselecteerde behandelde groep bestaat uit de 25 tot 30 jarigen die tussen juli en oktober 2004 een aankondigingsbrief kregen, de vergelijkingsgroep bestaat uit de 30 tot 35 jarigen die gedurende dezelfde periode ook een aankondigingsbrief zouden hebben gekregen, als ze jonger waren geweest dan 30 (door het feit dat de voorraad van langdurig werklozen binnen iedere leeftijdsgroep geleidelijk *in functie van hun leeftijd* een brief werd toegestuurd, verdween daar de discontinuïteit: de nog net geen 30 jarige langdurig werklozen kregen hun aankondigingsbrief op het einde van het eerste jaar, dat wil zeggen, vlak voor het moment waarop de net iets ouder dan 30 jarige langdurig werklozen hun brief zullen ontvangen, nl. bij de start van het tweede jaar, waardoor een continuïteit in de relatie tussen leeftijd en behandeling ontstaat. Daarom kon de voorraad niet worden meegenomen in de studie).

De schattingen van de auteurs suggereren dat de aankondigingsbrief de kans op werk met 4% deed toenemen, althans wat betreft Vlaanderen en Wallonië. Dit homogeen effect is evenwel statistisch niet te onderscheiden van nul. Als men evenwel heterogeniteit in de effecten toelaat, stelt men voor Vlaanderen een beduidend sterker effect vast, dat evenwel enkel geldt voor hoger opgeleiden. Zoals te verwachten, stijgt het effect ook naarmate de datum van het opvolgingsgesprek nadert.

3.3.5 Het verschil van de verschillen

Een wat naïeve benadering van behandelingseffecten gaat uit van de overigens legitieme claim dat een persoon nog het best vergeleken kan worden met zichzelf. Men kan weliswaar niet tegelijkertijd behandeld en niet behandeld zijn, maar men kan wel op verschillende tijdstippen in de verschillende toestanden zitten. De beste vergelijkingsgroep voor een groep van behandelde personen, aldus deze redenering, bestaat dan ook uit diezelfde groep, zij het op een vroeger tijdstip, voorafgaand aan de behandeling. Dit is het zogenaamde voor-na-design.

Weze Y_{1t} de uitkomst na deelname aan een programma, en $Y_{0t'}$ de uitkomst vooraleer men deelnam, en veronderstel dat het programma plaatsvindt op moment k, met $t' < k < t$, dan zal een voor-na-schatter de uitkomst $Y_{0t'}$ (de uitkomst zonder behandeling in een periode voorafgaand aan de behandeling) gebruiken als een benadering voor wat de uitkomst zou zijn geweest zonder behandeling, in een periode na behandeling. De gemaakte veronderstelling is dan ook dat voor de groep van behandelden (i.e. D=1) geldt:

$$E(Y_{0t} - Y_{0t'} \mid D=1) = 0 \quad (44)$$

De voor-na-schatter wordt dan gegeven door

$$E(Y_{1t} \mid D=1) - E(Y_{0t'} \mid D=1) \quad (45)$$

of nog, door het verschil te maken van de gemiddelde uitkomst in een periode na, en de gemiddelde uitkomst in een periode voor de deelname en dit telkens voor de groep van behandelden. In de praktijk wordt gewerkt met steekproeven, i.p.v. te werken met de verwachte waarde werken we dan met het steekproefgemiddelde. Als het steekproefgemiddelde van Y_{1t} wordt voorgesteld door $\underline{\mathbf{Y}}_{\mathbf{1t}}$ enz., en als het subscript '1' verwijst naar D=1, wordt de voor-na-schatter gegeven door:

$$\Delta^{VN} = (\underline{\mathbf{Y}}_{\mathbf{1t}} - \underline{\mathbf{Y}}_{\mathbf{0t'}})_1 \quad (46)$$

Om de winst van deelname op individueel niveau bekijken, zal een individu na periode t het verschil maken tussen wat haar uitkomst voor die periode t zou zijn met deelname (nl. Y_{1t}), en wat haar uitkomst voor die periode t zou zijn zonder deelname (nl. Y_{0t}). Aangezien in een gelijkheid aan één kant iets mag worden bijgeteld als het ook wordt afgetrokken, kunnen we dit verschil ook schrijven als volgt:

$Y_{1t} - Y_{0t} = Y_{1t} - Y_{0t} + (Y_{0t'} - Y_{0t'})$, of nog,

$$Y_{1t} - Y_{0t} = (Y_{1t} - Y_{0t'}) + (Y_{0t'} - Y_{0t}) \quad (47)$$

De tweede term ter rechterzijde is een benaderingsfout. De voor-na-schatter legt niet op dat deze term op individueel niveau 0 is, maar wel dat het gemiddelde van deze term, berekend over alle deelnemers, nul is.

In dat geval zal het effect van deelname gegeven worden door van de gemiddelde post-programmauitkomst de gemiddelde pre-programmauitkomst af te trekken. Als de uitkomst het inkomen is, is dit zeer handig. Het effect van deelname wordt dan gemeten door het gemiddeld inkomen na deelname bij de deelnemersgroep te minderen met het gemiddeld inkomen van dezelfde groep voor de deelname.

Hiervoor zijn op het eerste zicht longitudinale gegevens nodig (twee metingen voor dezelfde groep, met name voor en na deelname), maar de schatter werkt ook met herhaalde doorsneden die een steekproef trekken uit dezelfde populatie op een ander moment, op voorwaarde dat benaderingsfout uitmiddelt tot nul. Dit is op zich een heel mooi resultaat, omdat het minder hoge eisen stelt aan het datainzamelingsproces.

Blijft natuurlijk dat in beide gevallen, of men nu beschikt over longitudinale gegevens dan wel over herhaalde doorsneden, opgelegd wordt dat de benaderingsfout gemiddeld nul is, d.w.z. dat de gemiddelde uitkomst in een toestand zonder behandeling hetzelfde is in periode t en t'. In het algemeen is dat een veronderstelling die moeilijk te maken is, omdat uitkomsten typisch ook samenhangen met andere fenomenen die wijzigen in de tijd, zoals de arbeidsmarktconjunctuur, of de positie in de levenscyclus binnen een cohorte van deelnemers (Heckman e.a. 1999).

Een mogelijke oplossing om uit deze impasse te geraken, is toch een beroep te doen op een vergelijkingsgroep van niet-deelnemers. Men moet dan wel bereid zijn om te veronderstellen dat datgene wat wijzigt tussen t' en t, de uitkomsten van de deelnemersgroep en de vergelijkingsgroep in een toestand zonder behandeling op een zelfde wijze beïnvloedt. Voor factoren zoals de arbeidsmarktconjunctuur (of leeftijdseffecten) is dit dikwijls een aanvaardbare stelling: de wijziging in de arbeidsmarktconjunctuur tussen t' en t is alleszins identiek voor beide groepen, en tot bewijs van

het tegendeel is het aanvaardbaar dat die identieke wijziging de evolutie in hun uitkomsten (zonder deelname) gemiddeld gezien ook identiek beïnvloedt. We krijgen dan als veronderstelling:

$$E(Y_{0t} - Y_{0t'} \mid D=1) = E(Y_{0t} - Y_{0t'} \mid D=0) \qquad (48)$$

Deze uitdrukking heeft betrekking op een toestand zonder deelname (enkel Y_0 uitkomsten), de linkerterm heeft betrekking op de deelnemers (wat zou hun gemiddelde evolutie geweest zijn als ze niet hadden deelgenomen?), de tweede term betreft de vergelijkingsgroep. In vergelijking (48) duikt weer de 'benaderingsfout' van boven op. Hier wordt evenwel niet meer opgelegd dat dit verschil gemiddeld nul moet zijn, hier wordt enkel geëist dat het gemiddelde van het verschil bij de deelnemersgroep gelijk is aan het gemiddelde van het verschil bij de vergelijkingsgroep.

Het is dan mogelijk om de voor-na-schatter zowel voor de deelnemersgroep als voor de vergelijkingsgroep te schatten. Door het verschil tussen beide voor-na-schatters te maken, maakt men eigenlijk een verschil tussen de verschillen, en krijgen we de DID-schatter ('difference-in-differences'):

$$\Delta^{DID} = (\underline{\mathbf{Y}}_{\mathbf{1t}} - \underline{\mathbf{Y}}_{\mathbf{0t'}})_1 - (\underline{\mathbf{Y}}_{\mathbf{0t}} - \underline{\mathbf{Y}}_{\mathbf{0t'}})_0 \qquad (49)$$

Eigenlijk wordt zo de wijziging over de tijd van de uitkomsten van de vergelijkingsgroep gebruikt als een soort van benchmark voor de gemeenschappelijke jaar of leeftijdseffecten. Of nog, de eerste verschil tussen haakjes in (49), het verschil van de deelnemers, wordt zowel bepaald door hun deelname aan het programma als door andere wijzigingen tussen t' en t. Voor dit laatste wordt gecorrigeerd door het tweede verschil ervan af te trekken.

