F.S. Kool

Een Statistische Analyse van Recidive Cijfers

Bachelorscriptie

Scriptiebegeleiders:
Prof.dr. A.W. van der Vaart & S.L. van der Pas, MSc MA

Datum Bachelorexamen: 3 juli 2014

Mathematisch Instituut, Universiteit Leiden

Inhoudsopgave

Inleiding

Een samengesteld dialoog uit een debat in 2011 over het voorkomen van recidive tussen Lillian Helder (PVV), Jeroen Recourt (PvdA) en Sharon Gesthuizen (SP):

> Meneer Recourt: *"We hebben al wat onderzoeken gehoord en de één is wat succesvoller dan de ander, maar ze geven allemaal als uitkomst dat voor het voorkomen van recidive de taakstraf beter werkt dan een gevangenisstraf."*
>
> Mevrouw Helder: *"Ik vind het een beetje appels met peren vergelijken. Niet appels met koeien, zover wil ik niet gaan, maar toch wel appels met peren. Niet ieder persoon is hetzelfde. Recidive slaat terug op de persoon zelf en iemand die een taakstraf opgelegd heeft gekregen en recidiveert is wel iemand anders dan iemand die een vrijheidsstraf opgelegd heeft gekregen en recidiveert. Diegene heeft een vrijheidsstraf ondergaan en geen taakstraf. Hoe moet ik die twee nu met elkaar vergelijken?"*
>
> [...] Mevrouw Gesthuizen: *"Is hiermee dan ook gezegd dat de PVV nooit gelooft in enige vorm van statistisch onderzoek? Want dat doen we namelijk met statistisch onderzoek op allerlei terreinen. Je vergelijkt altijd groepen mensen met andere groepen mensen, anders is het niet meer te doen. Gelooft mevrouw Helder niet in statistisch onderzoek?"*
>
> Mevrouw Helder: *"Persoon A is niet persoon B. Ik kan iemand met een gevangenisstraf en iemand met een taakstraf niet vergelijken."*
>
> Mevrouw Gesthuizen: *"Ik vind het echt een kolderredenering en ik vind het heel erg verdrietig dat ik op deze manier moet debatteren."*

In bovenstaand dialoog probeert mevrouw Helder een beschrijving te geven van een veel voorkomend probleem in de zogenoemde observationele studies. Hierbij lijkt mevrouw Helder de enige van de drie politici te zijn die zich bewust is van de moeilijkheden die kunnen ontstaan wanneer we groepen mensen in een experiment willen vergelijken, terwijl zij eigenlijk niet vergelijkbaar zijn. Zij wijst hier indirect op het feit dat we niet altijd te maken hebben met een gerandomiseerd experiment, waarbij deze problemen niet optreden. In deze scriptie geven we antwoord op de vraag van mevrouw Helder hoe we personen kunnen vergelijken, terwijl de één een gevangenisstraf heeft ondergaan en de ander een werkstraf.

Om deze situatie nader te bekijken, zoomen we in op het artikel 'Recidive na werkstraffen en na gevangenisstraffen' van de auteurs Wermink et al. (2009). In dit artikel wordt geprobeerd antwoord te geven op de vraag of werkstraffen een goed alternatief zijn voor gevangenisstraffen in relatie tot de recidive van de gestraften na afloop van hun straf. Door middel van de propensity score methode wordt geconcludeerd dat werkstraffen voor minder recidive zorgen in vergelijking tot gevangenisstraffen.

In Hoofdstuk 1 beschouwen we het probleem dat mevrouw Helder probeert te formuleren, waarbij de propensity score methode als mogelijke oplossing wordt beschreven. Hoofdstuk 1 zal daarmee de kern van deze scriptie weergeven. In Hoofdstuk 2 wordt een uiteenzetting gegeven over logistische regressie wat gebruikt wordt in de methode uit Hoofdstuk 1 en vormt daarmee een aanvulling op het eerste hoofdstuk.

De bovenstaande dialoog is niet de enige ophef die omtrent dit onderwerp is ontstaan. Zo is er specifiek over het artikel dat we bestuderen een discussiestuk geschreven, waar de auteurs vervolgens een weerwoord op hebben gegeven. In deze scriptie bekijken en beoordelen we dit discussiestuk van Groenendijk en van Delft (2013a) en het weerwoord van de auteurs, deels aan

de hand van een opgezette simulatie. Hoofdstuk 3 geeft instructief de opgezette simulatie weer, zodat in Hoofdstuk 4 een aantal van de discussiepunten kan worden besproken.

In de conclusie wordt een mening gevormd over het uitgevoerde onderzoek naar het effect van een werkstraf op recidive aan de hand van de bestudeerde methode en de geschreven discussie daarop. Er wordt geconcludeerd dat er in de strijd over dit gevoerde onderzoek geen echte winnaar of verliezer is aan te wijzen. Het onderzoek van Wermink et al. toont een aantal tekortkomingen, maar het bestudeerde discussiestuk van Groenendijk en van Delft geeft niet direct voldoende aanleiding om het gevoerde onderzoek van tafel te vegen. Daarnaast wordt benadrukt dat een enkel onderzoek niet voldoende bewijs levert om de bevindingen toe te passen in de praktijk.

1 Propensity score matching

1.1 Inleiding

De vraag die Wermink et al. in het artikel 'Recidive na werkstraffen en na gevangenisstraffen' (hierna gerefereerd als Artikel I) proberen te beantwoorden, is in hoeverre werkstraffen een goed alternatief zijn voor gevangenisstraffen in relatie tot de recidive van de gestraften na afloop van hun straf. Om dit te bepalen wordt het gemiddeld aantal recidive vergeleken van veroordeelden tot een gevangenisstraf met veroordeelden tot een werkstraf. Wanneer men een directe vergelijking maakt zonder rekening te houden met een selecte toedeling tot de straffen, kan het gemeten verschil in recidive te wijten zijn aan een verzameling van verstorende variabelen. Om de invloed van dergelijke variabelen uit te schakelen is het wenselijk om een gerandomiseerd experiment met een controlegroep uit te voeren. Hierbij is er sprake van een willekeurige toewijzing van deelnemers aan ofwel een controlegroep ofwel een experimentele groep. Echter, dit is niet direct mogelijk aan de hand van de observationele data die is verkregen. Het is voor de hand liggend dat individuen die een werkstraf of gevangenisstraf hebben gekregen niet vergelijkbaar zijn, waardoor de toedeling tot één van de straffen wordt bepaald door factoren buiten onze controle. Dit probleem treedt op bij observationele studies, waarbij sprake is van ongewenste selectie. Er zijn methoden om voor deze selectie te controleren. In Artikel I wordt gekozen voor propensity score matching.

1.2 Een structuur voor de bepaling van causale effecten

Deze paragraaf bekijkt enkele definities, probleemstellingen en oplossingen in de zoektocht naar het omschrijven van causale effecten. In deze studie is er sprake van een causaal effect wanneer het verkrijgen van een werkstraf noodzakelijkerwijs wordt gevolgd door bijvoorbeeld een vermindering in het aantal keer recidiveren. De beoordeling van een gevonden verband in termen van causaliteit is van belang wanneer het uiteindelijke doel van een onderzoek is om de verkregen resultaten toe te passen in de praktijk. Men moet zich ervan bewust zijn dat de uitkomst uit een studie zoals hier beschreven niet direct een causaal verband hoeft te impliceren, maar dat hier meer onderzoek voor nodig is.[1]

Om mogelijke causale effecten te kunnen bepalen van een werkstraf op recidive leiden we in deze paragraaf bruikbare uitdrukkingen af. In paragraaf 1.2.1 worden twee uitdrukkingen gegeven om het effect van een werkstraf op recidive te bepalen. We bekijken de moeilijkheden bij de realisatie van deze uitdrukkingen, welke ontstaan nu er geen sprake is van een gerandomiseerd experiment. In paragraaf 1.2.2 zien we dat conditionering op de verstorende variabelen, de variabelen die ervoor zorgen dat we geen gerandomiseerd experiment hebben, hier een oplossing voor geeft. Echter, de toepassing van deze conditionering is in de praktijk vrijwel onmogelijk. Paragraaf 1.2.3 introduceert daartoe de propensity score, een functie van de verstorende variabelen. De conditionering op de propensity score, in plaats van de conditionering op alle variabelen, geeft de uiteindelijke oplossing van het probleem, zodat het causale effect kan worden bepaald.

Het mag benadrukt worden dat de toepassing hier ligt in de onderzoeksrichting van Artikel I, maar dat dit kader wel degelijk een universele methode betreft.

[1] Om meer te lezen over over het causaliteitsbegrip en het herkennen van een causaal verband wordt de lezer verwezen naar bijvoorbeeld Bijma et al. (2013) of Wasserman (2004).

1.2.1 Gemiddelde causale effecten

Laat Z de variabele zijn die aangeeft welk van de twee straffen is toegewezen aan een individu

$$Z = \begin{cases} 0 & \text{als gevangenisstraf,} \\ 1 & \text{als werkstraf.} \end{cases} \tag{1.1}$$

Wanneer Z de waarde nul aanneemt, zeggen we dat een dader zich in de *controlegroep* bevindt. In het andere geval is de dader toegewezen aan de *experimentele groep*. In het algemeen gaat men ervan uit dat de controlegroep geen 'behandeling' of 'de behandeling' zoals gebruikelijk ondergaat. De experimentele groep krijgt de nieuwe behandeling, waarvan we het gewenste effect willen bepalen. In deze situatie beschouwen we een gevangenisstraf dus als de gebruikelijke straf voor veroordeelden, waarbij een werkstraf als een nieuwe mogelijkheid kan worden beschouwd.

Definieer de waargenomen recidive Y als

$$Y = \begin{cases} Y_0 & \text{als } Z = 0, \\ Y_1 & \text{als } Z = 1, \end{cases} \tag{1.2}$$

waarbij Y_i het aantal keer recidiveren van de daders na de veroordeling aangeeft met $Y_i \in \mathbb{N}_0$.

Het is voor de hand liggend om het effect van een werkstraf op recidive te meten door voor elk individu het volgende verschil te bepalen:

$$Y_1 - Y_0.$$

Dit verschil is echter niet meetbaar, omdat voor elk individu slechts één van beide variabelen wordt waargenomen. Een veel gebruikte oplossing is om op zoek te gaan naar het gemiddelde verschil in gemeten recidive van de veroordeelden. In de literatuur wordt veelal onderscheid gemaakt tussen het gemiddelde effect van de werkstraf op alle daders (τ) en het gemiddeld effect van de werkstraf op enkel de werkgestraften (τ_e).[2] Voor τ gebruiken we de volgende uitdrukking:

$$\tau = \mathbb{E}[Y_1] - \mathbb{E}[Y_0]. \tag{1.3}$$

Hiervan onderscheiden we τ_e als volgt

$$\tau_e = \mathbb{E}[Y_1|Z = 1] - \mathbb{E}[Y_0|Z = 1]. \tag{1.4}$$

De keuze of τ ofwel τ_e moet worden bepaald, kan afhankelijk zijn van het uitgangspunt van het onderzoek. Denk bijvoorbeeld aan de bepaling van het effect van abortus op lichamelijke klachten van vrouwen. De interpretatie voor τ is dan als volgt: τ is het gemiddelde van wanneer alle zwangere vrouwen een abortus laten plegen ($Z = 1$) min het gemiddelde wanneer geen van de zwangere vrouwen abortus heeft laten plegen ($Z = 0$). Het is meer voor de hand liggend om een bepaling te doen van τ_e, omdat het interessanter lijkt te zijn wat het effect van abortus is op vrouwen die daadwerkelijk abortus hebben gepleegd. Omdat het plegen van abortus een doelbewuste keuze is, kunnen we aannemen dat het voor vrouwen die de baby houden niet van belang is om te weten wat de effecten van abortus geweest zouden zijn. We conditioneren daarom op $Z = 1$ in de uitdrukking voor τ_e.

[2] De uitdrukkingen τ en τ_e worden ook wel het 'average treatment effect' en het 'average treatment effect on the treated' genoemd, respectievelijk. De geïntereseerde lezer wordt verwezen naar Austin (2011) of Caliendo en Kopeinig (2005).

Bij de bepaling van τ en τ_e treedt er een soortgelijk probleem op. In vergelijking (1.4) is de moeilijkheid echter eerder zichtbaar. Gezien dat de term $\mathbb{E}[Y_0|Z=1]$ het verwachte aantal recidive van een werkgestrafte is, terwijl hij of zij een gevangenisstraf heeft gehad, is dit een waarde die we niet kunnen observeren. Het doel is daarom een waarde voor $\mathbb{E}[Y_0|Z=1]$ te substitueren die als realistisch kan worden beschouwd. Bij de bepaling van τ stuiten we op hetzelfde, omdat we hierbij niet alleen voor de hele experimentele groep willen weten hoeveel er was gerecidiveerd als er een gevangenisstraf had plaatsgevonden, maar ook andersom.

Om een oplossing te vinden voor de substitutie van termen die niet kunnen worden waargenomen beschouwen we eerst het geval waarin elk individu willekeurig wordt toegewezen aan één van de straffen, zoals in een gerandomiseerd experiment. In deze situatie is het aannemelijk dat de variabele Z onafhankelijk is van het paar (Y_0, Y_1), omdat er voor elk individu met behulp van een zuivere muntworp bepaald kan worden welke Y_i zal plaatsvinden. Met behulp van deze eigenschap geldt voor de verwachting van recidive

$$\begin{aligned}\mathbb{E}[Y_i] &= \mathbb{E}[Y_i|Z=i] \\ &\overset{(1.2)}{=} \mathbb{E}[Y|Z=i].\end{aligned}$$

Er volgt nu met behulp van het bovenstaande dat vergelijking (1.3) gelijk is aan

$$\mathbb{E}[Y_1] - \mathbb{E}[Y_0] = \mathbb{E}[Y|Z=1] - \mathbb{E}[Y|Z=0]$$

en vergelijking (1.4) voldoet aan

$$\mathbb{E}[Y_1|Z=1] - \mathbb{E}[Y_0|Z=1] = \mathbb{E}[Y|Z=1] - \mathbb{E}[Y|Z=0].$$

We zien dat in dit geval τ gelijk is aan τ_e. Door de willekeurige toewijzing tot één van beide groepen ontstaan er geen systematische verschillen tussen de groepen. Daardoor verwachten we geen verschil in het effect als we alle deelnemers bekijken of slechts een deel hiervan. De bevinding dat τ gelijk is aan τ_e loopt dus in één lijn met onze verwachting.

Echter, de gedachte van aselecte toewijzing tot één van beide straffen is niet realistisch. Het is goed mogelijk dat daders die een werkstraf hebben gekregen niet vergelijkbaar zijn met de gevangenisgestraften. In deze situatie kan gedacht worden aan een onzuivere muntworp die zal bepalen welk van de twee straffen wordt toegewezen. Het is daardoor mogelijk dat er van te voren al meer waarschijnlijkheid is dat een gevangenisgestrafte zal recidiveren.[3] De onafhankelijkheidsaanname tussen Z en het paar (Y_0, Y_1) kan daardoor niet meer worden toegepast.

1.2.2 Observationele studies en verstorende variabelen

Bij een directe vergelijking, zoals we hierboven hebben beschreven, kan het gemeten verschil in recidive te wijten zijn aan een aantal verstorende variabelen. Deze variabelen zijn verstorend in de zin dat ze de keuze voor een bepaalde straftoewijzing hebben beïnvloed, waardoor er systematische verschillen tussen de groepen kunnen ontstaan. De verstorende variabelen worden *covariaten* genoemd. De vector van p verschillende covariaten noteren we als $X = (X_1, \ldots, X_p)$.

Het doel is om een geschikte uitdrukking te vinden voor τ of τ_e, waarbij de moeilijkheid ligt bij een term die niet waarneembaar is, zoals bovenaan deze pagina staat beschreven. Laten

[3] In Artikel I wordt vermeld dat bij een directe vergelijking gevangenisgestraften al op voorhand een hoger risico hebben op herhaald crimineel gedrag dan werkgestraften, doordat de rechter bij de toewijzing bijvoorbeeld rekening houdt met de ernst van de criminele activiteiten en/of dit de eerste veroordeling is.

we een geschikte substitutie zoeken voor de term $\mathbb{E}[Y_0|Z = 1]$. We hebben gezien dat het in een gerandomiseerd experiment volstaat om hiervoor de verwachte uitkomst van de daders uit de gevangenisgestrafte groep te nemen. Ondanks dat de willekeurige toewijzing niet meer van toepassing is, ligt hier toch de oplossing. Het idee is dat wanneer de beslissing over de straf willekeurig is voor individuen met gelijke waarden voor de covariaten, de gemiddelde uitkomst van vergelijkbare individuen kan worden genomen die geen werkstraf hebben gehad.

In Rosenbaum en Rubin (1983) worden twee aannames voorgelegd om deze oplossing te kunnen bereiken.