Lost deze benadering het probleem van selectievertekening op? Dit vergt een genuanceerd antwoord. Een erg algemene uitdrukking voor de uitkomsten, die verder bouwt op (20), wordt gegeven door:

$$Y_{it} = g^0_t(X_i) + \Delta_{it}(X_i)\, D_{it} + (K_i + L_t + M_{it}) \qquad (50)$$

Deze uitdrukking geldt zowel voor deelnemers als voor vergelijkingsgroep. Blundell en Costa Dias (2002) ontleden de niet-geobserveerde restterm U_{it} in drie componenten. K_i heeft betrekking op persoonlijke kenmerken die constant zijn in de tijd ('individual specific fixed effect'). Een kandidaatlid voor deze categorie zijn zeker de aangeboren vaardigheden. Maar ook elementen zoals de voorkeuren van de persoon en haar motivatie kunnen hier onder vallen, zo lang ze maar ongewijzigd blijven binnen de bestudeerde tijdspanne (zeg van t' tot t). Vervolgens is er L_t, wat door de geciteerde auteurs het gemeenschappelijk macro-economisch effect wordt genoemd. Dit heeft betrekking op niet-geobserveerde wijzigingen over de tijd, die wel identiek zijn voor iedereen. Een voorbeeld is uiteraard de arbeidsmarktconjunctuur, maar ongetwijfeld ook de leeftijd: tussen t' en t wordt iedereen ouder, en wel met eenzelfde aantal jaren.

Een laatste term is dan M_{it}, de tijdelijke individueel-specifieke effecten. Als de motivatie van een persoon wijzigt over de tijd, hoort die hier thuis. Maar als bijvoorbeeld een wijzigende arbeidsmarktconjunctuur sommigen groepen anders treft dan andere, moet dat wellicht hier worden opgenomen, en niet bij de vorige term (als vb. hoger opgeleide autochtonen meer profiteren van een krappe arbeidsmarkt dan laaggeschoolde migranten, en dit niet kan worden geobserveerd).

Terugkerend naar de DID-schatter, is dan duidelijk dat de eerste verschillen (de voor-na-schatters voor respectievelijk deelnemers en vergelijkingsgroep) er voor zorgen dat de K_i effecten verdwijnen. Selectie op niet-geobserveerde verschillen is dus geen probleem, zolang ze maar constant zijn in de tijd (in die zin blijft selectievertekening als dusdanig bestaan, maar ze stelt geen probleem op voorwaarde dat ze constant blijft). Dit is op zich al een mooi resultaat. Het verschil van de ver-

schillen zal er vervolgens voor zorgen dat ook de L_t effecten wegvallen. Waar de DID-schatter evenwel niets tegen vermag, is selectie op M_{it} variabelen. Als er toch M_{it} invloeden aanwezig zijn die het resultaat mee bepalen, dan moet worden verondersteld dat ze niet meespeelden bij de selectie (opdat DID zou zijn geïdentificeerd). Of omgekeerd, als een variabele meespeelt in de selectie, dan moet worden verondersteld dat die niet wijzigt over de tijd. Dit kan iets meer compact worden geschreven als volgt:

$$E(U_{it}^0 \mid X_i, D_i) = E(K_i \mid X_i, D_i) + L_t \qquad (51)$$

Zoals gezegd, is het mogelijk dat sommige groepen anders reageren op gelijke macro-economische wijzigingen, en dit wordt een probleem als er systematisch meer van een bepaalde soort deelnemen. Bell e.a. stellen een interessante uitbreiding voor die hier in sommige gevallen rekening mee kan houden bij het gebruik van DID (Bell et al. 1999). Veronderstel dat er twee groepen zijn (Q=0 en Q=1) die op een verschillende manier reageren op de macro-economische evolutie L_t. Dit verschil wordt weergegeven door k^Q, zodanig dat het macro-effect voor de eerste groep gelijk is aan k^0L_t, en voor de tweede groep gelijk is aan k^1L_t. De selectie blijft zoals voorheen onafhankelijk van het tijdelijk individueel-specifieke effect M_{it}, zodat nu moet worden opgelegd dat:

$$E(U_{it}^0 \mid X_i, D_i) = E(K_i \mid X_i, D_i) + k^Q L_t \qquad (51')$$

De DID-schatter schat nu, zoals (49) het ATT, maar daar komt nog een term bij:

$$E(\Delta^{DID}) = \Delta^{ATT} + (k^1 - k^0)\,[L_t - L_{t'}] \qquad (52)$$

Dit zal alleen het ATT schatten als geldt dat $k^0 = k^1$, wat bij veronderstelling niet het geval is.

De oplossing voor dit probleem is vergelijkbaar met de overgang die werd gemaakt van de voorna-schatter naar de DID-schatter: daar werd de wijziging over de tijd van de uitkomsten van de vergelijkingsgroep gebruikt als een soort van benchmark voor de conjunctuur. Hier gaan we dan vervolgens een benchmark moeten zoeken, niet voor de wijziging in de conjunctuur, maar wel voor de wijze waarop de deelnemers- en vergelijkingsgroepen (ten gevolge van een andere samenstelling op het vlak van de k^Q) verschillend reageren op een wijziging in de conjunctuur.

De strategie bestaat er dan uit om een periode te zoeken in het verleden (en liefst niet al te lang geleden), waarin tussen de momenten t''' en t'' (met t''' < t'' < t' < k < t) de conjunctuur op een soortgelijke manier evolueerde als tussen de momenten t' en t. De tweede helft van vergelijking (53) geeft de gewenste benchmark:

$$[(\underline{\mathbf{Y}}_{1t} - \underline{\mathbf{Y}}_{0t'})_1 - (\underline{\mathbf{Y}}_{0t} - \underline{\mathbf{Y}}_{0t'})_0] - [(\underline{\mathbf{Y}}_{0t''} - \underline{\mathbf{Y}}_{0t'''})_1 - (\underline{\mathbf{Y}}_{0t''} - \underline{\mathbf{Y}}_{0t'''})_0] \qquad (53)$$

waarbij $(\underline{\mathbf{Y}}_{0t''} - \underline{\mathbf{Y}}_{0t'''})_1$ voor de groep die later zal gaan deelnemen, aangeeft hoe de gemiddelde uitkomsten evolueerden tussen t''' en t'', $(\underline{\mathbf{Y}}_{0t''} - \underline{\mathbf{Y}}_{0t'''})_0$ hetzelfde geeft voor de vergelijkingsgroep, en waarbij de veronderstelling is dat het verschil dat tussen deze twee verschillen werd opgetekend in een periode (t'', t''') en in afwezigheid van een programma, identiek is aan het verschil dat tussen deze twee verschillen zou worden opgetekend in een andere periode (t',t), met een gelijklopende conjunctuur, én eveneens in de veronderstelling van afwezigheid van een programma (het blijft een counterfactual). Op deze wijze krijgt met een verschil van het verschil van de verschillen.

Als de DID-schatter vervolgens wordt gecombineerd met de matchingschatter, ontstaat de zogenaamde DID-matching-schatter. Dit geeft een design dat rekening houdt met selectie op geobserveerde verschillen, selectie op niet-geobserveerde verschillen die niet wijzigen in de tijd, en desgevallend, selectieve verschillen in de mate waarin wordt gereageerd op macro-economische schokken.

Ook hier kan, zoals bij de voor-na-schatter, naast longitudinale gegevens de schatter worden toegepast op herhaalde doorsnedegegevens. Het is dus niet noodzakelijk om te beschikken over gegevens op t en t' van dezelfde personen, het volstaat om voor beide tijdstippen gegevens te hebben over personen uit dezelfde populaties. In dat geval zal men bij een combinatie van matching en DID wel driemaal moeten matchen voor elke behandelde persoon na de behandeling: één keer om een vergelijkbare behandelde te vinden voor de behandeling (i.e. op moment t'), en tweemaal om vergelijkbare niet-behandelden te vinden, met name één voor (i.e. op t') en één na (i.e. op t) het programma.

Job-search assistance

In Centeno e.a. (Centeno et al. 2005) wordt de impact bestudeerd van een programma dat werklozen ondersteuning biedt bij het zoeken van werk. Een interessant aspect aan dit programma is dat het geleidelijk werd ingevoerd, eerst in een aantal regio's, later in andere regio's. De auteurs benadrukken dat de keuze van de regio's waar werd gestart met de invoering, toevallig was. Het is dus niet zo dat eerst werd gestart daar waar de werkloosheid het grootst is, of het meest hardnekkig. Dit is interessant, want het biedt de mogelijkheid om werklozen te selecteren voor de vergelijkingsgroep die volgens de gangbare selectiecriteria met een grote kans in het programma zouden terecht gekomen zijn, maar dit (nog) niet zijn omdat ze in een regio wonen waar het programma nog niet werd ingevoerd. Dit noemt men de pseudobehandelden.

De auteurs stellen dan een difference-in-difference-in-differences-matching-benadering voor. Er is vooreerst de groep van werklozen die in aanmerking komen voor het programma. Deelname is verplicht. In Regio 1, waar het programma al werd ingevoerd, bestaat die groep van in aanmerking komende werklozen dan ook volledig uit effectief behandelden. In Regio 0, waar het programma nog niet werd ingevoerd, bestaat de groep van in aanmerking komende werklozen volledig uit niet-behandelden. Men kan dan voor de behandelden uit Regio 1 een voor-na-effect schatten, en idem voor de in aanmerking komende, maar nog niet behandelden uit regio 0. Het verschil van die twee verschillen geeft dan een DID-schatter die controleert voor verschillen in K_i en L_t, maar wellicht ook voor M_{it} (als verplichte deelname echt verplicht is, als de keuze van de regio's echt toevallig is, kortom in de mate waarin de invoering van het programma in Regio 1 en niet in Regio 0 een proces van toevalstoewijzing benadert). Op te merken valt dat de auteurs een beroep doen op de herhaalde doorsnede variant van de schatter.

Omdat niet kan worden uitgesloten dat er verschillen zijn tussen Regio 0 en Regio 1 die niet gerelateerd zijn aan het programma, maar wel invloed kunnen hebben op de arbeidsmarktuitkomsten, wordt een derde trap toegevoegd. Daartoe worden twee nieuwe groepen aan het design toegevoegd, met name de niet voor het programma in aanmerking komende werklozen uit Regio 0, en de niet voor het programma in aanmerking komende werklozen uit Regio 1. Ook voor deze twee groepen wordt weer telkens het voor-na-verschil geschat. Het verschil van die twee geeft een DID, dat dan als benchmark kan dienen om te corrigeren voor mogelijke regionale verschillen. Samen geeft dit dan een verschil van het verschil van de verschillen.

De resultaten suggereren dat er een eerder beperkt effect uitgaat op de werkloosheidsduur, i.e. behandelden zien hun werkloosheidsduur gemiddeld met een maand dalen. Daarnaast zou er sprake zijn van een negatief effect op het loon bij wedertewerkstelling van behandelden. De auteurs stellen dan ook de effectiviteit van het programma ter discussie.

3.3.6 Duurmodellen

Duurgegevens hebben betrekking op duurtijden, die al dan niet voltooid zijn: als een werkloze na 10 maand de werkloosheid verlaat, is sprake van een voltooide duur van 10 maanden. Als we daarentegen een werkloze observeren die al 10 maanden werkloos is, is er sprake van een onvoltooide duur: mogelijk blijft de persoon nog lang werkloos, mogelijk is morgen de werkloosheidperiode afgelopen. In beide gevallen is de duur gelijk aan 10 maanden, maar toch is er een duidelijk verschil. Dit fundamentele kenmerk van duurgegevens heeft aanleiding gegeven tot de ontwik-

keling van een afzonderlijke klasse van statistische modellen: duurmodellen. Meestal wordt er gewerkt met het concept 'hazard'. De hazard θ(t) is een voorwaardelijke kans, met name de kans dat men een bepaalde toestand verlaat na een duurtijd t, gegeven dat men tot t in de toestand zit (vb. de kans op het verlaten van de werkloosheid na 11 maanden, gegeven dat men al 11 maanden werkloos is). Er kunnen dan hazardmodellen worden geschat, waarin typisch (1) een globaal patroon wordt geschat dat weergeeft hoe de (voorwaardelijke) uitstroomkans evolueert met de duur, en (2) waarin wordt gespecifieerd hoe dit globaal patroon (de zogenaamde baseline hazard) wordt beïnvloed door verklarende veranderlijken, al dan niet geobserveerd.

Binnen de evaluatie van arbeidsmarktmaatregelen zijn hazardmodellen populair, omdat ze toelaten om te schatten in welke mate deelname aan een programma het verblijf in de werkloosheid verkort (of verlengt).

Abbring en van den Berg hebben een methodologie ontwikkeld die toelaat om behandelingseffecten te identificeren bij duurgegevens (Abbring and van den Berg 2003). Daarbij wordt gewerkt met twee verschillende duren. De duur T_m meet hoe lang het duurt vooraleer een individu overgaat naar een bepaalde gewenste toestand (vb. duur tot het verlaten van de werkloosheid, of duur tot het vinden van werk). De duur T_p meet hoe lang het duurt vooraleer het individu start met de behandeling. Beide duurtijden starten op hetzelfde moment (moment nul, de start van de werkloosheidsperiode). Deze twee duren zijn toevalsvariabelen, en concrete realisaties (de waarden voor een bepaald individu) worden aangeduid met t_m en t_p. Daarnaast zijn er geobserveerde kenmerken X, en niet-geobserveerde kenmerken V (respectievelijk V_m en V_p).

De kans dat een individu op tijdstip t wordt behandeld, gegeven dat zij al een duur t werkloos is, wordt dan gegeven door:

$$\theta_p(t \mid X,V_p) = \lambda_p(t) \exp[X\beta_p + V_p] \qquad (54)$$

De λ term is de baseline hazard, en geeft weer hoe de hazard evolueert met de duur. De tweede term (exp[...]) geeft weer wat de invloed is van de geobserveerde en niet-geobserveerde kenmerken op deze baseline hazard. De specificatie is multiplicatief: als bij wijze van voorbeeld $\exp(X\beta_p)$ voor de ene groep van mensen gelijk is aan 1, en voor een ander groep gelijk aan 2, geldt dat voor de eerste groep de baseline hazard van toepassing is, terwijl voor de tweede groep de uitstroomkansen op iedere t dubbel zo groot zullen zijn (men noemt dit een proportionele hazard specificatie).

De kans dat het individu op tijdstip t naar een gewenste toestand gaat, gegeven dat zij tot t werkloos blijft, heeft een bijkomende term, aangezien deze ook beïnvloed wordt door de behandeling (en het tijdstip van de behandeling).

$$\theta_m(t \mid t_p,X,V_m) = \lambda_m(t) \exp[X\beta_m + V_m + \mu(t-t_p)I(t > t_p)] \qquad (55)$$

Deze hazard in qua opbouw zeer gelijkaardig met de voorgaande. Aan de linkerkant wordt er t_p bijgeschreven, om aan te geven dat de uitstroomkans ook afhankelijk is van het moment waarop men aan een behandeling begon. Aan de rechterkant vinden we terug de baseline hazard, de rol van de niet-geobserveerde kenmerken $\exp(V_m)$ en tot slot de invloed van de geobserveerde kenmerken. Naast de invloed van de X kenmerken is er nu een bijkomend kenmerk opgenomen, met name het behandelingseffect $\exp[\mu(t-t_p)I(t > t_p)]$. Hier is $I(t > t_p)$ een indicator die de waarde 1 aanneemt als $t > t_p$, d.w.z. wanneer men begint aan een behandeling vooraleer men uitstroomt. Als, bij wijze van spreken, men zo vlug uitstroomt dan men niet de kans zag om een behandeling te ondergaan, is I gelijk aan nul, en valt de laatste term in (55) uiteraard weg.

Als men wel een behandeling start, zal de baseline hazard worden vermenigvuldigd met exp[μ(t-t_p)]. Er wordt met andere woorden geen homogeen behandelingseffect exp(μ) gespecifieerd, maar wel een behandelingseffect dat kan variëren, afhankelijk van de tijd die verloopt tussen de start van de behandeling en de uitstroom.