1. *Conditionele onafhankelijkheid*
$$(Y_i \perp Z)|X.$$
Gegeven de covariaten X is de uitkomst, recidive, onafhankelijk van de straftoewijzing Z. Deze sterke aanname gaat enkel volledig op wanneer de vector X de juiste covariaten bevat. Over de keuze van de covariaten is meer te lezen in paragraaf 2.4.

2. *Common support*
$$0 < P(Z = 1|X = x) < 1 \qquad \forall x.$$
Dit is een voor de hand liggende aanname, welke aangeeft dat voor ieder individu met covariaten X er zowel een kans op gevangenisstraf als een kans op werkstraf moet zijn. Veronderstel dat voor een zekere dader geldt dat $P(Z = 1|X = x) = 1$ dan is het niet realistisch om dit individu te vergelijken met een gevangenisgestrafte dader.

Laten we nu met bovenstaande aannames in gedachte naar de verwachting van recidive gegeven de covariaten kijken, ofwel

$$\begin{aligned}\mathbb{E}[Y_i|X] &= \sum_y yP(Y_i = y|X = x) \\ &= \sum_y y\frac{P(Y_i = y|X = x)P(Z = i|X = x)}{P(Z = i|X = x)}.\end{aligned}$$

Met behulp van de conditionele onafhankelijkheidsaanname volgt dat deze uitdrukking gelijk is aan

$$\mathbb{E}[Y_i|X] = \sum_y y\frac{P(Y_i = y, Z = i|X = x)}{P(Z = i|X = x)}. \tag{1.5}$$

Voor gebeurtenissen A, B en C kunnen we schrijven:

$$P(A|B, C) = \frac{P(A, B|C)}{P(B|C)}. \tag{1.6}$$

Een bewijs van deze regel is te vinden in Appendix B. Vergelijking (1.6) kunnen we gebruiken om vergelijking (1.5) te vereenvoudigen tot het volgende

$$\begin{aligned}\mathbb{E}[Y_i|X] &= \sum_y yP(Y_i = y|Z = i, X = x) \\ &= \mathbb{E}[Y_i|Z = i, X = x] \\ &= \mathbb{E}[Y|Z = i, X = x].\end{aligned}$$

De laatste gelijkheid verkrijgen we met behulp van de definite van recidive Y in (1.2). De verkregen uitdrukking voor de conditionele verwachting is intuïtief wat we verwachten. Door de onafhankelijkheid tussen de straftoewijzing en de uitkomst recidive, gegeven de covariaten, kunnen we conditioneren op Z, omdat dit geen extra informatie toevoegt. Gezien het feit dat we op zoek zijn naar τ of τ_e is bovenstaande slechts een hulpmiddel. Er geldt: (Rice (2007), p.148)

$$\mathbb{E}[Y_i] = \mathbb{E}_X[\mathbb{E}(Y_i|X)]. \tag{1.7}$$

Dus met behulp van (1.7) hebben we voor τ een nieuwe uitdrukking gevonden.

$$\begin{aligned}
\tau &= \mathbb{E}[Y_1] - \mathbb{E}[Y_0] \\
&= \mathbb{E}_X[\mathbb{E}(Y_1|X)] - \mathbb{E}_X[\mathbb{E}(Y_0|X)] \\
&= \int \mathbb{E}[Y|Z=1, X=x]p_X(x)dx - \int \mathbb{E}[Y|Z=0, X=x]p_X(x)dx \\
&= \int (\mathbb{E}[Y|Z=1, X=x] - \mathbb{E}[Y|Z=0, X=x])p_X(x)dx.
\end{aligned}$$

Het verschil met het bepalen van τ_e is dat we niet de verwachting moeten nemen over de verdeling van X van alle daders samen, maar over de verdeling van X slechts in de experimentele groep. Voor de afleiding van een uitdrukking voor τ_e gebruiken we daarom de volgende vergelijking

$$\mathbb{E}[Y_i|Z=1] = \mathbb{E}_{X|Z=1}[\mathbb{E}(Y_i|X)],$$

waardoor voor τ_e de volgende uitdrukking kan worden gebruikt.

$$\begin{aligned}
\tau_e &= \mathbb{E}[Y_1|Z=1] - \mathbb{E}[Y_0|Z=1] \\
&= \mathbb{E}_{X|Z=1}[\mathbb{E}(Y_1|X)] - \mathbb{E}_{X|Z=1}[\mathbb{E}(Y_0|X)] \\
&= \int (\mathbb{E}[Y|Z=1, X=x] - \mathbb{E}[Y|Z=0, X=x])p_{X|Z}(x|1)dx.
\end{aligned}$$

Merk op dat wanneer het aantal covariaten stijgt het goed mogelijk is dat in de twee groepen geen individuen kunnen worden gevonden met gelijke waarden voor alle covariaten. Het is bijvoorbeeld vrijwel onmogelijk om voor een werkgestrafte een gevangenisgestrafte te vinden met precies dezelfde leeftijd, geslacht, criminele geschiedenis enzovoort. Wanneer het aantal covariaten stijgt kunnen de bovenstaande uitdrukkingen voor τ en τ_e dus niet of niet eenvoudig gerealiseerd worden. De volgende paragraaf geeft een oplossing voor dit probleem.

1.2.3 De propensity score

In Rosenbaum en Rubin (1983) wordt een oplossing gegeven voor het probleem wanneer de conditionering op een gehele vector van covariaten niet meer te realiseren is. In dit artikel wordt laten zien dat we in plaats van conditionering op een vector van covariaten kunnen conditioneren op een functie van de geobserveerde covariaten, de propensity score. Rosenbaum en Rubin definiëren daartoe eerst een balancing score en laten ons kennismaken met een aantal aangename eigenschappen.

Definitie 1.1. (balancing score) Een *balancing score* $b(x)$ is een functie van geobserveerde covariaten x zodanig dat de conditionele verdeling van x gegeven $b(x)$ gelijk is voor de experimentele groep en de controlegroep.

Vervolgens laten Rosenbaum en Rubin zien dat de grofste balancing score de *propensity score* is en definiëren deze als volgt.

Definitie 1.2. (propensity score) Zij de *propensity score* $\pi(x)$ de kans om toegedeeld te worden tot de experimentele groep gegeven de geobserveerde covariaten x

$$\pi(x) = P(Z = 1|X = x).$$

Een mogelijkheid is om de propensity score te schatten uit de data met behulp van een logistisch regressiemodel. Dit model wordt verder toegelicht in Hoofdstuk 2.

We bekijken nu twee stellingen waarmee we kunnen we laten zien dat het voldoende is om op een balancing score, in het bijzonder de propensity score, te conditioneren.

Stelling 1.3. [Rosenbaum en Rubin (1983), stelling 1] De toewijzing tot één van beide groepen en de geobserveerde covariaten zijn conditioneel onafhankelijk gegeven de propensity score, ofwel

$$(X \perp Z)|\pi(X).$$

Bewijs. Om de conditionele onafhankelijkheid tussen x en Z gegeven de propensity score aan te tonen, willen we laten zien dat er aan de volgende vergelijking wordt voldaan

$$P(X = x, Z = z|\pi(X) = \pi(x)) = P(X = x|\pi(X) = \pi(x))P(Z = z|\pi(X) = \pi(x)). \tag{1.8}$$

Met behulp van (1.6) kunnen we de linkerzijde van vergelijking (1.8) schrijven als

$$P(Z = z|X = x, \pi(X) = \pi(x))P(X = x|\pi(X) = \pi(x)).$$

Dus nu volgt met behulp van vergelijking (1.8) dat het volstaat het volgende aan te tonen:

$$P(Z = z|X = x, \pi(X) = \pi(x)) = P(Z = z|\pi(X) = \pi(x)). \tag{1.9}$$

Omdat Z enkel de waarde 1 of 0 kan aannemen, is het voldoende om te laten zien dat er geldt:

$$P(Z = 1|X = x, \pi(X) = \pi(x)) = P(Z = 1|\pi(X) = \pi(x)). \tag{1.10}$$

Aangezien de propensity score een functie van de covariaten is, voegt het condtioneren op de propensity score geen extra informatie toe. De linkerzijde van vergelijking (1.10) is dus te schrijven als

$$P(Z = 1|X = x). \tag{1.11}$$

Merk op dat vergelijking (1.11) gelijk is aan de propensity score. Dus met behulp van vergelijking (1.10) volgt dat we moeten laten zien dat er geldt:

$$P(Z = 1|\pi(X) = \pi(x)) = \pi(x).$$

Schrijf daartoe

$$\begin{aligned} P(Z = 1|\pi(X) = \pi(x)) &= 1 \cdot P(Z = 1|\pi(X) = \pi(x)) + 0 \cdot P(Z = 0|\pi(X) = \pi(x)) \\ &= \mathbb{E}[Z|\pi(X) = \pi(x)]. \end{aligned}$$

Om deze uitdrukking te herschrijven, gebruiken we dat voor stochastische variabelen X, Y en Z geldt:

$$\mathbb{E}[X|Y] = \mathbb{E}[\mathbb{E}(X|Y,Z)|Y]. \tag{1.12}$$

Een bewijs van deze regel is te vinden in Appendix B. Aan de hand van vergelijking (1.12) kunnen we schrijven

$$\begin{aligned} P(Z=1|\pi(X)) &= \mathbb{E}[\mathbb{E}[Z|\pi(X),X]|\pi(X)] \\ &= \mathbb{E}[\mathbb{E}[Z|X]|\pi(X)]. \end{aligned}$$

Gezien dat Z een binaire variabele is, volgt

$$\begin{aligned} P(Z=1|\pi(X)) &= \mathbb{E}[P(Z=1|X)|\pi(X)] \\ &= \mathbb{E}[\pi(X)|\pi(X)] \\ &= \pi(X). \end{aligned}$$

□

Aan de hand van stelling 1.3 is nu eenvoudig in te zien dat de conditionele verdeling van x gegeven de balancing score, dus in het bijzonder de propensity score, daadwerkelijk gelijk is voor beide groepen, zoals beschreven in definitie 1.1. Aan de hand van vergelijking (1.6) kunnen we namelijk schrijven

$$P(X=x|Z=1,\pi(X)=\pi(x)) = \frac{P(X=x,Z=1|\pi(X)=\pi(x))}{P(Z=1|\pi(X)=\pi(x))}.$$

Met behulp van stelling 1.3 volgt dan dat bovenstaande gelijk is aan

$$\begin{aligned} P(X=x|Z=1,\pi(X)=\pi(x)) &= \frac{P(X=x|\pi(X)=\pi(x))P(Z=1|\pi(X)=\pi(x))}{P(Z=1|\pi(X)=\pi(x))} \\ &= P(X=x|\pi(X)=\pi(x)) \\ &= \frac{P(X=x|\pi(X)=\pi(x))P(Z=0|\pi(X)=\pi(x))}{P(Z=0|\pi(X)=\pi(x))} \\ &= P(X=x|Z=0,\pi(X)=\pi(x)). \end{aligned}$$

We zullen nu een stelling bewijzen die van groot belang is voor de propensity score methode.

Stelling 1.4. Als aan de conditionele onafhankelijkheidsaanname wordt voldaan dan geldt

$$(Y_i \perp Z)|\pi(X).$$

Bewijs. Het bewijs van deze stelling gaat soortgelijk aan het bewijs van stelling 1.3.
We willen laten zien dat er geldt:

$$P(Y_i=y_i, Z=z|\pi(X)=\pi(x)) = P(Y_i=y_i|\pi(X)=\pi(x))P(Z=z|\pi(X)=\pi(x)).$$

Met behulp van (1.6) kunnen we de linkerkant van bovenstaande vergelijking schrijven als

$$P(Z=z|Y_i=y_i,\pi(X)=\pi(x))P(Y_i=y_i|\pi(X)=\pi(x)).$$

Dus we willen laten zien:

$$P(Z = z|Y_i = y_i, \pi(X) = \pi(x)) = P(Z = z|\pi(X) = \pi(x)).$$

Echter, het is voldoende om te laten zien:

$$P(Z = 1|Y_i = y_i, \pi(X) = \pi(x)) = P(Z = 1|\pi(X) = \pi(x)). \tag{1.13}$$

In het bewijs van stelling 1.3 hebben we gezien dat er geldt

$$P(Z = 1|\pi(X) = \pi(x)) = \pi(x).$$

Dus er volgt nu uit vergelijking 1.13 dat we de volgende gelijkheid moeten aantonen:

$$P(Z = 1|Y_i = y_i, \pi(X) = \pi(x)) = \pi(x).$$

Aan de hand van de definitie van Z, vergelijking (1.12) en het feit dat de propensity score een functie van geobserveerde covariaten is, kunnen we schrijven

$$\begin{aligned} P(Z = 1|Y_i, \pi(X)) &= \mathbb{E}[Z|Y_i, \pi(X)] \\ &= \mathbb{E}[\mathbb{E}[Z|Y_i, \pi(X), X]|Y_i, \pi(X)] \\ &= \mathbb{E}[\mathbb{E}[Z|Y_i, X]|Y_i, \pi(X)] \\ &= \mathbb{E}[P(Z = 1|Y_i, X)|Y_i, \pi(X)] \\ &= \mathbb{E}[P(Z = 1|X)|Y_i, \pi(X)] \\ &= \mathbb{E}[\pi(X)|Y_i, \pi(X)] \\ &= \pi(X). \end{aligned}$$

□

Als aan de common support aanname wordt voldaan volgt direct dat de common support aanname gegeven de propensity score ook geldt:

$$0 < P(Z = 1|\pi(X) = \pi(x)) < 1 \qquad \forall \pi(x). \tag{1.14}$$

We kunnen dankzij stelling 1.4 voor τ en τ_e een gelijke afleiding doen als in paragraaf 1.2.2, door te conditioneren op de functie $\pi(X)$ in plaats van op de vector X. Veronderstel dat we iemand uit de experimentele groep en iemand uit de controlegroep vinden met een exact gelijke propensity score, maar mogelijk andere waardes voor X, dan volgt dat het gemiddelde effect van een werkstraf op recidive gelijk wordt aan het gemiddelde verschil in recidive van de op propensity score gekoppelde paren:

$$\tau = \mathbb{E}[\mathbb{E}(Y|Z = 1, \pi(X)) - \mathbb{E}(Y|Z = 0, \pi(X))]. \tag{1.15}$$

Zo volgt ook dat het gemiddelde effect van een werkstraf op de werkgestrafte gezien recidive gelijk is aan

$$\tau_e = \mathbb{E}[[\mathbb{E}(Y|Z = 1, \pi(X)) - \mathbb{E}(Y|Z = 0, \pi(X))]|Z = 1]. \tag{1.16}$$

Aan de hand van vergelijking (1.14) gaan we ervan uit dat de voorwaardelijke verwachting van Y gegeven de propensity score $\pi(x)$ bestaat voor elke $\pi(x)$. Als de common support aanname niet geldt voor een zekere $\pi(x)$ dan is het niet mogelijk om zowel $\mathbb{E}[Y|Z = 1, \pi(X) = \pi(x)]$ als $\mathbb{E}[Y|Z = 0, \pi(X) = \pi(x)]$ te vinden, omdat er voor deze waarde van $\pi(X)$ slechts een controle- of experimentele individu aanwezig is.

Dus we zien nu dat we τ en τ_e kunnen bepalen met behulp van de propensity score, welke we kunnen schatten. Dit is prettig, want we hoeven dus niet te conditioneren op de hele vector van covariaten X maar enkel op een functie van X, de propensity score.

1.3 Een match strategie

In het voorgaande zijn we ervan uitgegaan dat er voor ieder individu een vergelijkbare dader uit de andere groep kan worden gevonden met exact dezelfde propensity score. Echter, in de praktijk is deze exacte match niet te vinden. Er wordt daarom gezocht naar koppels die zo goed mogelijk vergelijkbaar zijn, wanneer men let op de propensity score. In paragraaf 1.3.1 bekijken we een mogelijk matching algoritme waarmee deze koppels kunnen worden gevonden. In paragraaf 1.3.2 worden mogelijke aanpassingen van de match strategie besproken, welke de strategie eventueel kunnen verbeteren. Paragraaf 1.3.3 bespreekt twee methodes waarmee kan worden beoordeeld of er ondanks het matchen op de propensity score, in plaats van het matchen op alle covariaten, twee groepen zijn verkregen waar geen systematische verschillen meer tussen zitten.

1.3.1 Nearest Neighbour matching

Er kunnen verschillende keuzes worden gemaakt voor een matching algoritme. In Artikel I wordt gekozen voor *Nearest Neighbour matching*, waarbij het principe berust op het zoeken van de dichtstbijzijnde buur gelet op de propensity score.

Definieer voor individu i de straftoewijzing als Z^i, de bijbehorende propensity score als π_i en de waargenomen recidive als Y^i. Zij N de oorspronkelijke grootte van de dataset, met N_0 de verzameling indices van de controlegroep en N_1 de verzameling indices van de experimentele groep. Laat M de indexverzameling zijn die alle gekoppelde daders bevat en $|M|$ de grootte van deze dataset, waarbij geldt $|M| \leq N$. Definieer de grootheid M_e als $M_e = \sum_{i \in M} Z^i$, zodat M_e het aantal gematchte daders uit de experimentele groep representeert.