Een fundamenteel verschil tussen deze benadering en de voorgaande, is dat in die voorgaande het krijgen van een behandeling als een binair gegeven wordt beschouwd (ja/nee), terwijl er hier expliciet wordt rekening gehouden met het moment waarop de behandeling plaatsvindt. Het is met name die laatste informatie die gebruikt wordt om het behandelingseffect te identificeren: de basisintuïtie is dat de realisatie van T_p de vorm van de hazard van T_m gaat wijzigen, *en wel vanaf het punt t_p en verder*, en, naar de auteurs veronderstellen, op een deterministische manier. Door te kijken naar de gemeenschappelijke (bivariate) verdeling van t_m en t_p, wordt een tweedimensionaal stochastisch (= toevals)proces beschreven, waarin het voorkomen van een gebeurtenis in de ene dimensie, de hazard rate in de andere dimensie beïnvloedt. Het op een bepaald moment starten met een behandeling (dus in de dimensie T_p), zorgt ervoor dat er een fundamenteel verschil is in de andere dimensie T_m tussen de tijd voor en de tijd na t_p. Als er voldoende observaties zijn, met voldoende variatie in de duurtijden waarop er wordt gestart met de behandeling, zal de 'breuk' in de hazard van T_m perfect te detecteren zijn. Voor een soortgelijke benadering in een andere context verwijzen de auteurs naar een onderzoek waarin de gemeenschappelijke duurtijden van huwelijk, niet-huwelijk en levensduur samen worden gemodelleerd, en waarin wordt toegestaan dat het patroon van de kans op overlijden wijzigt op momenten van huwelijk of huwelijksontbinding. Ook hier is er een 'behandeling' die de duur in een andere dimensie (namelijk de levensduur) wijzigt zodanig dat het effect van de behandeling op de dimensie levensduur kan worden nagegaan. Er wordt overigens niet bij vermeld wat het teken is van het effect van de respectievelijke behandelingen.

Natuurlijk kan er hier ook selectiviteit optreden. Dit zal o.m. het geval zijn als personen met een relatief hoge uitstroomkans ook een hoge kans hebben om aan het programma deel te nemen. Een hoge uitstroomkans kan in dat geval natuurlijk gewoon te verklaren zijn door een positief behandelingseffect, maar het is mogelijk ook te verklaren doordat de deelnemers gemiddeld een hogere V_m hebben, en dus sowieso sneller uitstromen. In dat laatste geval zal er een positieve correlatie zijn tussen V_m en V_p zodanig dat ook de gemeenschappelijke verdeling $G(V_m, V_p)$ moet worden gespecifieerd.

De hele aanpak is niet-parametrisch. De baseline hazard wordt m.a.w. vrijgelaten en niet gedwongen om een bepaald vaststaand patroon te volgen. Hetzelfde geldt m.b.t. de specificatie van de niet-geobserveerde heterogeniteit (V_m en V_p).

In het basismodel wordt opgelegd dat t_p geen invloed mag hebben op $\theta_m(t \mid t_p, X, Vm)$ voor waarden van t die voorafgaan aan t_p. Dit sluit anticiperingseffecten uit. Het aankondigingseffect dat werd bestudeerd in het onderzoek van Cockx e.a. dat werd besproken in sectie 3.3.4 is een voorbeeld van zo een anticiperingseffect. Het model kan overigens wel uitgebreid worden om hier rekening mee te houden.[11]

Het aantrekkelijke aan deze benadering is dat er geen straffe 'exclusiebeperkingen' moeten worden opgelegd aan de x. Een typische 'exclusiebeperking' die in wat voorafging regelmatig

11 Te noteren valt dat anticiperingseffecten bij de voorgaande designs ook dikwijls moeten worden uitgesloten. Zo kunnen er problemen ontstaan bij een DID-schatter als toekomstige deelnemers, die weten dat ze in een latere periode zullen deelnemen, in de periode voorafgaand aan de deelname gemiddeld wat minder hard gaan werken, en bijgevolg het gemiddelde inkomen in de periode voor deelname daalt (t.g.v. het programma). Dit probleem, gekend als "Ashenfelters dip", blijkt in de praktijk dikwijls voor te komen.

onder een of andere vorm terugkwam, was dat er een variabele moest zijn die weliswaar samenhangt met de deelname maar niet met de uitkomst. Een andere exclusiebeperking was de Conditional Independence Assumption bij het matchingverhaal, waar met name moest worden verondersteld dat de beschikbare data alle systematische determinanten van het deelnameselectieproces bevatten.

3.3.7 Tijdveranderlijke behandelingsindicator

De tijd tot het moment van deelname duikt ook op een andere manier op. In Sianesi (Sianesi 2004) wordt de vergelijkingsgroep getrokken uit de personen die niet deelnamen voor een bepaald tijdstip.

Volgens Sianesi is het voor een werkloze niet zo zeer de vraag of men al dan niet zal deelnemen aan het programma, maar veeleer of ze nu dan wel later zullen participeren (als er 'locking in'-effecten zijn, kan het voor een werkloze meer optimaal zijn om eerst nog een tijd op eigen kracht naar werk te zoeken). De te bestuderen populatie op tijdstip u bestaat uit al wie na u maanden nog werkloos is. Als men op tijdstip u een behandeling ondergaat, krijgen we dat $D^u = 1$. De vergelijkingsgroep bestaat dan uit alle personen die minstens tot op het moment u nog niet beslisten om deel te nemen, voor hen geldt dan dat $D^u = 0$.

De uitkomst wordt dan ook gedefinieerd over de tijd. Als een individu op moment u deelneemt, wordt de uitkomst genoteerd als Y_t^{1u} , als een individu daarentegen minstens tot op moment u nog niet heeft deelgenomen, is er de uitkomst Y_t^{0u}. Voor ieder punt van reeds verlopen werkloosheidsduur kan er dan een impact worden omschreven als volgt:

$$\Delta_t^u = E(Y_t^{1u}-Y_t^{0u} \mid D^u=1) = E(Y_t^{1u} \mid D^u=1) - E(Y_t^{0u} \mid D^u=1) \qquad (56)$$

Dit geeft de gemiddelde impact, met name dat wat zij die op moment t, na u maanden werkloosheid, deelnemen aan het programma, winnen t.o.v. waar ze zouden zijn uitgekomen als ze op moment t, na u maanden werkloosheid, (nog) niet zouden toegetreden zijn. Uiteraard is de tweede term in (56) een counterfactual, en moeten de nodige veronderstellingen worden gemaakt om deze te kunnen identificeren:

$$Y_t^{0u} \perp D^u \mid X=x \text{ voor } t = u, u+1,...,T \qquad (57)$$

Rekening houdende met de geobserveerde kenmerken X, moet de counterfactuele verdeling van Y_t^{0u} voor personen die deelnemen na u maanden werkloosheid gelijk zijn aan de feitelijke verdeling van zij die (nog) niet deelnemen na u (er wordt gesproken over een verdeling omdat deze voorwaarde wordt opgelegd aan alle toekomstige uitkomsten Y_t^{0u}, Y_t^{0u+1}, ..., Y_t^{0T}). De vergelijking gebeurt dus steeds tussen personen met een gelijke reeds verlopen werkloosheidsduur.

4. Meta-analyse

Er worden jaarlijks tientallen impactevaluaties gemaakt van arbeidsmarktprogramma's is tal van landen. Als men al die resultaten op een verstandige wijze samenbrengt, kan men een soort van meta-schatter formuleren, waarbij de populatie dan niet langer betrekking heeft op personen die deelnemen aan een arbeidsmarktprogramma, maar wel op de gepubliceerde of op andere wijze beschikbare evaluatieresultaten. Er zijn daar uiteraard wel wat bedenkingen bij te formuleren. Een eerste bedenking is dat de impact van een bepaald programma mogelijk afhankelijk is van tijd en ruimte. Om met het ruimtelijk aspect te beginnen, tussen landen of arbeidsmarkten onderling kunnen er erg grote verschillen zijn wat betreft de sectorale samenstelling (vb. meer industrieel

versus meer tertiair), het gemiddeld opleidingsniveau van het arbeidsaanbod, de demografie, en de (werking) van de arbeidsmarktinstituties. Dit alles maakt dat het niet vanzelfsprekend is dat wat in het ene land werkt, ook in een ander land zal werken. Er is vb. dikwijls een sterke interactie tussen de organisatie van het stelsel van werkloosheidsuitkeringen en de impact van een arbeidsmarktprogramma. Wat betreft de tijdsdimensie is het vooral de stand van de arbeidsmarktconjunctuur die invloed kan hebben op de impact van programma's. Een arbeidsmarkt die naar een situatie van volledige werkgelegenheid tendeert, zal mogelijk anders reageren op een programma dan een arbeidsmarkt met een hoge werkloosheid.

Een tweede bedenking bij een meta-analyse is het feit dat men aangewezen is op gepubliceerde bronnen. Dit zou kunnen aanleiding geven tot een soort van publicatievertekening, die optreedt als vb. studies die een positief effect vinden, een grotere kans hebben op publicatie dan studies die geen effect vinden.