In de voorgaande paragraaf hebben we gezien dat we, onder de conditionele onafhankelijkheidsaanname, voor de ontbrekende uitkomst van een individu de uitkomst van een vergelijkbaar individu uit de tegengestelde groep mogen substitueren, zodat een mogelijk causaal effect bepaald kan worden. Definieer voor $i \in N_1$

$$j(i) = \arg\min_{j \in N_0} \{|\pi_i - \pi_j|\}. \tag{1.17}$$

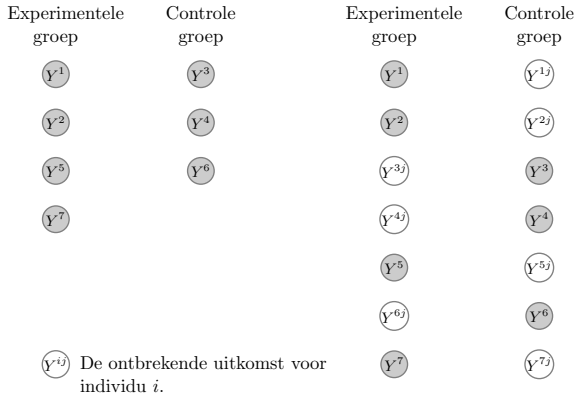
Op gelijke wijze kan $j(i)$ worden bepaald voor $i \in N_0$ met $j \in N_1$. Laat Y^{ij} de ontbrekende uitkomst zijn voor individu i als individu j als match is gebruikt, ofwel

$$Y^{ij} = \{Y^j | j \in j(i)\}. \tag{1.18}$$

In het algemeen bestaat $j(i)$ uit één waarde, zodat er geldt

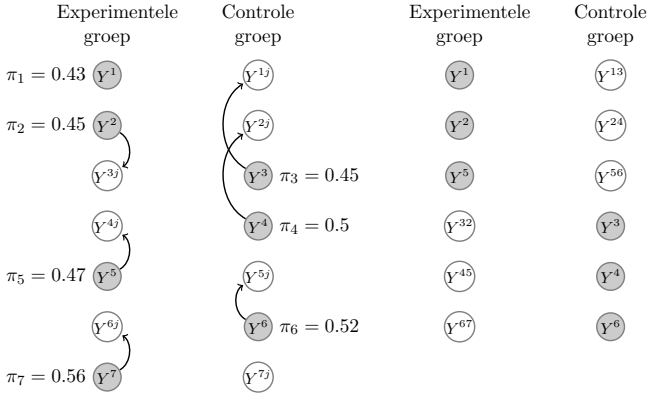
$$Y^{ij} = Y^{j(i)}.$$

We zullen het gebruik van Y^{ij} in het matchen illustreren aan de hand van een voorbeeld. Beschouw 8 individuen, waarvan er 5 zijn toegewezen tot de experimentele groep, zie Figuur 1.1. Elk individu heeft een ontbrekende uitkomst Y^{ij}. We mogen nu een individu j zoeken uit de tegengestelde groep als waarin individu i zich bevindt en zijn uitkomst Y^j substitueren voor Y^{ij}.



Figuur 1.1: Een illustratief voorbeeld voorafgaand aan het matchen, $N = 7$ en $N_1 = \{1, 2, 5, 7\}$. Links staan alle individuen toegewezen tot een groep. Rechts zijn de individuen weergeven met de bijbehorende ontbrekende uitkomst.

Beschouw eerst een mogelijke matchingsprocedure als we τ willen benaderen, dit is weergegeven in Figuur 1.2. We gaan er van uit dat elk individu slechts één keer gebruikt mag worden als match, dit wordt nader toegelicht in paragraaf 1.3.2.



Figuur 1.2: Een voorbeeld van een matchingprocedure voor benadering van τ is links weergeven. Een pijl van Y^3 naar Y^{1j} betekent dat de waarde voor Y^3 voor de ontbrekende uitkomst van Y^1 wordt gesubstitueerd, ofwel individu 3 is gevonden als match voor individu 1. Merk op dat voor individu 7 geen match kan worden gevonden, omdat alle controle individuen reeds gebruikt zijn. Individu 7 wordt daardoor buiten beschouwing gelaten. Rechts staan de gevonden koppels, met ingevulde ontbrekende uitkomsten.

We zien in Figuur 1.2 dat er geldt $|M| = 6$ en $M = \{1, 2, 3, 4, 5, 6\}$. Omdat we voor τ op zoek zijn naar het verschil van recidive van werkgestraften min de recidive van gevangenisgestraften zoeken we nu de ontbrekende uitkomsten van de individuen i uit beide groepen. We kunnen nu uit uitdrukking (1.15) een propensity score matching schatter voor τ afleiden. Vervang daartoe de verwachtingen over de recidive door de steekproefgemiddeldes. Dat wil zeggen dat we het gemiddelde van de verschillen in recidive kunnen nemen over het aantal gematchte daders $|M|$. De conditionering op $\pi(x)$ ontstaat door elke gestrafte te matchen aan een individu uit de tegengestelde groep met behulp van vergelijking (1.17). We kunnen dan de volgende schatter voor τ gebruiken, welke is geïntroduceerd in Abadie en Imbens (2012):

$$\hat{\tau} \quad = \quad \frac{1}{|M|} \sum_{i \in M} (2Z^i - 1)(Y^i - Y^{ij}). \tag{1.19}$$

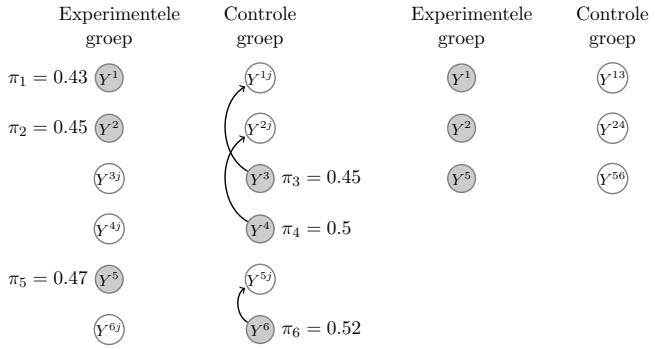
Merk op dat wanneer individu i uit de controlegroep komt, de ontbrekende uitkomst eigenlijk hoort bij de uitkomsten van de experimentele groep, waardoor we de uitkomsten moeten sorteren. De term $(2Z^i - 1)$ in vergelijking (1.19) ordent de uitkomsten zo dat voor individuen uit de controlegroep de uitkomsten ook bij de juiste groep worden ingedeeld.

In Figuur 1.2 zien we dat de volgorde van matchen invloed heeft op de koppels die er gemaakt worden. Het is daarom een voor de hand liggende eis dat de matchingprocedure in willekeurige volgorde wordt uitgevoerd.

De benadering van τ_e verschilt in dat we nu op zoek zijn naar het verschil in recidive van werkgestraften min de recidive van gevangenisgestraften gegeven dat de daders een werkstraf hebben gehad. Daarom zoeken we nu enkel de ontbrekende uitkomsten van de individuen i in de experimentele groep. We zijn dus geïnteresseerd in het verschil $Y^i - Y^{ij}$ enkel als een dader een werkstraf heeft gehad. Als een dader een gevangenisstraf heeft gehad laten we het verschil buiten bechouwing. We kunnen nu op gelijke wijze als de afleiding van vergelijking (1.19) een schatter vinden voor τ_e uit vergelijking (1.16). Echter, we nemen nu het gemiddelde van de uitkomsten over het aantal paren dat in dit geval van belang zijn, ofwel M_e. Dan verkrijgen we de volgende schatter voor τ_e:(Abadie en Imbens (2012))

$$\hat{\tau_e} \quad = \quad \frac{1}{M_e} \sum_{i \in M} Z^i (Y^i - Y^{ij}). \tag{1.20}$$

In Figuur 1.2 zien we dus dat we enkel de deelverzameling $\{1, 2, 5\} \subset M$ selecteren om de uitkomsten te gebruiken. We kunnen opmerken dat we de procedure om $\hat{\tau_e}$ te bepalen enigszins kunnen vereenvoudigen door enkel voor individuen i waarvoor geldt $Z^i = 1$ een geschikte individu j te zoeken, zodanig dat we de waarde Y^{ij} kunnen substitueren. Een voorbeeld hiervan naar aanleiding van Figuur 1.1 is weergeven in Figuur 1.3.



Figuur 1.3: Een voorbeeld van een matchingprocedure voor bepaling van $\hat{\tau_e}$ is links weergeven. Rechts staan de gevonden koppels, met ingevulde ontbrekende uitkomsten.

Door deze strategie hoeven we geen onderscheid te maken tussen $|M|$ en M_e, omdat in het bijzonder geldt $|M| = M_e$. We kunnen vergelijking (1.20) dan ook schrijven als

$$\hat{\tau_e} = \frac{1}{|M|} \sum_{i \in M} (Y^i - Y^{ij}).$$

Bovenstaande vergelijking is equivalent aan vergelijking (1.20), maar wanneer de interesse in τ_e van te voren is vastgesteld kan op deze manier werk worden bespaard in de matchingprocedure. In figuur 1.3 zien we $|M| = 3$. We kunnen opmerken dat $|M|$ gelijk is aan de grootte van de

controlegroep, wanneer dit aantal kleiner is dan de grootte van de experimentele groep. Wanneer de controlegroep groter is dan de experimentele groep zal gelden dat $|M|$ gelijk is aan de grootte van de experimentele groep.

Mogelijk kan de keuze om τ_e te bepalen dus niet alleen afhangen van de interesse van het onderzoek, maar ook van een vereenvoudigde matchingprocedure. In Artikel I wordt aangegeven dat voor elk individu uit de experimentele groep een koppel wordt gezocht uit de controlegroep. Dit duidt dus op bepaling van τ_e. De keuze voor τ_e en of deze te maken heeft met de interesse van het onderzoek of het matchen wordt in Artikel I niet toegelicht.

1.3.2 Mogelijke aanpassingen van de match strategie

In de methode van Nearest Neighbour matching zijn er vier keuzes die gemaakt kunnen worden. We zullen deze aanpassingen nu kort beschouwen.

Met of zonder teruglegging.

In bovenstaande uiteenzetting is rekening gehouden met de keuze van de auteurs van Artikel I om teruglegging niet toe te staan. Dit houdt in dat de uitkomst recidive van elke individu j maximaal één keer gebruikt mag worden om voor Y^{ij} te substitueren. We hebben gezien in Figuur 1.3 dat $M_e = 3$, wat precies het aantal controle individuen is. Echter, wanneer bij de bepaling van $\hat{\tau}_e$ teruglegging wordt toegestaan zal M altijd gelijk zijn aan het aantal individuen in de experimentele groep. Wanneer $\hat{\tau}$ wordt bepaald zoals in Figuur 1.2 zullen er met teruglegging altijd N matches kunnen worden gevormd, terwijl dit er zonder teruglegging maximaal N min het verschil tussen de groepgroottes zijn.

Er zijn twee redenen waarom aan teruglegging gedacht kan worden. De eerste mogelijkheid is om betere matches te kunnen maken. We kunnen bijvoorbeeld iemand uit de controlegroep opnieuw aan iemand uit de experimentele groep koppelen als dit meer dan één keer de beste match is. Op deze wijze voorkom je dat er slechte matches worden gemaakt. Echter, we zullen onder het kopje 'caliper matching' zien dat teruglegging niet de enige wijze is waarop dit kan worden voorkomen.

Een tweede mogelijkheid is de situatie, zoals in Artikel I, wanneer het aantal individuen in de experimentele groep groter is dan in de controlegroep. Het terugleggen kan dan een mogelijkheid zijn om ook voor de overige individuen een match te kunnen maken.

Wanneer teruglegging is toegestaan, kunnen we er vanuit gaan dat de matches van betere kwaliteit zijn dan zonder teruglegging. Het verschil tussen de echte waarde van het effect, τ, en de schatting, $\hat{\tau}$, zal naar verwachting dalen. Echter, doordat we individuen vaker dan één keer kunnen gebruiken zal de variantie van $\hat{\tau}$ of $\hat{\tau}_e$ stijgen. Merk op dat wanneer teruglegging is toegestaan de willekeurige volgorde van het matchen minder van belang is als alvorens, waar het te bepalen effect afhing van de matches die er waren gemaakt.

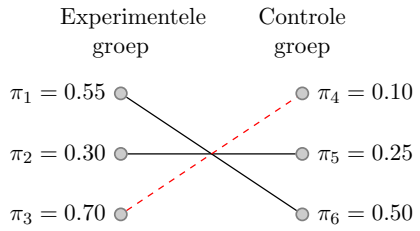
Caliper matching.

Een soortgelijke keuze als 'met teruglegging' is het instellen van een *caliper*, een grenswaarde die het maximale verschil tussen de propensity scores van twee daders aangeeft. Beide keuzes hebben het gezamenlijke doel dat het maken van slechte matches wordt voorkomen. Bekijk daartoe het eenvoudige voorbeeld in Figuur 1.4. We zien dat het individu uit de experimentele groep met een propensity score van 0.7 gekoppeld zou worden aan een dader met propensity score 0.1. Er ontstaat daardoor een match tussen twee daders die niet vergelijkbaar zijn. Om dit te voorkomen kan er worden gekozen om een caliper in te stellen.

Door het instellen van een caliper wordt er een extra restrictie opgelegd aan het vinden van de ontbrekende uitkomst Y^{ij} voor individu $i \in N_1$:

$$Y^{ij} = \{Y^j | \min_{j \in N_0} |\pi_i - \pi_j| < \varepsilon\},$$

waarbij $\varepsilon > 0$ de van te voren vastgestelde caliper is. Y^{ij} kan op gelijke wijze worden gedefinieerd voor individu $i \in N_0$ met $j \in N_1$. In het algemeen bestaat de verzameling Y^{ij} slechts uit één element. In Artikel I wordt gekozen voor $\varepsilon = 0.05$. Hoewel over de keuze van een caliper veel geschreven is,[4] wordt er hier vanuit gegaan dat wanneer twee individuen minder dan 0.05 in propensity score verschillen, de individuen als vergelijkbaar kunnen worden beschouwd.



Figuur 1.4: Een voorbeeld van gemaakte koppels, wanneer er geen gebruik wordt gemaakt van een caliper.

In het algemeen zijn de eerder genoemde uitspraken over het aantal matches dat kan worden gemaakt onder het kopje 'met of zonder teruglegging', op pagina 15, slechts een bovengrens van het aantal matches dat kan worden gemaakt bij instelling van een caliper. Wanneer de mogelijke matches niet aan de caliper voldoen, worden deze koppels buitengesloten.

De uitdrukking voor τ en τ_e zijn gebaseerd op de common support aanname. Hiermee eisen we dat voor elk individu met zekere propensity score een vergelijkbaar individu kan worden gevonden in de andere groep. Met de instelling van een caliper houd je indirect rekening met deze aanname, doordat individuen waarvoor geen vergelijkbaar individu kan worden gevonden, worden buitgengesloten.

Het voordeel van een caliper is, net als bij 'met teruglegging', dat er betere matches kunnen worden gemaakt, waardoor de zuiverheid van de schatting zal toenemen. Het nadeel is dat er in het algemeen minder matches kunnen worden gemaakt, waardoor de variantie waarschijnlijk zal stijgen.

Oversampling.

Er kan worden gekozen om voor elk individu meerdere buren te kiezen uit de andere groep en hier bijvoorbeeld een gemiddelde van te nemen. In Abadie en Imbens (2012) wordt laten zien dat de schatter van τ er dan als volgt uitziet:

$$\hat{\tau} = \frac{1}{M} \sum_{i=1}^{M} (2Z^i - 1) \left(Y^i - \frac{1}{|Y^{ij}|} \sum_{j \in Y^{ij}} Y^j \right).$$

Er geldt nu $|Y^{ij}| \geq 1$ in tegenstelling tot vergelijking (1.18), waarbij we ervan uitgingen dat deze verzameling slechts één element bevat. De keuze voor oversampling lijkt voor de hand liggend als er een overschot is aan individuen in de groep waar een geschikte ontbrekende uitkomst moet

[4] De geïnteresseerde lezer wordt verwezen naar Austin (2010a).

worden gezocht. Omdat dit in Artikel I niet het geval is, wordt er niet voor oversampling gekozen en laten we dit verder buiten beschouwing.

Variable matching.

Naast de keuze voor een caliper kan er nog een andere aanpassing gemaakt worden om koppels te vinden die zo goed mogelijk vergelijkbaar zijn. De oorspronkelijke gedachte is om koppels te maken die vergelijkbaar zijn op alle covariaten, zie paragraaf 1.2.2. Gezien de moeilijkheid van dit probleem hebben we in paragraaf 1.2.3 gezien dat de oplossing ligt bij het maken van koppels aan de hand van de propensity score. Wanneer men van een aantal variabelen zeker is van de invloed op de uitkomsten kan ervoor gekozen worden om extra op deze covariaten te matchen. In Artikel I wordt er bijvoorbeeld voor gekozen om extra te matchen op geslacht, leeftijdscategorie en straflengte. We conditioneren dan dus niet enkel op $\pi(x)$ maar ook op drie covariaten. Om hiervan een voorbeeld te geven veronderstellen we dat X_1, X_2, X_3 de covariaten voor geslacht, leeftijdscategorie en straflengte respectievelijk weergeven. We maken eerst voor elke individu $i \in N_1$ een verzameling J^i, welke de individuen $j \in N_0$ bevat die vergelijkbaar zijn op de drie covariaten:

$$J^i = \{j : |x_1^i - x_1^j| < \theta, |x_2^i - x_2^j| < \lambda, |x_3^i - x_3^j| < \rho, j \in N_0\},$$

waarbij x_p^i de geobserveerde waarde van covariaat X_p is voor individu i. De calipers $\theta, \lambda, \rho \geq 0$ zijn voorafgaand bepaald, waarbij voor een binaire variabele zoals X_1 het aannemelijk is dat θ=0. J^i kan op gelijke wijze gedefinieerd worden voor $i \in N_0$ met $j \in N_1$. Vervolgens kan Y^{ij} gedefinieerd worden als

$$Y^{ij} = \{Y^j | \min_{j \in J^i} |\pi_i - \pi_j| < \varepsilon\}.$$

Door extra te matchen op variabelen die van belang zijn, kan er mogelijk een nog betere balans worden gevonden op de covariaten tussen de twee groepen. Doordat we nu extra vergelijkbare individuen vinden, kan de onzuiverheid van de schatting dalen. Mogelijk vinden we minder matches waardoor de variantie kan stijgen, echter vermoedelijk blijft dit beperkt omdat de propensity score een functie is van de covariaten. In Hoofdstuk 4 zullen we de invloeden van variabele matching nader bekijken.

1.3.3 Beoordeling van de balans na het matchen

Na het matchen op de propensity score zijn er twee gereduceerde groepen over van gelijke grootte, waarvan de kansverdeling van de propensity scores gelijkwaardig is, i.e.,

$$\pi(X)|Z = 0 \sim \pi(X)|Z = 1.$$

Echter, de oorspronkelijke wens is het verkrijgen van gelijke verdelingen van alle covariaten in de twee groepen. Dit wil zeggen dat we na het matchen geen systematische verschillen meer willen zien tussen beide groepen, i.e., met $(X \perp Z)|\pi(X)$ volgt

$$X|Z = 0 \sim X|Z = 1.$$

Wanneer er voor het matchen veel meer vrouwen in de experimentele groep zitten dan in de controlegroep hopen we dat na het matchen de verdeling tussen mannen en vrouwen in beide groepen ongeveer gelijk is. Omdat we in het algemeen niet hebben kunnen matchen op alle covariaten moet

er gecontroleerd worden of we, aan de hand van het matchen op de propensity score, voldoende balans hebben verkregen in de verdelingen van de covariaten. Een voor de hand liggende keuze is het uitvoeren van een t-toets, maar in Artikel I wordt aanvullend nog gekozen om gestandaardiseerde verschillen te bekijken. De t-statistieken en de gestandaardiseerde verschillen uit Artikel I zijn weergeven in Tabel 2 (Appendix A).

Twee steekproeven t-toets.

Na het matchen is het wenselijk dat er geen significante verschillen meer zijn in de verdelingen van de covariaten. Als we met het blote oog geen verschil zien in de absolute verschillen van de gemiddeldes van de covariaten rest ons de vraag of deze gelijkheid berust op toeval of dat het daadwerkelijk zo lijkt te zijn dat beide steekproeven gelijkwaardig zijn.

Beschouw een covariaat X_i welke we waarnemen in de experimentele groep en de controlegroep, respectievelijk, als $X_{i,e}$ en $X_{i,c}$ voor $i \in \{1, \ldots, p\}$ met p het aantal covariaten. Zij $\mu_{i,e}$ de verwachting van de covariaat in de experimentele groep en $\mu_{i,c}$ de verwachting van de covariaat in de controlegroep. We testen nu $H_0 : \mu_{i,e} = \mu_{i,c}$ tegen het tweezijdig alternatief $\mu_{i,e} \neq \mu_{i,c}$. De nulhypothese is equivalent met $\mu_{i,e} - \mu_{i,c} = 0$. De test statistiek is dan ook gebaseerd op de schatting $\overline{X}_{i,e} - \overline{X}_{i,c}$ voor $\mu_{i,e} - \mu_{i,c}$ die we standaardiseren met een schatting van de bijbehorende standaard afwijking

$$T = \frac{\overline{X}_{i,e} - \overline{X}_{i,c}}{\sqrt{\frac{s^2_{i,e}}{n} + \frac{s^2_{i,c}}{n}}},$$

met $s^2_{i,e}$ en $s^2_{i,c}$ de steekproef varianties van $X_{i,e}$ en $X_{i,c}$ respectievelijk.

Omdat de steekproeven voldoende groot zijn volgt aan de hand van de Centrale Limietstelling dat we de verdeling van T onder de nulhypothese kunnen benaderen met een standaard normale verdeling.(Bijma et al. (2013), p.136) We kunnen de nulhypothese toetsen tegen het tweezijdig alternatief bij onbetrouwbaarheidsdrempel α door H_0 te verwerpen als $|T| \geq \zeta_{1-\alpha/2}$, waarbij $\zeta_{1-\alpha/2}$ het $(1-\alpha/2)$-kwantiel is van de standaard normale verdeling.(Bijma et al. (2013), p.111) Wanneer de waarden voor T zijn gegeven, kunnen we opmerken dat waarden voor T dichtbij nul in het voordeel zijn voor $H_0 : \mu_{i,e} = \mu_{i,c}$.

Gestandaardiseerde verschillen.

Een gestandaardiseerd verschil[5] kan worden gebruikt om de gemiddeldes van de covariaten tussen twee groepen te vergelijken. Het gestandaardiseerde verschil is het verschil in de gemiddeldes als percentage van de gemiddelde standaard afwijking

$$d = \frac{\overline{X}_{i,e} - \overline{X}_{i,c}}{\sqrt{\frac{1}{2}(s^2_{i,e} + s^2_{i,c})}} \cdot 100.$$

De auteurs van Artikel I hebben ervoor gekozen om naast de t-statistiek dit percentage op te nemen in de resulaten. Hoewel er geen standaard criterium bestaat voor dit gestandaardiseerde verschil, wordt in Austin (2011) gesuggereerd dat er voor $|d| < 10\%$ een verwaarloosbaar verschil in gemiddeldes is.

In Austin (2011) wordt tevens aangedragen dat er, ongeacht de uitkomst van de t-toets en het

[5]In Artikel I wordt dit aangeduid met gestandaardiseerde verschiltoets (D). In Nieuwbeerta et al. (2007) wordt verwezen dat hiermee 'standardized differences' wordt bedoeld wat in Rosenbaum en Rubin (1985) wordt besproken.

gestandaardiseerde verschil, stil moet worden gestaan bij het feit dat matchen op propensity score ervoor moet zorgen dat de gehele verdeling van de covariaten tussen beide groepen gelijk moet zijn. Het bekijken van de gemiddeldes van de verdelingen kan mogelijk niet volstaan. Er kan daartoe een extra vergelijking worden gemaakt tussen hogere orde termen en interacties tussen covariaten. We gaan hier verder niet op in, de geïnteresserde lezer wordt verwezen naar Austin (2011).

Wanneer geen goede balans is verkregen in de covariaten is het mogelijk dat het logistisch regressiemodel niet voldoende is. In Caliendo en Kopeinig (2005) wordt in dit geval voorgesteld om hogere orde termen of interactie tussen verschillende covariaten aan het model toe te voegen. Als hierna nog niet voldoende balans is verkregen, kan dit op een mislukking van de conditonele onafhankelijkheidsaanname duiden en moet er gedacht worden over het gebruik van een andere methode.

1.4 Bepaling van het causale effect

Wanneer er voldoende balans is verkregen in de covariaten kunnen we de resultaten gebruiken om een bepaling te doen van het gewenste effect. Gezien er in Artikel I een bepaling wordt gedaan van τ_e richten we ons hierop. De gevonden resultaten van Artikel I staan vermeld in Tabel 3 (Appendix A). Definieer ter vereenvoudiging van de notatie:

$$\begin{aligned} \mu &= \mathbb{E}[Y_1|Z=1], \\ \nu &= \mathbb{E}[Y_0|Z=1]. \end{aligned}$$

Omdat we op zoek zijn naar τ_e, moet er een bepaling worden gedaan van μ en ν. We hebben een schatting voor μ en ν kunnen doen aan de hand van de uitdrukking voor $\hat{\tau_e}$. Schrijf namelijk aan de hand van vergelijking (1.20),

$$\begin{aligned} \hat{\tau_e} &= \frac{1}{M_e}\sum_{i\in M} Z^i(Y^i - Y^{ij}) \\ &= \frac{1}{M_e}\sum_{i\in M} Z^i Y^i - \frac{1}{M_e}\sum_{i\in M} Z^i Y^{ij} \\ &= \hat{\mu} - \hat{\nu}. \end{aligned}$$

We nemen dus als schatting voor μ en ν de waarden $\hat{\mu}$ en $\hat{\nu}$, wat neerkomt op het nemen van de gemiddelde uitkomsten van recidive die we hebben gevonden met behulp van het matchen.

In Tabel 3 (Appendix A) staan in de eerste twee kolommen de gevonden waarden voor $\hat{\mu}$ en $\hat{\nu}$, waardoor het absolute verschil neerkomt op de waarde van het op recidive geschatte effect van de werkstraf op de werkgestraften, ofwel $\hat{\tau_e}$. We willen bepalen of het negatieve verschil wat we waarnemen, wat minder recidive ten gunste van de werkstraf kan aanduiden, berust op toeval of dat het verschil wel degelijk significant is. Dit kunnen we doen aan de hand van een t-toets voor gepaarde waarnemingen.(Bijma et al. (2013), p.133) Gezien het negatieve verschil wat we waarnemen, is de toets die we willen doen $H_0 : \hat{\tau_e} \geq 0$ tegen $H_1 : \hat{\tau_e} < 0$. Om te concluderen dat het effect in het voordeel werkt voor de werkstraf is het nodig de nulhypothese te verwerpen.

We hebben door het matchen twee gepaarde steekproeven van recidive verkregen. De paren zijn dus (Y_1^i, Y_0^{ij}), waarbij $i = 1, \ldots, M_e$ met M_e de grootte van de gematchte groep en j vastgesteld voor elke i zoals beschreven is in paragraaf 1.3.1. We werken met de verschillen $D_i = Y_1^i - Y_0^{ij}$. Voor de toepassing van de t-toets veronderstellen we dat de verschillen $D_1, \ldots, D_{M_e}$ onafhankelijk

en normaal verdeeld zijn met verwachting τ_e en variantie σ^2. Als alle Y_1^i en Y_0^{ij} onafhankelijk en normaal verdeeld zijn, dan volgt met behulp van lineairiteit dat de verschillen dat ook zijn. Echter, omdat we matches hebben gemaakt aan de hand van covariaten, zullen Y_1^i en Y_0^{ij} in het algemeen niet onafhankelijk zijn. "Gelukkig is ook zonder die onafhankelijkheid de normaliteit en onafhankelijkheid van de verschillen een redelijke aanname."(Bijma et al. (2013), p.142)

De t-toets voor paren is dan de gewone t-toets toegepast op de verschillen $D_1, \ldots, D_{M_e}$. We maken daarom gebruik van de toetsingsgrootheid

$$T = \sqrt{M_e}\frac{\overline{D}}{s_D},$$

waarbij s_D de steekproef standaarddeviatie is van de verschillen.(Bijma et al. (2013), p.129) We verwerpen H_0 als $\sqrt{M_e}\frac{\overline{D}}{S_D} \leq t_{M_e-1,\alpha} = -t_{M_e-1,1-\alpha}$, waarbij $t_{M_e-1,1-\alpha}$ het $(1-\alpha)$-kwantiel van de t-verdeling is met $M_e - 1$ vrijheidsgraden en α de onbetrouwbaarheidsdrempel.

In Tabel 3 (Appendix A) is een deel van de uitkomsten van Artikel I te zien. De vijfde kolom geeft de waarde voor T weer. Voor de hand liggend is dat we de nulhypothese verwerpen voor negatieve waarden van deze grootheid. In Tabel 3 is te zien dat voor het eerste jaar geldt $T = -3.229$. Daarnaast lezen we af dat $M_e = 2123$, zodat bepaald kan worden $t_{2122,0.01} = -2.328$. De p-waarde van de eenzijdige test is dus kleiner dan 0.01, wat in de kolom van significantie wordt weergeven met drie sterren. Hoe kleiner de p-waarde des te sterker het bewijs tegen de nulhypothese is. Dit is dus ten gunste voor het effect op recidive van een werkstraf op de werkgestraften, ten opzichte van een gevangenisstraf.

De auteurs van Artikel I hebben er voor gekozen om ook het relatieve verschil op te nemen in de resultaten. Wermink et al. (2009): "Het relatieve verschil van werkstraf wordt berekend door het quotiënt te nemen van de recidive na werkstraf en de recidive na gevangenisstraf. Het relatieve verschil geeft daarmee de afwijking van de controlegroep weer."

Echter, de auteurs van Artikel I doen niet precies wat hier wordt beweerd. Zij nemen het quotiënt van de recidive na werkstraf en na gevangenisstraf, wat voor het eerste jaar neerkomt op 0.273/0.683=0.399. Dus ongeveer 40% van het "oorspronkelijke" aantal recidiveren is nog over. De afwijking van de controlegroep komt dan neer op $0.399 - 1 \approx -0.6$. Er wordt vervolgens geconcludeerd dat werkgestraften 60% minder recidiveren. Men kan zich afvragen of de redenatie op deze wijze gepast is.

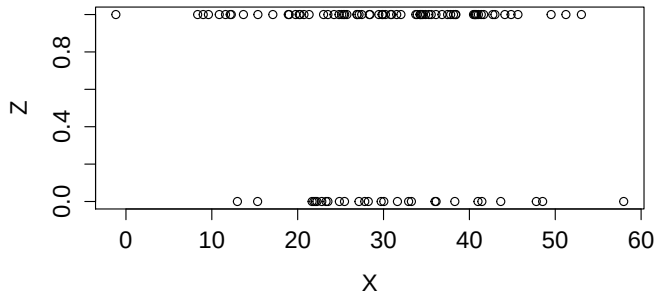
2 Schatten van de propensity score

2.1 Inleiding

De propensity score is de conditionele kans van toewijzing tot een werkstraf gegeven een vector van geobserveerde covariaten, zie definitie 1.2 op pagina 9. Om een schatting van deze kans te maken is *logistische regressie* een veel gebruikte methode. Dit hoofdstuk beschrijft de schatting van de propensity score aan de hand van logistische regressie. Deze uiteenzetting is voornamelijk gebaseerd op Hosmer (2000).

Het doel is een relatie te beschrijven tussen een *uitkomstvariabele* en een verzameling van *covariaten*. Wanneer een verzameling onafhankelijke waarnemingen in een scatterplot een lineair verband aanduidt, kan lineaire regressie worden gebruikt.(Rice (2007), p.542) Het komt echter voor dat een uitkomst variabele niet continu is, maar discreet. In dit geval kan aan logistische regressie worden gedacht. In het bijzonder wordt bij logistische regressie gebruik gemaakt van een *dichotome variabele*, een variabele die slechts twee mogelijke uitkomsten kan aannemen. Denk hierbij aan de situatie beschreven in paragraaf 1.2.1, waarbij wordt aangenomen dat de uitkomstvariabele straf slechts 'werkstraf' of 'gevangenisstraf' kan zijn. Een mogelijkheid is om op een dichotome variabele een binaire codering aan te brengen, zoals in vergelijking (1.1) op pagina 5.

Beschouw nu een dergelijke binaire variabele Z, waarbij aangenomen wordt dat deze van één enkele covariaat X afhangt. Gegeven een dataset (x_i, z_i), waarbij $i = 1, \ldots, N$, zal een scatterplot er op een soortgelijke manier uitzien als figuur 2.1.



Figuur 2.1: Scatterplot van een simulatie van toewijzing tot een werkstraf, $Z = 1$, tegen leeftijd, X.

Het probleem is dat er niet zo eenvoudig een relatie is af te lezen als in lineaire regressie. Een oplossing hiervoor is het logistische regressiemodel.

2.2 Een eenvoudig logistisch regressiemodel

Veronderstel dat Z de uitkomstvariabele is die aangeeft tot welke straf een dader is veroordeeld, waarbij de variabele de waarde 1 aanneemt als de dader een werkstraf heeft gekregen. Neem aan dat de uitkomst variabele afhangt van een enkele covariaat X, de leeftijd van de dader. Een lineair regressie model ziet er dan als volgt uit:

$$\mathbb{E}[Z|X = x] = \beta_0 + \beta_1 x,$$

waarbij $\beta_0, \beta_1 \in \mathbb{R}$ en $-\infty < \mathbb{E}[Z|X = x] < \infty$.