Ten derde kan er sprake zijn van een soort van sneeuwbaleffect. Een metastudie maakt een stand van zaken op basis van bestaande studies, zonder zelf nieuwe gegevens toe te voegen. Na verloop van tijd kan dan een situatie ontstaan waarin diverse studies naar elkaar verwijzen, zodat het aantal referenties naar een bepaald resultaat (vb. 'opleiding werkt niet') toeneemt, en dus (schijnbaar) meer ondersteuning biedt aan dit resultaat, terwijl dat in werkelijkheid misschien niet het geval is.

Meta-analyse van het Europees actief arbeidsmarktbeleid

Een recente studie naar de effectiviteit van arbeidsmarktprogramma's in Europa werd uitgevoerd door Kluve (Kluve 2006). Er wordt vastgesteld dat het nogal meevalt met het verband tussen enerzijds de effectiviteit en anderzijds land, periode, conjunctuur of institutionele setting.

De resultaten suggereren dat traditionele opleidingen een weliswaar positieve maar toch eerder beperkte invloed hebben op de post-programma-werkgelegenheidsgraad. Klassieke directe tewerkstelling in de publieke sector doet het in vergelijking met opleiding beduidend slechter, hetzelfde geldt in het algemeen ook voor programma's gericht op jongeren. In vergelijking met opleidingsprogramma's doen 'private sector incentive programs' en 'services and sanctions' het dan weer aanmerkelijk beter. Onder 'private sector incentive programs' begrijpt de auteur programma's die werkzoekenden en/of werkgevers stimuleren met het oog op het verhogen van de werkgelegenheid. Voorbeelden zijn programma's van loonkostensubsidies, of programma's die werklozen helpen bij het opzetten van een eigen zaak. Onder 'services and sanctions' wordt alles begrepen wat te maken heeft met het ondersteunen en verhogen van het werk-zoek-efficiëntie. Voorbeelden zijn werk-zoek-opleidingen en job clubs, beroepsoriëntering, begeleiding en controle van het zoekgedrag, met inbegrip van sancties in het geval waar men niet voldoet aan de gestelde voorwaarden.

De beleidsaanbevelingen zijn voor Kluve dan ook duidelijk. Opleidingsprogramma's kunnen worden verder gezet, 'private sector incentive programs' mogen eventueel nog aan belang winnen, maar vooral moet ook veel aandacht gaan naar 'services and sanctions', deze laatste categorie van maatregelen is niet alleen effectief maar bovendien ook relatief goedkoop. Directe publieke tewerkstelling zou daarentegen moeten worden afgebouwd.

De studie van Kluve toont dat metastudies zeker een meerwaarde kunnen hebben, maar toont ook de beperkingen aan van een dergelijk design. Om een enigszins ruime steekproef te hebben voor de analyse, worden hier evaluaties m.b.t. een veelheid van zeer heterogene arbeidsmarktprogramma's samengenomen. Die heterogeniteit wordt dan vervolgens in een beperkt aantal categorieën geduwd. Verwijzend naar het toch wel belangrijke onderscheid tussen het effect van de dreiging dat men zal gecontroleerd worden en het effect van de controle zelf (zie sectie 3.3.4), kan men alleen maar vaststellen dat de categorie 'services and sanctions' toch wel wat aan subtiliteit verliest. Hetzelfde geldt ook voor de andere categorieën. Er zou wel eens een verschil kunnen

zijn tussen het effect van een opleiding gericht op een knelpuntberoep (en waar het knelpunt zich vooral manifesteert omwille van een gebrek aan gekwalificeerd aanbod) en het effect van een werkervaringsprogramma voor langdurig werklozen. Ook dat zit hier binnen eenzelfde categorie. Dit staat toch wel in contrast met het gebruik van meta-analyse in vb. de biomedische wetenschappen. Klinische studies, vb. met betrekking tot het effect van de toediening van een bepaalde molecule, hebben dikwijls betrekking op een relatief kleine groep. Om de 'power' van eventuele conclusies te vergroten, gaat men dan verschillende studies combineren in een meta-analyse. Het gaat dan echter wel over studies die alle dezelfde molecule toedienden, en waar eventuele variatie tussen de studies vb. betrekking heeft op het feit dat in de ene studie 25 mg en in een andere 50 mg werd toegediend.

5. Uitdagingen

5.1 Een kijk in de zwarte doos?

In wat vooraf ging, werden een aantal recente ontwikkelingen op het vlak van de effectiviteitsmeting overlopen. Daarbij kunnen zeker de volgende belangrijke innovaties worden vermeld:

- toenemende aandacht voor heterogeniteit van het behandelingseffect: ook het behandelingseffect wordt gekenmerkt door een bepaalde verdeling, waarvan voorheen typisch enkel het gemiddelde werd geschat (een identiek, homogeen effect voor iedereen), maar waar nu meer en meer aandacht wordt geschonken aan het feit dat de grootte van het behandelingseffect mogelijk samenhangt met geobserveerde (of niet-geobserveerde) kenmerken.

- aandacht voor de verdelingsaspecten op zich, die immers ook belangrijk kunnen zijn. Zo zullen, vanuit maatschappelijk standpunt bekeken, twee programma's die bij een zelfde populatie een zelfde gemiddelde impact bekomen, niet als gelijkwaardig worden beschouwd als het ene programma die gemiddelde impact bekomt door bij een beperkt aandeel van de deelnemers hele grote winsten te boeken, terwijl het ander programma dit zelfde gemiddelde bereikt door voor de meeste deelnemers relatief beperkte winsten boekt.

- in toenemende mate aandacht voor de dimensie tijd en de dimensie duur: het belang van de timing van de behandeling (reeds verlopen werkloosheidsduur bij de start van de behandeling), aandacht voor de duur van de behandeling zelf (locking in effecten), ook meer en meer studie van de evolutie van het behandelingseffect over een langere termijn, de zogenaamde duurzaamheid (wat bvb. in het geval van bepaalde opleidingen het negatieve effect van locking-in zou kunnen compenseren).

- stilaan wordt ook het klassieke twee of drie-periodemodel verlaten waarin de periode van behandeling en het leven na de behandeling centraal staat, (en, als de data het toelaten, ook nog het leven voor de behandeling wordt meegenomen). In de plaats daarvan wordt geëvolueerd naar een meer realistisch levensloopmodel, waarin een behandeling een soort van schok is die de levensloop al dan niet (en zo ja, tijdelijk of permanent) verlegt. Binnen zo een model kunnen dan ook andere schokken, waaronder vroegere behandelingen, worden geïncorporeerd.

- meer en meer aandacht voor alternatieve invullingen van het begrip vergelijkingsgroep: i.p.v. een groep van personen die de behandeling niet ondergaan, een groep van personen die de behandeling nog niet hebben ondergaan, of, meer radicaal, door een relatief i.p.v. een absoluut effect te schatten door het effect van alternatieve behandelingen tegen elkaar af te zetten.

Toch blijven er nog een aantal duidelijke lacunes bestaan in het effectiviteitsonderzoek. Zo blijft de behandeling zelf nog al te veel een zwarte doos, zodanig dat wel wordt gemeten welke behandeling effect heeft, maar daarom nog niet duidelijk is waarom deze behandeling effectief is, en een andere niet. Zoals bij heel wat evaluatieproblemen behelst ook hier een eerste stap in de goede richting het inzamelen van betere data: meer en betere gegevens over kenmerken van de behandeling. In de voorgaande sectie werd het voorbeeld gegeven van de categorie 'opleiding'. In de praktijk bestaan er bijzonder veel verschillende types van opleiding, alleen al als men de *inhoud* van de opleiding bekijkt. Andere dimensies zijn de *finaliteit* van de opleiding (een algemene opleiding versus een beroepsspecifieke opleiding, en in het geval van beroepsspecifiek, basisvoorbereiding op dat beroep versus specialisatie), de *opleidingsvorm* (werkplekopleiding, stage, werkervaring, klassikaal, afstandsleren, combinaties van de voorgaande, ...), de *intensiteit* en de *duur* van de opleiding, etc. Men kan deze lijst wellicht nog aanvullen. Het punt is dat met een variabele die aangeeft of iemand al dan niet een opleiding volgde, toch maar een beperkt stuk van de werkelijkheid wordt gevat. De kritiek dat bij een meta-analyse al deze verschillen worden verwaarloosd, betekent overigens helemaal niet dat de gangbare praktijk bij een klassieke evaluatie (d.w.z. op individueel niveau i.p.v. op het niveau van onderzoeken, zoals bij een meta-analyse) zoveel anders zou zijn. Het verschil is dat die verwaarlozing in het geval van een meta-analyse uit de aard van de aanpak volgt, terwijl men bij een klassieke evaluatie potentieel wel kan werken met een meer gedifferentieerde kijk op de aard van de behandeling.