Nu de variabele Z een binaire variabele is, kan bovenstaand model niet worden gebruikt. Voor de binaire variabele moet voldaan worden aan de volgende vergelijking

$$0 \leq \mathbb{E}[Z|X = x] \leq 1. \tag{2.1}$$

Een soortgelijk model als het lineaire regressie model lijkt niet onredelijk. Het is mogelijk dat de conditionele verwachting of de dader een werkstraf krijgt toeneemt of afneemt met de leeftijd. Om bovenstaand model nu zodanig aan te passen dat aan vergelijking (2.1) wordt voldaan is het nodig een transformatie toe te passen van de volgende vorm:

$$\pi : \mathbb{R} \quad \rightarrow \quad [0, 1].$$

De logistische functie lijkt geschikt te zijn om de transformatie op te baseren.(Hosmer (2000)) Definieer daarom het logistische regressie model als volgt:

$$\begin{aligned} \mathbb{E}(Z|X = x) &= \pi(x), \\ \pi(x) &= \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}. \end{aligned}$$

Omdat Z een binaire variabele is, geldt:

$$\begin{aligned} \mathbb{E}(Z|X = x) &= \sum_z z p_{Z|X}(z|x) \\ &= 0 \cdot P(Z = 0|X = x) + 1 \cdot P(Z = 1|X = x) \\ &= P(Z = 1|X = x). \end{aligned}$$

Merk op dat dit precies de propensity score is, waarnaar we op zoek zijn. De binaire stochast Z heeft dus een Bernoulli verdeling met parameter $\pi(x)$,

$$Z \sim Ber(\pi(x)). \tag{2.2}$$

In het lineaire regressie model wordt aangenomen dat de geobserveerde waarde z van Z gegeven covariaat X geschreven kan worden als:

$$z = \mathbb{E}(Z|X = x) + e,$$

waarbij de term e de *ruis* wordt genoemd, de afwijking tussen een observatie en de conditionele verwachting. De meest gebruikelijke aanname is dat e een normale verdeling heeft met een gemiddelde gelijk aan nul en een bepaalde constante variantie.(Rice (2007), p.547) Echter, bij logistische regressie is dit niet het geval. De ruis term neemt hier de volgende waarde aan:

$$e = z - \pi(x)$$

Met behulp van (2.2) volgt nu dat voor de ruis geldt:

$$e = \begin{cases} 1 - \pi(x) & \text{met kans } \pi(x), \\ -\pi(x) & \text{met kans } 1 - \pi(x). \end{cases}$$

2.3 Een meervoudig logistisch regressiemodel

Als de uitkomst aangeeft of een dader een werkstraf krijgt is het aannemelijk dat de uitkomst niet enkel afhangt van de leeftijd, maar bijvoorbeeld ook van het geslacht van de dader. In Artikel I wordt ervan uitgegaan dat er naast de genoemde covariaten rekening gehouden moet worden met criminele geschiedenis, type delict en of de dader autochtoon is.

In het algemeen gaan we ervan uit dat de uitkomst variabele afhangt van een verzameling X van p onafhankelijke covariaten $X_1, \ldots, X_p$. De afleiding van het meervoudige logistische regressiemodel gaat op gelijke wijze als het eenvoudige model. Er geldt nu:

$$\begin{aligned} &\pi : \mathbb{R}^{p+1} \to [0,1]. \\ &\pi(x) = \frac{e^{x^T\beta}}{1+e^{x^T\beta}}, \end{aligned} \tag{2.3}$$

waarbij $\beta = (\beta_0, \beta_1 \ldots \beta_p) \in \mathbb{R}^{p+1}$ en $x = (1, x_1 \ldots x_p) \in \mathbb{R}^{p+1}$.

2.4 De keuze van de covariaten

In de literatuur bestaat geen algemene overeenstemming welke variabelen meegenomen moeten worden in het logistisch regressiemodel. Omdat de propensity score de kans is op een bepaalde toewijzing, in dit geval de conditionele kans op een werkstraf, is het voor de hand liggend om ervan uit te gaan dat het model de variabelen moet bevatten die de toewijzing beïnvloeden. Echter, dit zijn niet de enige variabelen die meegenomen kunnen worden in het model. In Austin (2011) wordt onderscheid gemaakt tussen vier verzamelingen: alle gemeten covariaten, alle covariaten die geassocieerd zijn met enkel de toewijzing, alle covariaten die enkel de uitkomst (recidive) kunnen beïnvloeden of alle covariaten die zowel aan de toewijzing als aan de uitkomst gerelateerd zijn. Wanneer een keuze is gemaakt welke verzameling(en) moet(en) worden gebruikt, kan het onduidelijk blijven tot welke van de verzamelingen de gemeten covariaten behoren. Een literatuuronderzoek kan hiervoor een uitkomst zijn. Brookhart et al. (2006) suggereerden dat de variabelen die enkel de uitkomst beïnvloeden in ieder geval in het model moeten worden meegenomen. Zij beweren bovendien dat het meenemen van variabelen die enkel de toewijzing beïnvloeden de variantie van de schatting kan laten stijgen, terwijl er geen reductie in de onzuiverheid van de schatting wordt waargenomen. De gemeten covariaten moeten in ieder geval voor het begin van het experiment zijn vastgesteld, zodat deze niet beïnvloed zijn door het experiment, bijvoorbeeld de ondergane werkstraf. Omdat de propensity score methode is gebaseerd op de conditionele onafhankelijkheidsaanname moet er worden nagestreeft dat de verzameling covariaten aan deze voorwaarde voldoet.

Gezien bovenstaande moeilijkheden bij de keuze van de variabelen is een voorafgaand onderzoek gewenst. In Hoofstuk 4 bekijken we mogelijke gevolgen bij verschillende keuzes van de variabelen.

2.5 Discrete covariaten

De verzameling van onafhankelijke covariaten hoeft niet enkel uit continue variabelen te bestaan. Een covariaat kan, net zoals de uitkomst variabele, discreet zijn. Stel dat X_1 een discrete covariaat is die het aantal feiten in de uitgangszaak aanneemt, zodat bijvoorbeeld geldt $X_1 \in (0, 10]$. We kunnen deze variabele toevoegen aan het model alsof het een continue variabele is. Echter, we kunnen ons ook covariaten inbeelden die geen geheeltalige waardes aannemen, maar ingedeeld kunnen worden in twee categorieën. Denk hierbij bijvoorbeeld aan het geslacht, X_2, waarbij $X_2 = 0$

wanneer de dader een man is en $X_2 = 1$ als de dader een vrouw is. De categorie 'man' wordt in dit geval ook wel de *referentiecategorie* genoemd. Hoewel de keuze van de referentiecategorie geen invloed heeft op de uitkomsten van het model kan men overwegen te refereren ten opzichte van de grootste categorie, omdat deze wellicht als het meest gebruikelijk kan worden beschouwd.

Wanneer een covariaat voorkomt die meer dan twee categorieën kan aannemen, kunnen we de bovenstaande wijze van coderen uitbreiden. Om dit te doen wordt ervoor gekozen om een verzameling *dummy variabelen* te gebruiken. In deze wijze van coderen geldt dat wanneer de discrete variabele k verschillende waarden kan aannemen, er $k-1$ dummy variabelen nodig zijn. Dit komt doordat we uitgaan van een referentiecategorie.

Ter illustratie bekijken we de covariaat 'type delict', X_j met $j \in \{1, \ldots, p\}$, en nemen we aan dat deze slechts drie waardes kan aannemen:[6]

$$X_j = \begin{cases} \text{overige wetten,} \\ \text{huis, lokaalvredebreuk,} \\ \text{openlijk geweld.} \end{cases}$$

De desbetreffende covariaat kan gecodeerd worden met behulp van twee dummy variabelen welke we noteren als x_{jl} met bijbehorende regressiecoëfficiënten β_{jl}, $l \in \{1, 2\}$. De codering is te zien in Tabel 2.1. De categorie 'overige wetten' wordt in dit geval de referentiecategorie genoemd.

Type delict	x_{j1}	x_{j2}
Overige wetten	0	0
Huis, lokaalvredebreuk	1	0
Openlijk geweld	0	1

Tabel 2.1: Een dummy codering voor een binaire covariaat.

Stel dat covariaat X_j binair gecodeerd is dan kan model (2.3) dus als volgt geschreven worden:

$$\begin{aligned} \pi(x) &= \frac{e^{x^T\beta}}{1 + e^{x^T\beta}}, \\ x^T\beta &= \beta_0 + \beta_1 x_1 + \cdots + \sum_{l=1}^{k-1} \beta_{jl} x_{jl} + \cdots + \beta_p x_p. \end{aligned}$$

2.6 Maximum likelihood schatter

Wanneer we het logistische model (2.3) willen fitten op een dataset moeten de coëficiënten $\beta \in \mathbb{R}^{p+1}$ geschat worden. We gebruiken hiervoor de maximum likelihood methode vanwege onder andere de prettige eigenschappen.(Rice (2007), p.277) Er volgt uit vergelijking (2.2) dat Z een Bernoulli verdeling heeft met parameter $\pi(x)$:

$$P(Z = z) = \pi(x)^z (1 - \pi(x))^{1-z}.$$

We nemen aan dat de waarnemingen $z_1, \ldots, z_N$ onafhankelijk zijn. Dan is de gezamelijke kansverdelingsfunctie het product van de marginale kansfuncties. De likelihood is dus als volgt te

[6] In Artikel I wordt het type delict onderverdeeld in 20 verschillende categoriën.(p.218)

schrijven

$$\text{lik}(\beta) = \prod_{i=1}^{N} \pi(x_i)^{z_i}(1-\pi(x_i))^{1-z_i},$$

waarbij x_i een vector is van de i-de waarneming van p covariaten, $x_i = (1, x_{i1}, .., x_{ip})$ met $i \in \{1 \dots N\}$. De log likelihood is dus

$$\begin{aligned} l(\beta) &= \log[\text{lik}(\beta)] \\ &= \sum_{i=1}^{N}[z_i \log(\pi(x_i)) + (1-z_i)\log(1-\pi(x_i))]. \end{aligned}$$

Om de coëfficiënten te kunnen schatten willen we nu de partiële afgeleides van de log likelihood functie gelijk stellen aan nul. We bekijken daarvoor eerst de volgende termen

$$\begin{aligned} \log(\pi(x_i)) &= \log\left(\frac{1}{1+e^{-x_i^T\beta}}\right) \\ &= -\log(1+e^{-x_i^T\beta}), \end{aligned}$$

en

$$\begin{aligned} \log(1-\pi(x_i)) &= \log\left(1-\frac{1}{1+e^{-x_i^T\beta}}\right) \\ &= -x_i^T\beta - \log(1+e^{-x_i^T\beta}). \end{aligned}$$

We kunnen nu van bovenstaande uitdrukkingen de partiële afgeleide nemen over een willekeurige parameter $\beta_j \in \beta$, dan verkrijgen we

$$\begin{aligned} \frac{\partial \log(\pi(x_i))}{\partial \beta_j} &= \frac{x_{ij}e^{-x_i^T\beta}}{1+e^{-x_i^T\beta}} \\ &= x_{ij}(1-\pi(x_i)), \end{aligned}$$

en

$$\begin{aligned} \frac{\partial \log(1-\pi(x_i))}{\partial \beta_j} &= -x_{ij} + x_{ij}(1-\pi(x_i)) \\ &= -x_{ij}\pi(x_i). \end{aligned}$$

Met behulp van deze afgeleides kunnen we nu de partiële afgeleide van de log likelihood over een $\beta_j \in \beta$ bepalen:

$$\begin{aligned} \frac{\partial l(\beta)}{\partial \beta_j} &= \sum_{i=1}^{n} y_i x_{ij}(1-\pi(x_i)) - (1-y_i)x_{ij}\pi(x_i) \\ &= \sum_{i=1}^{n} x_{ij}(y_i - \pi(x_i)). \end{aligned}$$

Dus nu volgt dat de volgende vergelijkingen moeten worden opgelost om de parameters β_0 en β_j met $j \in \{1, \dots, p\}$ te vinden, respectievelijk:

$$\sum_{i=1}^{N}(y_i - \pi(x_i)) = 0,$$

en

$$\sum_{i=1}^{N} x_{ij}(y_i - \pi(x_i)) \quad = \quad 0.$$

Omdat bovenstaande vergelijkingen niet lineair zijn in de gezochte parameters wordt in Hosmer (2000) gesuggereerd dat er gebruik kan worden gemaakt van speciale iteratieve methoden om een oplossing te vinden.

2.7 Interpretatie van de coëfficiënten

Na een schatting te hebben gemaakt van de coëfficiënten kunnen we ons afvragen hoe we deze coëfficienten kunnen interpreteren. In Tabel 1 (Appendix A) zijn de uitkomsten van het logistisch regressiemodel van Artikel I gegeven. Gebruikelijk is om een hypothesetoets te doen waarbij je probeert te toetsen of een covariaat X_j van betekenis is, door $H_0 : \beta_j = 0$ te toetsen tegen de alternatieve hypothese $H_1 : \beta_j \neq 0$. Gegeven de eigenschap van asymptotische normaliteit van de maximum likelihood schatter weten we dat de geschatte coëfficiënten standaard normaal verdeeld zijn en kan een gebruikelijke z-toets worden toegepast.(Rice (2007), p.277)

Voor de covariaat 'openlijk geweld' is het in Tabel 1 (Appendix A) eenvoudig te zien dat de waarde 0 binnen de door de auteurs gekozen betrouwbaarheidsintervallen van de geschatte coëficient valt, waarmee de nulhypothese niet wordt verworpen. De auteurs willen hiermee aangeven dat de covariaat openlijk geweld mogelijk geen significante betekenis heeft.

Naast deze gebruikelijke hypothesetoets wordt er bij logistische regressie ook vaak voor gekozen om de *odds ratio*, vanwege de prettige interpretatie, op te nemen in de presentatie van de resultaten van het model. Dit kan voor de lezer extra ondersteuning geven in het interpreteren van de covariaten en of deze mogelijk van belang zijn. Om de odds ratio te introduceren is het handig om de *logit transformatie* van $\pi(x)$ te bekijken. Voor het eenvoudige logistische regressiemodel ziet dit er als volgt uit:

$$\begin{aligned}
\text{logit}(\pi(x)) &= \log\left(\frac{\pi(x)}{1-\pi(x)}\right) \\
&= \log\left(\frac{\frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}}{1-\frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}}\right) \\
&= \log\left(\frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}-e^{\beta_0+\beta_1 x}}\right) \\
&= \log(e^{\beta_0+\beta_1 x}) \\
&= \beta_0 + \beta_1 x \\
&:= g(x).
\end{aligned}$$

De logit transformatie van het meervoudige model gaat op precies gelijke wijze.

Beschouw eerst de situatie van het eenvoudige logistische regressiemodel met een binair gecodeerde covariaat X die slechts twee verschillende waardes kan aannemen. In dit geval representeert de richtingscoëfficiënt de verandering in de logit als de variabele X van categorie verandert. Ofwel $g(1) - g(0)$, waarbij

$$X = \begin{cases} 1 & : g(1) = \beta_0 + \beta_1, \\ 0 & : g(0) = \beta_0. \end{cases}$$

Om de coëfficient β_1 te interpreteren bekijken we de odds ratio. Als we $X = 1$ als succes beschouwen dan is de odds van $X = 1$ de kans op succes gedeeld door de kans op geen succes. Merk op dat dit gelijk is aan $e^{g(1)}$. De odds ratio wordt in het algemeen gedefinieerd als de odds van $X = 1$ ten opzichte van de referentiecategorie, $X = 0$. Dus er volgt nu dat de odds ratio gelijk is aan:

$$\begin{aligned} \text{OR} &= \frac{e^{g(1)}}{e^{g(0)}} \\ &= \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} \\ &= e^{\beta_1}. \end{aligned}$$

De odds ratio benadert hoeveel meer waarschijnlijk het is voor de uitkomst om voor te komen gegeven $X = 1$ ten opzichte van $X = 0$. Dus als X de covariaat geslacht representeert, met $X = 1$ als de dader een vrouw is, dan is de odds ratio te interpreteren als hoeveel meer waarschijnlijk het is om een werkstraf te krijgen als je een vrouw bent ten opzichte van wanneer je een man bent. Merk op dat de interpretatie evengoed is toe te passen op continue covariaten en discrete covariaten, waarbij e^{β_j} de odds ratio geassocieerd met één eenheid verandering in X_j is.

De odds ratio is op gelijke wijze toepasbaar voor een binair gecodeerde covariaat X_j met dummy variabelen, zoals beschreven in tabel 2.1 op pagina 24. Dan volgt

$$X_j = \begin{cases} x_{j1} = 0 \quad x_{j2} = 0 & : g(x_j) = \beta_0, \\ x_{j1} = 1 \quad x_{j2} = 0 & : g(x_j) = \beta_0 + \beta_{j1}, \\ x_{j1} = 0 \quad x_{j2} = 1 & : g(x_j) = \beta_0 + \beta_{j2}. \end{cases}$$

Waardoor ook hier geldt dat de odds ratio van 'huis, lokaalvredebreuk' ten opzichte van de referentiecategorie gelijk is aan $e^{\beta_{j1}}$.

Dit kan uitgebreid worden naar een situatie met p verschillende covariaten. Voor een gegeven X_j geeft de coëfficient β_j de verandering in de logaritme van de odds ratio weer voor gefixeerde waarden van $x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_p$.

Waar het voor de lezer van Artikel I mogelijk lastig te interpreteren is wat het verschil in significantie precies aangeeft, kan de odds ratio een hulpmiddel zijn. In de tabel 1 van Artikel I is te zien dat de odds ratio voor een vrouw gelijk is aan 1.5. Het kan zo voor de lezer aannemelijk worden gemaakt dat het waarschijnlijk is dat het geslacht van invloed is op de toewijzing tot een werkstraf.

Echter, het weglaten van een covariaat wanneer deze niet significant lijkt te zijn of wanneer de odds ratio rond de waarde 1 ligt, kan in dit soort studies niet direct worden toegepast, onder andere wegens de belangrijke rol van de covariaten in de propensity score methode, zie aanname 1 op pagina 7. In Hoofdstuk 4 gaan we hier nog verder op in.

3 Opzet van een Monte Carlo simulatie

Het doel van deze simulatie is een illustratieve data set te vormen die vergelijkbaar is met de gebruikte data set uit Artikel I. Vervolgens kan het werkelijke effect τ_e worden berekend, welke in Hoofdstuk 4 op verschillende wijzen wordt benaderd. Aan de hand van deze simulatie kunnen geen uitspraken over de data uit Artikel I worden gedaan. Het voordeel van een simulatie is dat het werkelijke effect τ_e bekend is, waardoor de methode 'getest' kan worden. In paragraaf 3.1 wordt een beschrijving gegeven over de wijze waarop de covariaten zijn gegenereerd. In paragraaf 3.2 wordt beschreven hoe de gesimuleerde daders kunnen worden toegewezen aan de experimentele groep of de controlegroep. Vervolgens wordt voor elke dader het aantal maal recidive gegenereerd, waarna het causale effect kan worden bepaald. De beschreven simulatie is uitgevoerd in R. De code is te vinden via `http://rpubs.com/FKool/21683`.

3.1 Een illustratieve data set

Eerst wordt 100 keer een data set van 1000 daders gegenereerd, waarbij er voor elke dader 10 covariaten zijn, $X_1, \ldots, X_{10}$. Hierbij worden de covariaten $X_1, \ldots, X_{10}$ onafhankelijk gegenereerd. In Tabel 3.1 is een overzicht te zien van deze gegenereerde variabelen. De keuze voor de covariaten is gebaseerd op Artikel I. De keuze voor de parameters voor de verdelingen waarmee we de covariaten genereren, is gebaseerd op de gegevens over de verkregen experimentele en controlegroep in Artikel I. De gemiddelde leeftijd ligt bijvoorbeeld rond 28 jaar, waardoor een keuze van 28 aannemelijk is als parameter voor de Poisson verdeling, waaruit X_2 wordt gegenereerd. Omdat in Artikel I enkel de gegevens van daders met een leeftijd tussen 18 en 50 jaar wordt gebruikt, wordt de gegenereerde data waarbij X_2 niet aan deze eis voldoet buiten beschouwing gelaten.

In het logistisch regressiemodel van Artikel I wordt de covariaat 'type delict' opgesplitst in 20 verschillende categorieën. Nu onderscheiden we slechts 'geweld' en 'vermogen' van 'overige delicten', omdat in het vervolg van Artikel I ook alleen onderscheid wordt gemaakt tussen deze drie type delicten.

Variabele	Gesimuleerde waardes	Regressiecoëfficiënt
X_1: geslacht	$x_1 \in \{0,1\}$ (x_1=0 als man)	0.41
X_2: leeftijd	$x_2 \in \{18, \ldots, 50\}$	-0.12
X_3: geboren in het buitenland	$x_3 \in \{0,1\}$ (x_1=0 als autochtoon)	-1.61
X_4: aantal feiten uitgangszaak	$x_4 \in \{0, \ldots, 3\}$	0.20
X_5: type delict	$x_5 \in$ {overig,geweld,vermogen} (dummy)	(-1.6,-0.6)
X_6: aantal vermogensdelicten afgelopen jaar	$X_6 \sim$ Pois(0.8)	-0.36
X_7: aantal geweld afgelopen jaar	$X_7 \sim$ Pois(0.7)	-0.28
X_8: aantal overig afgelopen jaar	$X_8 \sim$ Pois(1)	0.15
X_9: aantal delicten afgelopen 10 jaar	$X_9 \sim$ Pois(2)	-0.10
X_{10}: strafdreiging	$X_{10} \sim$N$(50, 15^2)$ met scheefheid $\gamma = 1.5$	-0.25
constante uit logistische regressie		6
Z: straftoewijzing	$z \in \{0,1\}$ ($z = 0$ als gevangenisstraf)	-0.15

Tabel 3.1: Gegenereerde variabelen met gekozen mogelijke waardes. De bijbehorende regressiecoëfficiënten worden gebruikt om voor elke dader zijn propensity score te berekenen.

In Artikel I is niet aangegeven wat realistische waarden zijn voor het aantal feiten in de uitgangszaak of het aantal delicten in de afgelopen jaren, dus moeten we hier zelf een schatting van maken. Wel is gegeven dat 0.81 en 0.9 de gemiddeldes zijn van het aantal feiten in de criminele geschiedenis in de experimentele en controlegroep, respectievelijk.

Vermoedelijk duidt de covariaat 'ernst uitgangszaak' uit Tabel 1 van Artikel I (Appendix A) op de later genoemde 'strafdreiging', waar tevens geen waarden voor zijn gegeven. Wel is gegeven dat

een werkstraf van 60 uur gelijk staat aan een gevangenisstraf van een maand. Artikel I vermeldt dat voor het matchen de gemiddeldes in de experimentele groep en controlegroep gelijk zijn aan 106 uur en 60.3 dagen. Laten we daarom zeggen dat de straflengte in de experimentele groep een gemiddelde heeft van 50 dagen. Daarnaast wordt aangegeven dat de verdeling van de straflengte in de controlegroep erg scheef is, bijna de helft van de daders heeft een gevangenisstraf tot 2 maanden. Daarom wordt gekozen om X_{10} uit een scheve normale verdeling te simuleren. In Artikel I worden na het matchen de gemiddeldes 4.3 en 4.1 verkregen, daarom worden de gesimuleerde waarden voor X_{10} door 10 gedeeld.

3.2 De berekening van het causale effect

Eerst moet er worden gesimuleerd of een individu een werkstraf of gevangenisstraf heeft gehad. Daartoe moeten we voor elk individu i zijn propensity score berekenen met behulp van vergelijking (2.3). Hierbij worden de gesimuleerde covariaten en de gekozen regressiecoëfficiënten gebruikt, zie Tabel 3.1. De keuze van de meeste regressiecoëfficiënten is gebaseerd op de gevonden coëfficiënten in Artikel I. De geschatte waardes van de coëfficiënten van logistische regressie zijn terug te vinden in de kolom "B"van Tabel 1 (Appendix A). Merk op dat de coëfficiënt van leeftijd 10 keer zo klein is, omdat in deze simulatie de waardes van leeftijd niet door 10 zijn gedeeld. Voor 'strafdreiging' is een negatieve regressiecoëfficiënt gekozen, omdat het waarschijnlijker is dat de kans op een werkstraf kleiner wordt wanneer de straflengte hoger wordt. De constante is zo aangepast dat er ongeveer gelijke groottes van de groepen ontstonden. Na de berekening van de propensity score kan nu voor elke dader gegenereerd worden welke straf hij/zij heeft gehad met behulp van de Bernoulli verdeling, zoals we hebben gezien in (2.2).

Laten we nu voor elk individu het aantal keer recidiveren, Y, genereren. Omdat $Y \in \mathbb{N}_0$ lijkt het redelijk een Poisson log lineair model te gebruiken.(Agresti (2007)) Neem hierbij aan dat Y een Poisson verdeling heeft en dat de logaritme van de verwachting gemodelleerd kan worden door een lineaire combinatie van de parameters:

$$\mathbb{E}[Y|Z=z, X=x] = e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p+\alpha z}, \tag{3.1}$$

waarbij $\alpha \in \mathbb{R}$ de gekozen coëfficiënt voor Z is, zie Tabel 3.1. De keuze voor α is gebaseerd op de gewenste uitkomst dat er gemiddeld met een werkstraf minder gerecidiveerd wordt, net zoals in Artikel I. We kunnen nu voor elke dader Y genereren aan de hand van een Poisson verdeling,

$$Y \sim \text{Pois}(\mathbb{E}[Y|Z=z, X=x]).$$

Een enkele keer ontstaat er data waarin een dader vaker dan 10 keer recidiveert. Deze data wordt buiten beschouwing gelaten, omdat dit naar mijn inzicht niet meer als realistisch kan worden beschouwd.

Het doel is om τ_e te berekenen. Deze berekening is gebaseerd op Austin (2010b). Als we de conditionele onafhankelijkheid en de common support aannemen, dan volgt uit paragraaf 1.2.2 dat τ_e berekend kan worden door

$$\tau_e = \mathbb{E}_{X|Z=1}[\mathbb{E}[Y|Z=1, X=x] - \mathbb{E}[Y|Z=0, X=x]], \tag{3.2}$$

waarbij uit vergelijking (3.1) volgt

$$\mathbb{E}[Y|Z=1, X=x] = e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p+\alpha}$$

en

$$\mathbb{E}[Y|Z=0, X=x] = e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}.$$

Vergelijking (3.2) wordt in twee delen bepaald, waarna het gewenste effect wordt verkregen door de twee delen van elkaar af te trekken. Er geldt

$$\begin{aligned}\mathbb{E}_{X|Z=1}[\mathbb{E}[Y|Z=1, X=x] &= \int_{X_p|Z=1} \dots \int_{X_1|Z=1} e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p+\alpha} \\ &\quad f_{X_1|Z}(x_1|1)\dots f_{X_p|Z}(x_p|1)dx_1\dots dx_p.\end{aligned}$$

Bovenstaande vergelijking kan met behulp van Monte Carlo integratie benaderd kan worden door:

$$I_1 = \frac{1}{|N_1|}\sum_{j=1}^{|N_1|} e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p+\alpha}.$$

Hierbij is $|N_1|$ de grootte van de experimentele groep ($Z=1$). Als waarden voor $x_1\dots x_p$ worden de gesimuleerde covariaten $X_1,\dots,X_p$ uit de experimentele groep gebruikt. Zo geldt ook

$$\begin{aligned}\mathbb{E}_{X|Z=1}[\mathbb{E}[Y|Z=0, X=x] &= \int_{X_p|Z=1} \dots \int_{X_1|Z=1} e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p} \\ &\quad f_{X_1|Z}(x_1|1)\dots f_{X_p|Z}(x_p|1)dx_1\dots dx_p.\end{aligned}$$

Op gelijke wijze als eerder kan bovenstaande uitdrukking bepaald worden door

$$I_0 = \frac{1}{|N_1|}\sum_{j=1}^{|N_1|} e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p},$$

waarbij eveneens de waarden voor $X_1,\dots,X_p$ uit de experimentele groep worden gebruikt. Er volgt nu dat het gemiddelde effect van een werkstraf op de werkgestraften, τ_e, gelijk is aan

$$I_1 - I_0 = (e^{\alpha}-1)I_0.$$

Merk op dat wanneer de waarde voor α positief wordt gekozen, het causale effect positief is en andersom. Voor $\alpha=-0.15$ volgt dat $\tau_e=-0.37$ het werkelijke causale effect is.

De benadering van τ_e wordt, aansluitend op Artikel I, gedaan met Nearest Neighbour matching waarbij $\varepsilon=0.05$ (paragraaf 1.3), tenzij anders staat vermeld. Hierbij kan het package 'Matching' in R worden gebruikt.(Sekhon (2013)) Het verschil tussen de echte waarde van τ_e en de schatting $\hat{\tau_e}$ als alles correct gespecificeerd is, is deels te verklaren door het gebrek aan teruglegging in de matching methode. Daarnaast zal de benadering van τ_e vermoedelijk verbeteren als het aantal maal dat een data set wordt gegenereerd groter wordt gemaakt dan 100.

Er is gekozen om in de simulatie enkel groepen van ongeveer gelijke grootte te bekijken, in tegenstelling tot Artikel I waar de experimentele groep uit ongeveer 70% van alle daders bestaat. Een ongelijke verdeling van de groepen heeft naar aanleiding van de simulatie vermoedelijk weinig invloed op de gevonden resultaten.

4 Discussie

Naar aanleiding van Artikel I is er een discussiestuk geschreven door Groenendijk en van Delft. Wermink et al. (de auteurs van Artikel I) hebben de gelegenheid gekregen een weerwoord te geven op dit discussiestuk. In dit hoofdstuk worden een aantal samengestelde discussiepunten besproken op basis van Groenendijk en van Delft (2013a). We bekijken daarbij het weerwoord in Wermink et al. (2013) en proberen, naar aanleiding van een simulatie en de verkregen kennis in de eerdere hoofdstukken, een mening te vormen over de gegeven kritiek. Niet alle kritiekpunten uit het geschreven discussiestuk van Groenendijk en van Delft worden besproken. We beperken ons tot de discussiepunten die direct gerelateerd zijn aan de gebruikte statistische methodiek, of meer specifiek aan propensity score matching. Discussiepunten die meer verbonden zijn aan criminologie, zoals de gehanteerde definitie van recidive, worden buiten beschouwing gelaten.

4.1 De gebruikte statistische methodiek

Het eerste discussiepunt wat we zullen bespreken betreft de gebruikte methodiek. Groenendijk en van Delft vinden dat een aantal stappen uit de propensity score matching methode op onvolledige wijze zijn besproken en/of weergegeven. De gegeven kritiek is opgesplitst, waarbij elk deel wordt aangegeven met de notatie [i].

> *"De onderzoekers stellen terecht vast dat er eigenlijk experimenten zouden moeten worden uitgevoerd met random toewijzing van taakstraf of gevangenisstraf. Nu was het in hun woorden cruciaal om voor selectieprocessen te controleren. De auteurs staan in algemene zin uitgebreid stil bij deze noodzaak, maar besteden geen woord aan de voorwaarden en beperkingen van de gekozen methodes voor die controle."[1]*
>
> [...] *"Bij de gehanteerde methodiek was sprake van zes afzonderlijke stappen. Twee daarvan zijn de keuze van de variabelen en het vaststellen van een model op basis van logistische regressie. Deze stappen zijn op een onvolledige wijze weergegeven in een tabel. Elke aanduiding ontbreekt van de gevonden fit van het verkregen model. Omdat er geen informatie verschaft wordt over het wel of niet 'passen' van het model, zijn er slechts twee regels in het artikel die de keuze van de variabelen onderbouwen."[2]*
>
> [...] *"De rechters deden hun best om taakstraffen en gevangenisstraffen te geven aan die verdachten bij wie die soort straf het beste paste. De inspanningen van de rechters zijn door de onderzoekers dus van tafel geveegd. "[3]*
>
> [...] *"De propensity score matching methode heeft met name opgang gemaakt in de context van onderzoek naar het gebruik van medicijnen. Dit onderzoek richtte zich op crimineel gedrag door mensen die voor de eerste keer veroordeeld waren voor een misdrijf. De groep die gevangenisstraf kreeg, is door de onderzoekers bestempeld tot controlegroep: alsof dit geen 'behandeling' is die effect zou kunnen hebben op de neiging tot criminaliteit nadien en de taakstraf wel."[4]*
>
> — Groenendijk & van Delft (2013a), p.60,61.

De auteurs van Artikel I gaven hierop het volgende weerwoord.

> *"Rechters streven naar het opleggen van een passende straf en zullen hun keuze voor een bepaalde strafmodaliteit daarom juist baseren op bepaalde daad- en daderkenmerken. Groenendijk en van Delft suggereren dat de gehanteerde matchingsmethode de inspanningen van rechter, om te komen tot een goed gemotiveerde strafoplegging,*

'van tafel veegt'. Dit is echter geenszins het geval. De door ons gehanteerde methode van propensity score matching maakt juist optimaal gebruik van de manier waarop de rechter tot een beslissing komt."[2]

"Afgaand op de uitkomsten van het beschikbare straftoemetingsonderzoek kunnen we concluderen dat onze propensity score is gebaseerd op factoren die een belangrijke rol spelen in de rechterlijke beslissing aangaande welke straf als passend wordt gezien. Zelfs al bevat ons logistisch regressiemodel voor de straftoemeting belangrijke daad- en daderkenmerken en wordt balans op alle meegenomen kenmerken bereikt, toch is het aannemelijk dat niet alle kenmerken die in de straftoemeting een rol spelen, konden worden meegenomen. Om na te gaan in hoeverre onze resultaten robuust zijn voor de mogelijke invloed van ongeobserveerde daad- of daderkenmerken voerden wij daarom een aantal sensitiviteitsanalyses uit."[1]

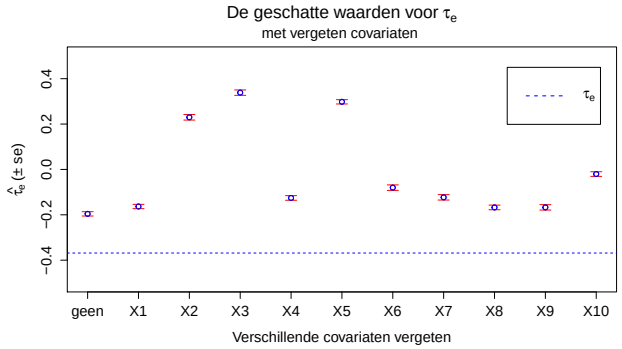
[...] *"In experimenten wordt gesproken van een experimentele en een controlegroep. De experimentele groep ondergaat de interventie waarvan men het effect beoogt vast te stellen. De controlegroep ondergaat doorgaans geen behandeling, een placebobehandeling, of de 'behandeling zoals gebruikelijk'. De keuze van de controlegroep bepaalt de uitspraken die op basis van het experiment kunnen worden gedaan. Ontvangt de controlegroep 'behandeling zoals gebruikelijk' dan vormt de gebruikelijke behandeling het basisniveau waartegen het effect van de experimentele interventie wordt afgezet."[4]*

— Wermink et al. (2013), p.70-72.