5.2 Externe validiteit

In wat voorafging, werd bijna exclusief aandacht besteed aan het omgaan met het bestaan van selectievertekening, en werd dus vooral gekeken naar de interne validiteit: het gemeten effect is volledig toe te schrijven aan de behandeling, en er mag met andere woorden worden uitgesloten dan het gemeten effect ook ten dele beïnvloed werd door andere factoren. Ook al is dat zeer waardevol, als het verhaal daar zou stoppen, zou evaluatieonderzoek vanuit beleidsstandpunt slechts een geringe toegevoegde waarde hebben: de vaststelling dat een concreet programma, op het moment waarop het werd geëvalueerd voor de personen die op dat moment deelnamen wel of juist geen effect had, is historische kennis, die hoogstens betekenis kan hebben bij het verantwoorden t.o.v. de maatschappij van de aan dat programma bestede budgetten.

Om meer beleidsrelevant te zijn, moet het mogelijk zijn om op basis van onderzoek dat uiteraard intern valide moet zijn, lessen te trekken die kunnen helpen bij het vormgeven van toekomstig beleid. Opdat zulks mogelijk zou zijn, moet het onderzoek ook extern valide zijn: de resultaten van het onderzoek moeten veralgemeenbaar zijn, eventueel naar andere doelgroepen, of naar andere conjunctuuromstandigheden.

Het tot hiertoe gehanteerde kader van potentiële uitkomsten zegt niets over veralgemeenbaarheid van de resultaten. Om dat mogelijk te maken, is er in essentie behoefte aan een verruiming van het causaal model. Het potentiële uitkomstenmodel bestudeert het *effect van een oorzaak*. De behandeling is de oorzaak, en er wordt dan vervolgens gekeken naar de gerealiseerde effecten. Er is dan sprake van 'causaliteit' omdat het effect volgt op de oorzaak. De vraag waarom het effect volgde op de behandeling, wordt niet gesteld, de zwarte doos wordt niet opengemaakt. Er wordt m.a.w. niet stilgestaan bij de wijze waarop, of het mechanisme waardoor het effect werd voortgebracht. Men modelleert de *effecten van oorzaken*, maar niet de *oorzaken van effecten* (Holland 1986). Dit laatste is dan in zekere zin het beschrijven van de causaliteit op een dieper niveau.

Het is natuurlijk duidelijk waarom het potentiële uitkomstenmodel zich beperkt tot de beperkte 'effect van een oorzaak'-causaliteit: een uitbreiding naar een 'oorzaken van effecten'-causaliteit betekent immers dat men een veel algemener model moet specifiëren, dat weliswaar op veel meer

situaties van toepassing is (externe validiteit!), maar waarvoor men ook veel meer theoretische keuzes moet maken en veronderstellingen moet opleggen.

Het potentiële uitkomstenmodel laat toe om een effect te schatten zonder dat al te veel veronderstellingen moeten worden gemaakt, maar de prijs die men daarvoor betaalt is dan wel dat het gemeten effect in principe niet mag worden veralgemeend naar andere situaties. Er is dus een duidelijke afruil ('trade-off') tussen de veralgemeenbaarheid van de resultaten enerzijds, en de mate waarin structuur moet worden opgelegd aan het model.

In 'The Scientific Model of Causality' onderzoekt Heckman op welke wijze men tot externe validiteit kan komen. Hij maakt daarbij de brug tussen het potentiële uitkomstenmodel, dat ontwikkeld werd in de wereld van de biostatistiek, en de econometrische traditie van structurele modellen (Heckman 2005).

Een van de klassieke manieren om op een meer algemene manier naar de werkelijkheid te kijken, betreft met name het opstellen en gebruiken van een zogenaamd structureel model. Dit is typisch een verzameling van relaties die een werkelijkheid (of een stuk van een werkelijkheid) beschrijven, met als één van de wezenlijke kenmerken dat ook rekening wordt gehouden met interdependenties: mogelijk zorgt een wijziging in A tot een verandering van B, maar die verandering van B kan op zijn beurt een wijziging in A bewerkstelligen (vb.: meer woestijnvorming leidt tot minder regen, minder regen leidt tot meer woestijnvorming). Bij het opstellen van het model moeten vele theoretische beslissingen worden genomen: welke relaties komen in het model, welke variabelen moeten in de respectievelijke relaties worden opgenomen en op welke manier (lineair verband, etc.), welke verbanden bestaan er tussen die relaties onderling, enz. Vervolgens wordt het model geschat op basis van bestaande data. Op basis van dergelijke modellen wordt het dan typisch mogelijk te voorspellen wat het effect zou zijn van een bepaalde interventie als men ze van een reeds gekende omgeving zou verplaatsten naar een andere omgeving (andere periode, andere doelgroepen), of zelfs, te voorspellen wat het effect zou zijn van een interventie die nog niet eerder werd uitgevoerd, in geen enkele omgeving.

De brug tussen het potentiële uitkomstenmodel en de benadering van het structurele modellen wordt gevormd door het concept Marginal Treatment Effect (Heckman 2005, Heckman and Vytlacil 2005). Daarbij wordt uitgegaan van het reeds eerder geïntroduceerde model van potentiële uitkomsten:

$$Y^1 = g^1(X) + U^1$$

$$Y^0 = g^0(X) + U^0$$

De link naar de structurele vergelijkingenliteratuur wordt gemaakt door een latente beslissingsregel mee op te nemen:

$$D^* = f(Z) - V$$

waarbij D=1 als de latente $D^* \geq 0$, en D=0 anders.

D^* wordt hier gezien als het netto-nut of de winst die de beslissingsmaker heeft bij het kiezen van toestand 1. Gezien het negatieve teken, hebben individuen met een hoge waarde voor V een kleinere kans op deelname dan individuen met een lage waarde voor V, ceteris paribus. Het MTE wordt dan als volgt gedefinieerd:

$$\Delta^{MTE}(x,v) = E(\Delta \mid X=x, V=v)$$

Dit MTE is een zeer flexibel concept. $\Delta^{MTE}(x,v)$ kan worden geïnterpreteerd als het gemiddeld effect van deelname voor deelnemers die op de marge zitten tussen deelname (D=1) en niet-deelname (D=0), of die dus indifferent zouden zijn tussen behandeld worden of niet als ze een exogeen bepaalde waarde van Z zouden toegewezen krijgen die er voor zorgt dat f(Z) = V (zodanig dat $D^*=0$). Voor kleine waarden v, is $\Delta^{MTE}(x,v)$ het gemiddeld effect voor individuen met niet-geobserveerde kenmerken die maken dat zij een heel grote kans maken om deel te nemen, voor grote waarden v is het dan weer juist het gemiddeld behandelingseffect voor individuen met een v die maakt dat de kans klein is dat zij zullen deelnemen.

Het aantrekkelijke van dit concept is dat alle vroeger beschouwde effecten kunnen worden gezien als gewogen gemiddelden van het MTE. Zo kan men het ATE beschouwen als een gewogen gemiddelde van de MTE's berekend over het gehele bereik van v, dus voor alle mogelijke waarden van v. Analoog is het ATT een gewogen gemiddelde van de MTE's berekend over de individuen waarbij de individuen die het meest geneigd zijn om deel te nemen (die een lage v hebben) het grootste gewicht krijgen bij de berekening van het gemiddelde. Bij het ATU zullen analoog juist de individuen met een hoge v het grootste gewicht krijgen. Het LATE kan dan weer gezien worden als een gewogen gemiddelde van de MTE's berekend over een bepaald interval [v',v] binnen de support van V (vandaar ook de verwijzing naar 'local').

Opleidingsprogramma's voor de beroepsrevalidatie van personen met een handicap

Een toepassing van de MTE-benadering wordt gemaakt door Aakvik e.a. (Aakvik et al. 2005). De bestudeerde programma's in Noorwegen richten zich op de opleiding van personen van wie de medische toestand een verlaagde productiviteit met zich meebrengt. Er wordt gewerkt met een steekproef van 1924 vrouwen, die allen interesse vertoonden in een deelname aan de opleiding, maar van wie er uiteindelijk slechts 1244 effectief deelnamen (de overigen werden hetzij geweigerd (o.m. wegens een tekort aan plaatsen, zie ook verder), of kozen zelf om toch niet deel te nemen).

Op basis van deze steekproef wordt een drievergelijkingenmodel opgesteld en geschat, dat bestaat uit de volgende vergelijkingen:

$D_i^* = Z_i\gamma - V_i$, met, zoals voorheen, $D_i = 1$ als $D_i^* \geq 0$ en $D_i = 0$ anders;

$Y_{1i}^* = X_i\beta_1 - U_{1i}$

$Y_{0i}^* = X_i\beta_0 - U_{0i}$

Er worden discrete uitkomsten geobserveerd: drie jaar later is de persoon wel aan het werk (Y=1) of niet aan het werk (Y=0). In het model wordt dan een onderliggende, latente uitkomst Y_{1i}^* verondersteld, waarbij $Y_{1i} = 1$ als $Y_{1i}^* \geq 0$, en $Y_{1i} = 0$ anders (en analoog wat betreft de relatie tussen Y_{0i} en Y_{0i}^*). In de deelnamevergelijking wordt de regionale mate van rantsoenering als instrument opgenomen. Het is met name zo dat in sommige regio's relatief gezien minder opleidingsplaatsen ter beschikking waren dan in andere. De auteurs vermelden dat in tegenstelling tot opleidingsprogramma's voor werklozen, waar de rantsoenering van opleidingsplaatsen typisch groter zal zijn in regio's met een hogere werkloosheid (en dan uiteraard geen geschikt instrument meer is, omdat het dan ook samenhangt met de uitkomst), de rantsoenering bij programma's voor beroepsrevalidatie-opleiding niet samenhangt met de lokale werkloosheid.

Na simultane schatting van de vergelijkingen, bekomt men dan geschatte waarden voor β_0, β_1, en γ. Op basis hiervan wordt dan ATE en ATT berekend. Voor ATE bekomt men een waarde van -0,014, wat aangeeft dat als de totale populatie zou deelnemen, het programma een licht negatief effect zou hebben (de totale populatie is hier gedefinieerd als al degenen die interesse vertoonden in deelname aan de opleiding). Voor ATT bedraagt de geschatte waarde -0,11. Voor degenen die effectief deelnamen aan het programma, wordt met andere woorden een beduidend grotere negatieve impact geschat. Als de MTE-parameter in een grafiek wordt gezet met op de horizontale as de waarde van V, bekomt men een stijgende lijn. Aangezien

hogere waarden van V impliceren dat men, op basis van zijn niet-geobserveerde kenmerken een kleinere kans heeft om deel te nemen aan het programma, suggereert een en ander dat degenen die het meest kans maken om deel te nemen aan het programma, er het minst baat bij hebben, en degenen die de kleinste kans hebben op deelname, het meest baat zouden hebben bij deelname. Hetzelfde werd al gesuggereerd door het feit dat de geschatte ATE groter was dan de geschatte ATT. De auteurs gewagen dan ook van een pervers afromingseffect.

Verder wordt er ook aandacht besteed aan heterogeniteit in de effecten. Deze is al groot in de geobserveerde kenmerken (vb. een hogere leeftijd leidt tot een groter behandelingseffect), maar is ook belangrijk in de niet-geobserveerde dimensie. Zo kan men het ATE van -0.014 beschouwen als het gemiddelde van een verdeling, en achter dit gemiddelde gaat schuil dat een kwart van de populatie werk vindt dank zij het programma, en een kwart van de populatie werkloos wordt door het programma (voor de overigen heeft het programma geen invloed). Dit is dan duidelijk gerelateerd aan de deelnamekans. Als men vb. de MTE evalueert op een waarde van V die 2 standaarddeviaties onder het gemiddelde ligt (zeer grote kans op deelname), blijkt dat slechts 11,9% van personen op deze waarde werk vindt dankzij het programma, maar dat 37,3% van de personen met een dergelijke V werkloos worden door het programma.

Als mogelijke verklaring voor het negatieve effect van de opleiding, wordt aangehaald dat deelname mogelijk een sterk negatief signaal geeft aan potentiële werkgevers: 'dit is een persoon die een grote kans maakt op langdurige periodes van afwezigheid door ziekte'.

De uitbreiding naar externe validiteit wordt in het besproken artikel niet gemaakt. Ook bij zo een operatie wordt het MTE als uitgangspuntgenomen, maar zullen in het algemeen nog extra veronderstellingen moeten worden gemaakt, zoals bvb. functionele-vorm-veronderstellingen die extrapolatie toelaten daar waar de support van de historische data niet (volledig) overeenstemt met die voor een nieuwe, andere populatie (voor concrete voorwaarden zie (Heckman and Vytlacil 2005)). Voor een uitbreiding naar het voorspellen op basis van historische informatie, van de effecten van een interventie die nog nooit werd uitgevoerd, is essentieel dat naast de geobserveerde kenmerken X en niet geobserveerde kenmerken U en V ook een nieuwe klasse van kenmerken mee wordt gemodelleerd, met name een klasse C die bestaat uit karakteristieken van een behandeling. In het geval van vb. een opleiding kan men het globaal gegeven 'opleiding' dan zien als een verzameling van kenmerken zoals duur, intensiteit, soort etc. (zie vroeger, noteer dat op deze manier de zwarte doos wordt opengemaakt). Een nieuwe, nog nooit eerder gedane interventie wordt dan gezien als een andere, nieuwe combinatie van karakteristieken. De veralgemeenbaarheid zal uiteraard op een grens botsen van zodra de nooit eerder bestudeerde interventie een kenmerk bezit dat geen enkel van de vroeger bestudeerde interventies bezat.

5.3 Algemeen evenwichtseffecten en andere nuisance-factoren

Bij de potentiële effectenbenadering wordt meestal (impliciet) uitgegaan van de veronderstelling dat is voldaan aan de zogenaamde SUTVA: 'Stable Unit Treatment Value Assumption'. Hier onder begrijpt men dat de behandeling van persoon i enkel en alleen een invloed heeft op de uitkomst van persoon i. Zo worden dus sociale interacties en algemene evenwichtseffecten uitgesloten.

Deze veronderstelling is eerder onrealistisch. Ze onderstelt o.m. dat de effecten van het programma hetzelfde zijn, ongeacht of het programma klein of heel groot is. Als er op een gegeven moment een knelpunt is rond lassers, zal een lasopleiding voor 100 werklozen mogelijk tot een goede uitkomst leiden. Een zelfde opleiding voor 10 000 werklozen had wellicht tot een andere gemiddelde uitkomst geleid. De veronderstelling gaat er ook van uit dat het programma enkel effecten heeft op deelnemers. Nochtans kan een programma sterke impact hebben op niet-deelnemers. Het jongerenbanenplan uit het laatste decennium van de vorige eeuw gaf aan de werkgever een (tijdelijke en degressieve) loonkostensubsidie bij aanwerving van een langdurig

werkloze jonger dan 26 jaar. Dit kan dan leiden tot substitutie-effecten (de langdurig werklozen van 27 jaar en ouder zagen hun kans op tewerkstelling verkleinen door de maatregel), tot verdringings-effecten (niet voor subsidie in aanmerking komende werknemers worden ontslagen en vervangen door wel in aanmerking komende personen). Aangezien een maatregel ook een zekere kostprijs heeft, is er een indirect effect op niet-deelnemers via de belastingen om die worden geheven om de kostprijs te kunnen betalen. Die hogere belastingen kunnen dan weer het gedrag beïnvloeden van niet-deelnemers. Voor een individuele maatregel is dit mogelijk verwaarloosbaar, maar macro-economisch is dit niet langer het geval. Zo zal de kostprijs van het activerend arbeidsmarktbeleid in zijn geheel mogelijk leiden tot een daling van de werkgelegenheid omdat een verhoogde belasting op arbeid leidt tot een reductie in het arbeidsaanbod.

Andere veronderstellingen (zie (Heckman 2005)) zijn o.m. dat de uitkomst van een behandeling hetzelfde is, ongeacht de wijze waarop de toewijzing gebeurt. Dit sluit o.m. vertekening t.g.v. randomisering uit.

Er kan besloten worden dat met de potentiële uitkomstenbenadering er, zeker bij wat grotere programma's, maar één kant van de effecten van het programma wordt belicht. In bepaalde omstandigheden is het zelfs mogelijk dat de boven opgesomde neveneffecten er voor zorgen dat de 'geen-programma-uitkomst' Y_0 van alle niet-deelnemers wordt beïnvloed. In dat geval wordt de potentiële uitkomstenbenadering zelf invalide.

6. Voorlopige conclusie

We eindigen met een 'voorlopige' conclusie. Voorlopig, omdat de meting van de effectiviteit van interventies een onderzoeksdomein is dat nog volop in ontwikkeling is, zoals de tekst ook heeft proberen aan te tonen. We sommen de elementen op die voor ons komen boven drijven na het maken van het bovenstaand overzicht.

- Er is niet zoiets als het ideale onderzoeksdesign met een manier van aanpakken die men in alle omstandigheden moet trachten te repliceren. Het is veeleer belangrijk om te vertrekken van de beschikbare data (wijze van inzameling, aantal observaties, aantal meetmomenten, rijkdom, etc.), en dan vervolgens in functie van de kenmerken van de data het meest geschikte design te selecteren.