Ten aanzien van [1] kan het volgende gezegd worden: De 'voorwaarden en beperkingen' waar Groenendijk en van Delft waarschijnlijk het meest op doelen is de conditionele onafhankelijkheidsaanname, een belangrijke aanname waar de gehele methode op is gebaseerd, zie paragraaf 1.2.2. Het is duidelijk dat de aanname alleen opgaat wanneer alle covariaten, die van invloed zijn op de uitkomsten, worden meegenomen. Zoals Wermink et al. aangeven is het ondanks uitgebreid onderzoek onmogelijk om zeker te zijn welke covariaten daadwerkelijk van invloed zijn geweest. De vraag is wat de invloed is op de schatting, wanneer we te maken hebben met ongeobserveerde covariaten. Het is daarom aanbevolen om een vorm van sensitiviteitsanalyse te verrichten, zie bijvoorbeeld Caliendo en Kopeinig (2005). Wermink et al. beweren in hun weerwoord dat zij een vorm van sensitiviteitsanalyse hebben uitgevoerd, waarin ze nagaan in hoeverre de resultaten robuust zijn voor de mogelijke invloed van ongeobserveerde covariaten. Echter, dit is niet wat er daadwerkelijk gebeurt. De auteurs maken onderscheid naar geslacht en verschillende leeftijdscategorieën en concluderen dat de verschillen in recidive dezelfde conclusies geven.(Wermink et al. (2009), p.223) We zien daardoor dat het gemeten effect niet verstoord is door, bijvoorbeeld, geslacht, omdat we voor beide een negatief effect observeren. Het is dus niet zo dat het effect voor vrouwen positief is, maar dat dit verborgen is doordat de groep mannen groter is. De auteurs maken hier dus helemaal geen gebruik van eventueel ongeobserveerde covariaten, zoals wel wordt beweerd.

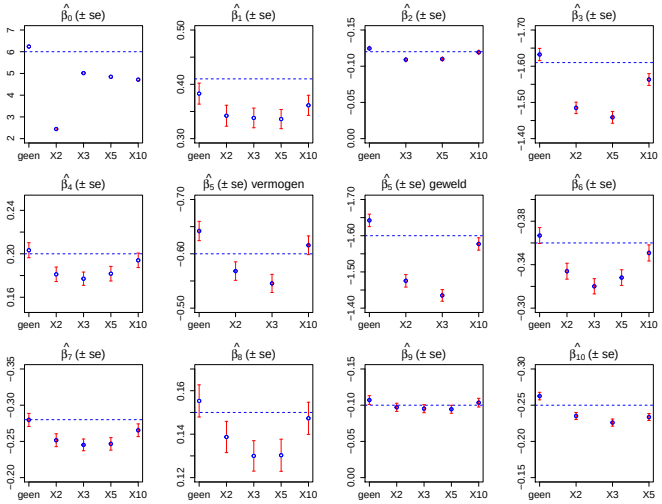
Groenendijk en van Delft lijken dus op indirecte wijze aan te duiden dat er in Arikel I niet genoeg stil wordt gestaan bij de conditionele onafhankelijkheidsaanname. Om meer inzicht te krijgen in de bovenstaande kritiek, bekijken we met een vorm van sensitiviteitsanalyse de invloed van de keuzes van de variabelen en het belang dat het regressiemodel de werkelijkheid goed benadert.

We beschouwen de invloed op de uitkomst van de schatting, wanneer er ongeobserveerde covariaten zijn. Dit wil zeggen dat we bekijken wat de invloed op de schatting is, wanneer we een covariaat vergeten mee te nemen in het regressiemodel, terwijl deze wel degelijk van invloed is op de uitkomst.



Figuur 4.1: Benadering τ_e, $N = 947.9(\pm 1.5)$ en $|N_1| = 472.7(\pm 5.1)$. De situatie 'geen' duidt op geen covariaten vergeten in het regressiemodel.

In Figuur 4.1 zien we dat wanneer we de covariaten X_2, X_3, X_5 of X_{10} vergeten $\hat{\tau_e}$ een positieve waarde heeft, wat duidt op gemiddeld minder recidive ten gunste van een gevangenisstraf. Wanneer deze covariaten worden vergeten zien we dus dat de uitkomst dermate verandert dat we een andere conclusie willen trekken. We kunnen ons afvragen hoe dit is veroorzaakt en waarom juist deze covariaten voor een groter verschil met de echte schatting zorgen, in vergelijking tot het vergeten van andere covariaten. Er is gebleken dat de t-toetsen en de gestandaardiseerde verschillen, zie paragraaf 1.3.3, beiden laten zien dat er voldoende balans is verkregen.[7] Omdat er voldoende balans is verkregen, keren we terug naar de bepaling van het logistisch regressiemodel. De schattingen voor de regressiecoëfficiënten zijn weergeven in Figuur 4.2.



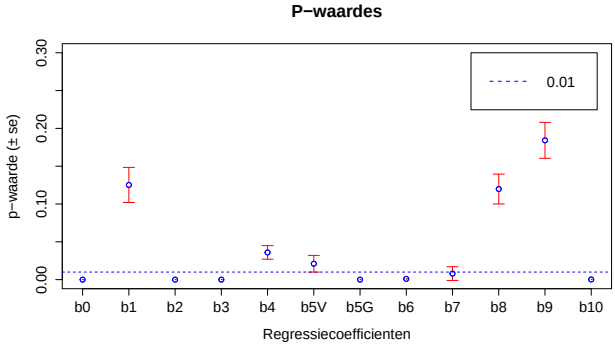
Figuur 4.2: Schattingen van de regressiecoëfficiënten, in de simulatie van Figuur 4.1, wanneer X_2, X_3, X_5 of X_{10} is vergeten in vergelijking tot wanneer er geen covariaten zijn vergeten. Coëfficiënt $\hat{\beta}_5$ is opgedeeld in twee coëfficiënten, omdat X_5 een dummy variabele is, zie Tabel 3.1. De stippellijnen geven de echte waarden van β_i weer, zie Tabel 3.1.

Er is een duidelijk verband te zien tussen de verkregen schattingen van τ_e en het logistisch regressiemodel. De slechtste schatting van τ_e werd verkregen wanneer X_3 werd vergeten. In Figuur 4.2 zien we dat de schattingen van de regressiecoëfficiënten over het geheel genomen ook het slechtste zijn wanneer we X_3 vergeten. We zien ook dat deze schattingen het beste waren wanneer we X_{10}

[7] Deze resultaten zijn wegens de ruimte niet opgenomen in de gepresenteerde uitkomsten.

vergeten, wat ook de beste van de slechtste schattingen was van τ_e in Figuur 4.1. We kunnen op grond hiervan dus concluderen dat in dit geval de schatting van τ_e slechter wordt doordat het logistisch regressiemodel de ware coëfficiënten niet meer goed benadert. Dus als het logistisch regressiemodel niet meer goed de waarheid benadert, kunnen we vrijwel zeker zijn dat de geschatte waarde $\hat{\tau_e}$ ook niet naar waarheid is.

We kunnen ons afvragen waarom de covariaten X_2, X_3, X_5 en X_{10} meer invloed op de schatting hebben als ze worden weggelaten dan de andere covariaten. Elke keer dat het regressiemodel in de simulatie wordt gefit, wordt voor de coëfficiënten een hypothese toets uitvoerd. Hiermee wordt gekeken of de covariaten 'van betekenis' zijn, zoals beschreven in paragraaf 2.7. We weten dat wanneer we een kleinere p-waarde hebben, er meer bewijs tegen de nulhypothese is.



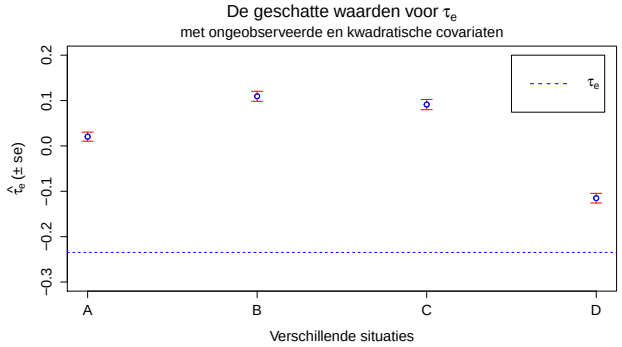
Figuur 4.3: De gemiddelde p-waardes van een hypothesetoets voor de regressiecoëfficiënten wanneer alle covariaten zijn meegenomen in het model.

Figuur 4.3 bevestigt dat de regressiecoëfficiënten van de covariaten X_2, X_3, X_5 en X_{10} gemiddeld genomen kleine p-waardes hebben, waardoor we ervan uitgaan dat deze covariaten van betekenis zijn voor het model. Het valt op dat de covariaat X_6 gemiddeld genomen ook een kleine p-waarde heeft, dit sluit aan bij het resultaat in Figuur 4.1 waar te zien is dat de benadering van τ_e bij het vergeten van X_6 ook niet meer voldoende is. We zien dus dat wanneer een covariaat een voorspelling kan hebben voor de uitkomst, ofwel wanneer er een kleine p-waarde is, het dermate veel invloed heeft voor de uitkomst dat de conclusies kunnen veranderen als we deze covariaat vergeten. Echter, we moeten ons niet teveel laten leiden door de uitgevoerde toets. Het is best mogelijk dat volgens de uitgevoerde hypothesetoets een covariaat niet zo van 'belang' is, zoals X_1, X_8 of X_9 (Figuur 4.3), maar dat deze wel degelijk in het model hoort. Wanneer we één van deze covariaten weglaten zien we in Figuur 4.1 dat de schatting slechter wordt. Het is dus van belang om uitgebreid onderzoek te doen naar welke covariaten van invloed zijn op de uitkomsten, zie ook paragraaf 2.4. Er volgt dus dat de kritiek van Groenenijk en van Delft, gezien de keuze van de variabelen en de presentatie daarvan, terecht is.

Er is nog een andere mogelijke verklaring waarom de covariaten X_2, X_3, X_5 en X_{10} meer invloed hebben op de schatting als ze worden weggelaten. Door de lineaire combinatie van de covariaten in het logistisch regressiemodel hebben deze covariaten mogelijk meer invloed op de uitkomsten, omdat deze variabelen de grootste waarden kunnen aannemen. In Tabel 3.1 wordt deze gedachte bevestigd. Er is te zien dat X_2 grotere waarden kan aannemen in vergelijking tot bijvoorbeeld X_3, maar dat X_3 juist een grotere regressiecoëffiënt heeft.

T.a.v. [2]: Groenendijk en van Delft maken onder andere een kanttekening bij het ontbreken van een gevonden fit van het verkregen regressiemodel. Het geven van informatie over het uiteindelijk verkregen regressiemodel en een fit van dit model kan inderdaad gewenst zijn.

Ter illustratie van deze kritiek kijken we, naar aanleiding van het regressiemodel van Artikel I, naar de invloeden van een kwadratisch gesimuleerde covariaat. In het logistisch regressiemodel van Artikel I is de covariaat 'leeftijd' kwadratisch meegenomen. Omdat de auteurs geen informatie verschaffen over de manier waarop het uiteindelijke regressiemodel is verkregen, is de motivatie voor deze kwadratische covariaat niet bekend. Het is bijvoorbeeld mogelijk dat er zonder deze kwadratische term geen balans werd verkegen na het matchen, zie paragraaf 1.3.3.



Figuur 4.4: De geschatte waarden voor τ_e, waarbij X_2 kwadratisch is gesimuleerd. **A**: X_2 lineair in regressiemodel, **B**: X_8, X_{10} ongeobserveerd, X_2 lineair in model, **C**:X_8, X_{10} ongeobserveerd, X_2 kwadratisch in model, **D**:X_2 kwadratisch in model.

In Figuur 4.4 bekijken we onder andere wat er met de schatting gebeurt wanneer de covariaat 'leeftijd' kwadratisch gesimuleerd wordt, terwijl een lineaire combinatie van de covariaten in het regressiemodel wordt gebruikt. De coëfficiënt voor X_2 is veranderd ten opzichte van Tabel 3.1 naar -0.004, omdat de covariaat nu kwadratisch gesimuleerd is. We hebben hiervoor gezien dat er meerdere moeilijkheden kunnen zijn in het regressiemodel, daarom zijn er een aantal situaties gecombineerd.

Figuur 4.4 geeft aan dat er een andere conclusie uit het onderzoek wordt getrokken als 'leeftijd' lineair in het model is meegenomen in plaats van kwadratisch, zoals gesimuleerd is. De schatting wordt slechter naarmate er meerde covariaten worden vergeten en X_2 lineair wordt meegenomen in het model, wat aansluit bij de verwachtingen. Daarnaast verwachten we dat de schattingen voor de bijbehorende regressiecoëfficiënt β_2 een stuk slechter zijn wanneer 'leeftijd' lineair is meegenomen. De simulaties bevestigen deze verwachting gezien $\hat{\beta}_2 = -0.23266(\pm 0.00209)$ als X_2 lineair werd meegenomen en $\hat{\beta}_2 = -0.00407(\pm 3e{-}05)$ wanneer X_2 kwadratisch werd meegenomen, wat de echte waarde wel goed benadert. De andere regressiecoëfficiënten konden wel goed geschat worden, maar de slechtere schatting van β_2 verklaart de slechtere schatting van τ_e.

We hebben nu meerdere malen gezien dat de invloed van de keuzes in het regressiemodel zo ingrijpend kunnen zijn dat de conclusies van het onderzoek kunnen veranderen. Het kan daarom zorgelijk zijn dat er in Artikel I geen informatie wordt verschaft over het uiteindelijk verkregen regressiemodel. We kunnen op zijn minst een goede argumentatie verwachten voor de keuze van de covariaten. Er is begrip voor de moeilijkheid van het probleem, maar ik sluit me bij Groenendijk en van Delft aan dat de argumentatie voor de keuze van de covariaten, zeker in combinatie met de beperkte wijze van presenteren van het regressiemodel, niet voldoende is.

T.a.v. [3]: Naar aanleiding van het weerwoord van Wermink et al. kunnen we concluderen dat zij het bij het juiste eind hebben. Door te conditioneren op de juiste covariaten gebruiken we het feit dat we weer een situatie hebben verkregen waarin de straftoewijzing willekeurig is verlopen, wat de oorspronkelijke wens is. De inspanningen van de rechters zijn in die zin dus van belang

voor de gebruikte methode. Dit is uitgebreid beschreven in paragraaf 1.2. Ook hier blijft een moeilijke factor het vinden van de "juiste"covariaten.

T.a.v. [4]: De keuze van de controlegroep maakt statistisch gezien geen verschil. Men kan in vergelijking (1.1) evengoed de codering omdraaien, waardoor $Z = 0$ correspondeert met een dader die een werkstraf heeft gehad. Het enige wat door deze keuze veranderd is, zoals Wermink et al. ook aangeven, de conclusie die uiteindelijk wordt gegeven. De controlegroep wordt als basis genomen, waarmee de "nieuwe" methode wordt vergeleken. De keuze van de controlegroep hangt dus enkel af van het uitgangspunt van het onderzoek. Wermink et al. staan dus volkomen in hun recht om als controlegroep de gevangenisgestraften te nemen.

4.2 De gekozen match strategie

Het volgende discussiepunt van Groenendijk en van Delft betreft grotendeels de keuze voor een caliper in de match strategie 'Nearest Neighbour matching'. Nadat deze kritiek is behandeld, bekijken we ter illustratie de invloeden van variabele matching, waarbij naast een match op de propensity score ook een match op één of meerdere covariaten wordt gezocht.

4.2.1 Nearest Neighbour matching met een caliper

De gegeven kritiek is opgesplitst, waarbij elk deel wordt aangegeven met de notatie [i].