- Dit is overigens geen pleidooi voor een passief-receptief toepassen van het meest geschikte onderzoeksdesign, gegeven de beschikbare data. Wel integendeel, reeds bij de conceptie van nieuw arbeidsmarktbeleid, dus vooraleer nieuwe maatregelen worden ingevoerd, zou er ook telkens op beleidsniveau moeten worden nagedacht welk soort van data het meest geschikt zijn om de effectiviteit van de nieuwe maatregel te bestuderen, en zou vervolgens ook het nodig moeten in het werk gesteld worden opdat dit soort van data ook zou kunnen worden ingezameld, vanaf dag 1.

- Een eenvoudige maar fundamentele vaststelling wordt gemaakt door Heckman e.a. 1999, na het bestuderen van eenzelfde programma waarvan de impact zowel op experimentele als niet-experimentele wijze werd gemeten. Gesteld dat de gebruikte data voldoen aan een bepaalde minimale kwaliteit, is de vaststelling met name dat selectievertekening verantwoordelijk is voor slechts een beperkt stuk van het verschil tussen het experimentele en het niet-experimentele resultaat. "A far more important bias arises from comparing non-comparable people" (p. 2 082). Drie belangrijke bronnen die tot niet-vergelijkbaarheid leiden, worden geciteerd: (1) het gebruik van data uit verschillende surveys of verschillende bronnen; (2) deelnemers en niet-deelnemers

komen uit een andere lokale arbeidsmarkt; (3) de match op basis van persoonlijke kenmerken is niet goed.

Dit zijn alvast vaststellingen die tot duidelijke aanbevelingen leiden: gebruik voor deelnemers en niet-deelnemers data uit dezelfde bron, controleer goed de lokale arbeidsmarktstatus van deelnemers en niet-deelnemers, en zorg dat de support voor beide groepen identiek is.

Bibliografie

Aakvik, A., Heckman, J.J. & Vytlacil, E.J. 2005. Estimating treatment effects for discrete outcomes when responses to treatment vary: an application to Norwegian vocational rehabilitation programs. ***Journal of Econometrics 125***:15-51.

Abbring, J.H. & van den Berg, G.J. 2003. The Nonparametric Identification of Treatment Effects in Duration Models. ***Econometrica 71***:1 491-1 517.

Aho, S. 2005. Measuring impact of ALPM participation on later employment - how to deal with the problem of universal participation and evaluation of targeting? Case a matched control group method application into Finnish register data. xerox.

Angrist J. 1990. Lifetime earnings and the vietnam era draft lottery: evidence from social security administrative records. ***American Economic Review 80***:313-336.

Barnow, B., Cain, G. & Goldberger, A.1980. Issues in the analysis of Selectivity Bias, pp. 290-317 In E. Stromsdorfer & G. Farkas [eds.], ***Evaluation Studies*** Vol. 5. Sage.

Bell, B., Blundell, R. & Van Reenen, J. 1999. Getting the unemployed back to work: The role of targeted wage subsidies. ***International Tax and Public Finance 6***:339-360.

Blundell, R. & Costa Dias, M. 2002. Alternative Approaches to Evaluation in empirical Microeconomics. ***Portuguese Economic Journal 1***:91-115.

Card, D. & Krueger, A. 1997. Myth and Measurement: The New Economics of the Minimum Wage.

Centeno, L., Centeno, M. & Novo, A. 2005. Evaluating the impact on wages and unemployment duration of a mandatory job search program. xerox.

Cockx, B. & Dejemeppe, M. 2007. Is the Notification of Monitoring a Threat to the Unemployed? A Regression Discontinuity Approach. EZA Discussion Paper No. 2854.

Heckman, J.J. 1979. Sample selection as a specification error. ECONOMETRICA 47:153-161.

Heckman, J.J. 2005. The Scientific Model of Causality. xerox.

Heckman, J.J. 1997. Instrumental Variables: A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations. ***Journal of Human Resources 32***:441.

Heckman, J.J., LaLonde, R.J. & Smith, J.A. 1999. The Economics and Econometrics of Active Labor Market Programs. ***Handbook of labor economics.*** Volume 3A. 19991865.

Heckman, J.J. & Robb, R. Jr. 1985. Alternative Methods for Evaluating the Impact of Interventions: An Overview. ***Journal of Econometrics 30***:239.

Heckman, J.J. & Smith, J.A. 1995. Assessing the Case for Social Experiments. ***Journal of Economic Perspectives 9***:85.

Heckman, J.J. & Vytlacil, E. 2005. Structural Equations, Treatment Effects, and Econometric Policy Evaluation. ***Econometrica 73***:669.

Heylen, V. & Bollens, J. 2006. Stromen tussen werkloosheid, werk en OCMW.

Holland, P. 1986. Statistics and causal inference. ***Journal of the American Statistical Association 81***:945-960.

Imbens, G.W. & Angrist, J. 1994. Identification and estimation of local average treatment effects. ***Econometrica 62***:467-475.

Keane, M. 2006. Structural vs. Atheoretic Approaches to Econometrics. xerox.

Kluve, J. 2006. The Effectiveness of European Active Labor Market Policy. IZA Discussion Paper No. 2018.

Lechner, M. 2002. Some practical issues in the evaluation of heterogeneous labour market programmes by matching methods. ***Journal-of-the-Royal-Statistical-Society 165***(1).

Lechner, M. 2004. Sequential Matching Estimation of Dynamic Causal Models. IZA Discussion Paper No. 1042.

Rosenbaum, P. & Rubin, D. 1983. The Central Role of the Propensity Score in Observational Studies for Causal Effects. ***Biometrika 70***:41-50.

Rosenzweig, M.R. & Wolpin, K.I. 2000. Natural 'Natural Experiments' in Economics. ***Journal of Economic Literature 38***:827.

Sianesi, B. 2004. An evaluation of active labour market programmes in Sweden. ***Review of Economics and Statistics 86***:133-155.

Smith, J. 2000. A Critical Survey of Empirical Methods for Evaluating Active Labor Market Policies. ***Schweizerische Zeitschrift für Volkswirtschaft und Statistik 136***:1-22.

Smith, J.A. & Todd, P.E. 2005. Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators? ***Journal of Econometrics 125***:305.

Appendix: Voorwaardelijke verwachtingen

De operator E(X) staat voor de verwachte waarde (of de 'verwachting') van X, dit is het statistisch equivalent van wat in een steekproef het steekproefgemiddelde is: de som van alle waarden x die voorkomen, gedeeld door het totaal aantal voorkomende waarden. Het is dikwijls zinvol om een verwachting te definiëren voor subgroepen. Men krijgt dan een voorwaardelijke verwachting. Stel dat de gemiddelde lengte van mannen verschilt van de gemiddelde lengte van vrouwen, dan kan gesteld worden dat de verwachting van de lengte, gegeven dat men ze alleen bij mannen meet, zal verschillen van de verwachting van de lengte, gegeven dat men ze alleen bij vrouwen meet. Als we de naam van een variabele in hoofdletters schrijven, en een bepaalde waarde van die variabele in kleine letters, wordt er dan geschreven: E(LENGTE│GESLACHT=man) ≠ E(LENGTE│GESLACHT=vrouw) ≠ E(LENGTE), waarbij de verticale streep kan gelezen worden als 'als men rekening houdt met', of 'gegeven dat', of 'voorwaardelijk op het feit dat' of 'conditionerend op'. In de gegeven voorbeelden wordt geconditioneerd op één waarde (de waarde 'man' of de waarde 'vrouw'). Men kan echter ook conditioneren op een volledige toevalsvariabele (of een vector van verschillende toevalsvariabelen), vb. E(LENGTE│LEEFTIJD), in het voorbeeld zoeken we dan de verwachte waarde van de lengte, rekening houdend met de leeftijd. Aangezien LEEFTIJD een toevalsvariabele is die verschillende waarden aanneemt, zal de verwachting niet langer, zoals in de bovenstaande voorbeelden, één getal zijn, maar zelf een toevalsvariabele worden met evenveel waarden als de variabele leeftijd (waarbij LEEFTIJD kan gemeten zijn in discrete stappen

(1 jaar, 2 jaar...) of als continue variabele). De betekenis van E(LENGTE│LEEFTIJD,GESLACHT=man) volgt logisch uit het voorgaande.

Tot slot, als twee toevalsvariabelen onderling onafhankelijk zijn, en vb. de lengte geen relatie heeft met het IQ, dan geldt dat E(LENGTE│IQ) = E(LENGTE), d.w.z., dan maakt het niet uit of men 'conditioneert' op het IQ, dus kan men het perfect laten vallen.