> *"Er is sprake van twee manieren van matchen: nearest neighbour matching en een maximaal verschil in propensity score. Wanneer eenduidig was gekozen voor de eerste manier had men aan ieder individu uit de kleinste groep een individu kunnen koppelen uit de grotere groep.[1] In plaats daarvan is gekozen voor een mengsel van beide methodes: Een persoon uit de controlegroep werd gekoppeld aan een individu uit de experimentele groep, wanneer het verschil in de geschatte kans op werkstraf voor beide personen niet meer bedroeg dan 0.05. Dit resulteert in het schrappen van personen in beide uiteinden van de gematchte groep.[2] Voor 39% van de personen in de controlegroep kon geen match gevonden worden. De aard van het delict is een zeer voor de hand liggende voorspeller. De groep van 39% is echter niet op basis van deze variabele geschrapt, maar op basis van de geconstrueerde propensity score, waar het delicttype slechts een bouwsteen van vormt.[3] Hiermee raken we aan een ander problematisch aspect van de opzet van de analyse: de bijna 73% van de werkgestraften en 39% van de gevangenisgestraften daders die buiten de vergelijking zijn gehouden, omdat ze niet passen vanwege het te grote verschil in propensity score, zijn wel meegenomen bij de opstelling van het model voor diezelfde propensity score.[4]"*
>
> — Groenendijk & van Delft (2013a), p.62.

De auteurs van Artikel I gaven hierop een weerwoord, waarin ze ingaan op [2] en [4].

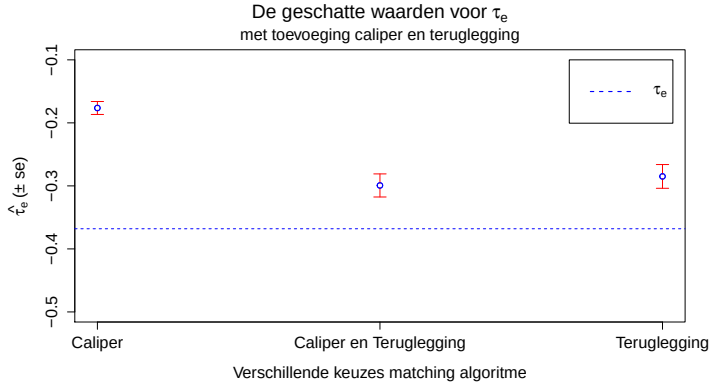
> *"Wij kozen voor caliper matching, een variant van nearest neighbour matching. Het gebruik van een caliper voorkomt dat leden van de experimentele groep uiteindelijk gekoppeld worden aan controlepersonen die weliswaar dichtsbijzijnd, maar desalniettemin behoorlijk verschillend zijn.[2] Caliper matching resulteert dan ook in het schrappen van experimentele en controlepersonen met extreme waardes op de propensity score: personen die, gegeven hun geobserveerde kenmerken, heel veel of juist heel weinig kans hebben om aan de experimentele conditie te worden toegewezen. Voor deze personen kan geen goede match in de data worden gevonden. Het weglaten van onvergelijkbare*

personen betekent dat we hele betrouwbare uitspraken kunnen doen over de groep waar we uitspraken over doen. De prijs die hiervoor wordt betaald, is echter dat het geschatte effect van werkstraf enkel geldt voor die werkgestraften die voldoende vergelijkbaar zijn met personen die een korte gevangenisstraf kregen opgelegd.[4] "

— Wermink et al. (2013), p.72-73.

Ten aanzien van [1] kan het volgende worden gezegd: Wanneer eenduidig was gekozen voor nearest neighbour matching is het niet het geval dat men aan ieder individu uit de kleinste groep een individu koppelt uit de grotere groep. De wijze van matchen is afhankelijk van of τ of τ_e moet worden bepaald, zie paragraaf 1.3.1. Het toelaten van een caliper wijzigt niet de keuze voor τ of τ_e, maar geeft enkel een restrictie aan het algoritme waarmee dit wordt gedaan.

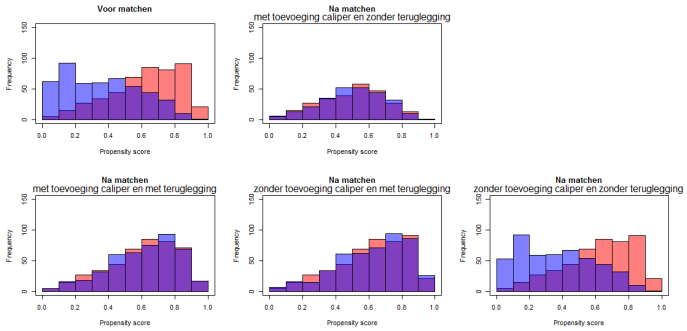
T.a.v. [2] : Met een mengsel van nearest neighbour matching en een maximaal verschil in propensity score wordt geduid op nearest neighbour matching, waarbij gebruik wordt gemaakt van een caliper (zie paragraaf 1.3.2). Wanneer er geen caliper wordt ingesteld, is het mogelijk dat er zoveel slechte koppels ontstaan dat het geschatte effect zodanig is beïnvloed dat de realiteit niet meer wordt benaderd. Dit kunnen we nader bekijken aan de hand van een simulatie. We benaderen daartoe τ_e met $\hat{\tau}_e$ door nearest neighbour matching toe te passen met verschillende aanpassingen. De uitkomsten zijn zichtbaar in Figuur 4.5.



Figuur 4.5: Benadering τ_e met 100 keer de gesimuleerde data set, $N = 953.86(\pm 0.62)$, $|N_1| = 466.58(\pm 1.60)$, caliper $\epsilon = 0.05$.

Wanneer er geen caliper is toegevoegd en geen teruglegging is toegestaan is de geschatte waarde voor τ_e gelijk aan $\hat{\tau}_e = 1.1881(\pm 0.0116)$. Deze waarde ligt zo ver van de de echte waarde af dat de benadering uit Figuur 4.5 is gelaten. Blijkbaar ontstaan er dermate veel slechte koppels dat het verschil tussen τ_e en $\hat{\tau}_e$ zal stijgen, in tegenstelling tot wanneer deze koppels worden buitengesloten. Het toevoegen van een caliper zorgt er dus voor dat de zuiverheid van de schatting toeneemt. Wanneer zowel een caliper als teruglegging is toegelaten, verbeteren de matches zodanig dat deze schatting het beste de echte waarde voor τ_e benadert. Echter, we zien wel dat in de schattingen 'met teruglegging' de standaardfouten vergroten. Deze resultaten komen overeen met de verwachtingen uit paragraaf 1.3.2.

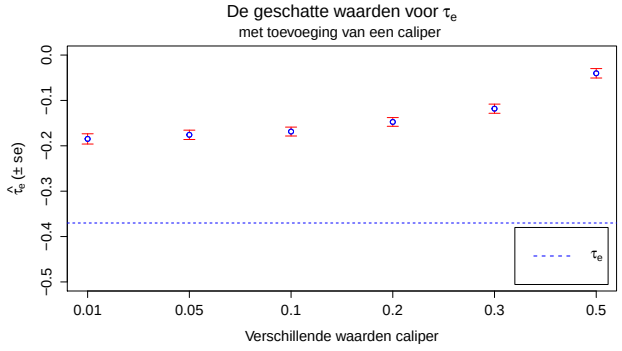
Door de verschillende match strategiën ontstaan er verschillen in de schattingen van $\hat{\tau}_e$, wat zichtbaar is in Figuur 4.5. Dit verschil wordt veroorzaakt door het verschil in matchen, gezien het feit dat in de simulatie voor alle strategiën hetzelfde logistisch regressiemodel is gebruikt. Figuur 4.6 geeft de propensity score verdelingen voor en na het matchen weer.



Figuur 4.6: Propensity score verdelingen van een enkele simulatie, waarbij blauw de controlegroep, rood de experimentele groep en paars de overlap. $N = 965$, $|N_1| = 488$,, caliper $\epsilon = 0.05$.

We zien dat wanneer teruglegging is toegestaan controle individuen met een hoge propensity score vaak hergebruikt worden, waardoor ook de waardes voor het aantal recidive in grote mate hergebruikt worden. Men kan zich afvragen of teruglegging daarom gepast is. In Figuur 4.6 valt op dat bij de instelling van een caliper individuen met extreme waardes voor de propensity score worden buitengesloten, waardoor niet vergelijkbare koppels worden uitgesloten. De plot van wanneer enkel teruglegging is toegelaten, ziet er soortgelijk uit als met teruglegging en met caliper, gezien de overeenkomst tussen deze twee restricties. Wat we uit Figuur 4.6 dus nogmaals kunnen concluderen is dat het verschil in de verdeling van de propensity scores, ofwel of er veel matches zijn gemaakt die niet voldoende vergelijkbaar zijn, zorgt voor een slechtere schatting als er geen caliper (of teruglegging) is gebruikt. We kunnen daarbij opmerken dat de vorm van de propensity score verdelingen ook samenhangt met de benadering van τ_e. In de plots waarbij teruglegging is toegestaan, is de verdeling van de propensity scores ongeveer gelijk aan de verdeling van de propensity scores van de experimentele groep voor het matchen. Dit komt omdat τ_e wordt bepaald, waardoor voor de daders uit de experimentele groep een match wordt gezocht. Wat opvalt is dat het toestaan van teruglegging zorgt dat de verdeling van de propensity scores in de experimentele groep meer wordt behouden en daarmee een betere benadering van τ_e geeft, zie Figuur 4.5.

Laten we nu de situatie uit Artikel I bekijken waarin we geen teruglegging gebruiken, maar wel een caliper. Verwacht kan worden dat wanneer de caliper groter wordt, de uitkomst van het matchen zonder caliper en zonder terugleggen benaderd zal worden. In Figuur 4.7 wordt dit vermoeden bevestigd.



Figuur 4.7: Benadering τ_e met 100 keer de gesimuleerde data set, $N = 953.60(0.71)$, $|N_1| = 465.38(1.40)$.

Merk op dat de schattingen voor kleine waarden van de caliper niet veel verschillen, terwijl deze daarna toeneemt. Op grond hiervan lijkt een keuze van 0.05 voor de caliper redelijk.

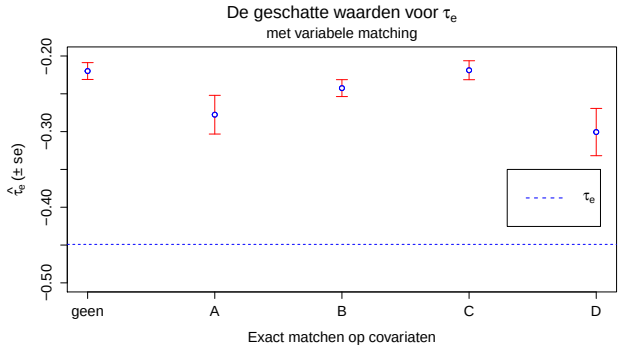
Door het uitsluiten van koppels waarvan de individuen niet voldoende vergelijkbaar zijn, verbetert de schatting aanzienlijk ten opzichte van wanneer we deze koppels wel toelaten. De keuze van een caliper in de match strategie lijkt dus terecht; men zal in het algemeen alleen nog een overweging moeten maken of teruglegging wordt toegelaten.

T.a.v. [3] : De groep van 39% is inderdaad geschrapt op basis van de geconstrueerde propensity score en niet enkel op basis van het 'type delict'. Het liefst wil men conditioneren op de gehele vector van covariaten, maar gezien de grootte van de vector is dit vaak niet mogelijk. In paragraaf 1.2.3 hebben we gezien dat het voldoende is om te conditioneren op slechts een functie van de covariaten, de propensity score. Wanneer men denkt dat een bepaalde covariaat van groot belang is, kan er extra op deze variabele worden gekoppeld, zie paragraaf 1.3.2.

T.a.v. [4] : De daders die niet gekoppeld konden worden zijn inderdaad wel meegenomen bij de opstelling van het logistisch regressiemodel. Dit heeft te maken met de volgorde van het uitvoeren van de methode, wat is beschreven in Hoofdstuk 1. Eerst wordt de propensity score geschat, waarna we de daders matchen op deze propensity score. Stel dat de daders die niet gekoppeld konden worden buiten beschouwing worden gelaten bij de opstelling van het logistisch regressiemodel. Op basis daarvan vinden we andere waarden voor de regressiecoëfficiënten, zo ook andere propensity scores, waarna de matching procedure opnieuw moet worden uitgevoerd. We komen zo in een cykel terecht. Daarnaast is het in het algemeen prettig om een grotere data set te hebben, zodat de propensity scores meer naar waarheid kunnen worden geschat. Ten aanzien van het weerwoord op [4] van Wermink et al. heeft het weglaten van vele daders tot gevolg dat het geschatte effect enkel geldt voor de personen die voldoende vergelijkbaar zijn. Naar mijn mening had in de conclusie van Artikel I meer benadrukt moeten worden dat we bijna drie kwart van de werkgestraften niet kunnen gebruiken, zodat we ons bewust kunnen zijn dat de generaliseerbaarheid van de resultaten mogelijk beperkt kan blijven.

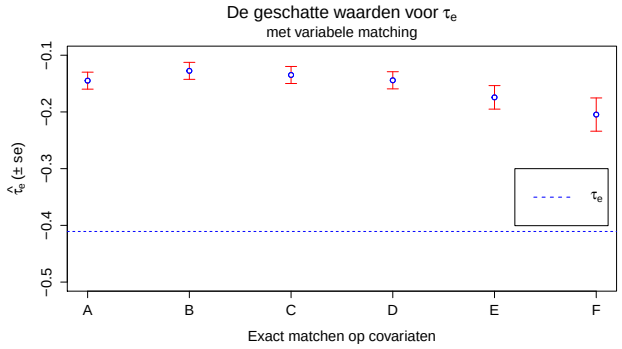
4.2.2 Variabele matching

In de vorige paragraaf hebben we de invloeden bekeken van Nearest Neighbour matching met het gebruik van een caliper. Een andere mogelijke keuze die in de match strategie kan worden gemaakt is variabele matching, waarvoor aanvullend wordt gekozen in Artikel I. In paragraaf 1.3.2 is vermeld dat voor variabele matching kan worden gekozen, wanneer een aantal covariaten van extra belang worden geacht. In deze paragraaf bekijken we daarom ter illustratie de keuze van variabele matching, ondanks dat dit geen discussiepunt van Groenendijk en van Delft is.



Figuur 4.8: De geschatte waarden voor τ_e met variabele matching. De situatie 'geen' duidt aan dat er geen variabele matching is uitgevoerd. **A**: match op X_2, X_3, X_5 , **B**: match op X_1, X_8, X_9, **C**: match op X_4, X_7, X_9, **D**: match op X_2, X_3, X_{10}. Calipers $(X_1, X_2, X_3, X_4, X_5, X_7, X_8, X_9, X_{10}) = (0, 0, 0, 1, 0, 0.5, 1, 1, 0.5)$.

Naar aanleiding van de eerdere uitkomsten, wanneer covariaten worden vergeten, kan geconcludeerd worden dat in deze simulatie de covariaten X_2, X_3, X_5 en X_{10} meer van betekenis zijn dan de andere covariaten. De vraag is of de schatting verbetert wanneer een exacte match op een combinatie van deze covariaten wordt gemaakt in tegenstelling tot een combinatie waar de covariaten minder belangrijk worden geacht, zie Figuur 4.3. Zoals te zien is in de figuur is het dus mogelijk dat de schatting wat slechter wordt, omdat je op een aantal variabelen matcht die niet van groot belang zijn.(C) Het idee dat de schatting mogelijk kan verbeteren als er op een aantal belangrijke covariaten wordt gematcht, is in dit geval bevestigd. Mogelijk is het matchen op een aantal covariaten nog meer van nut wanneer bijvoorbeeld een kwadratische term in het model zit, welke lineair wordt meegenomen (Figuur 4.9). Wanneer we op deze variabele gaan matchen verwachten we dat daardoor de schatting weer wat beter wordt, alhoewel de fit van de regressiecoëfficiëten nog steeds minder zal zijn, zoals we eerder hebben kunnen zien.



Figuur 4.9: De geschatte waarden voor τ_e met vergelijking variabele matching, als X_2 kwadratisch is gesimuleerd. **A**: X_2 kwadratisch in model meegenomen, **B**: X_2 lineair meegenomen, **C**: X_2 lineair en variabele match, caliper $\theta = 1$, **D**: Zelfde als 'C', caliper $\theta = 0.5$, **E**: Zelfde als 'C', caliper $\theta = 0$, **F**: Zelfde als 'C', maar nog extra variabele match op X_3 en X_{10} met calipers $\lambda = 0$ en $\rho = 0.5$.

Situatie 'F' uit Figuur 4.9 is vergelijkbaar aan situatie 'D' uit Figuur 4.8, maar verschilt omdat X_2 in Figuur 4.9 kwadratisch is gesimuleerd. Door op X_2 te matchen als deze lineair in het model zit, maar kwadratisch is gesimuleerd, zien we dat de schatting erop vooruit gaat als we gaan matchen op deze variabele. Echter doordat de regressiecoëfficiënten minder goed geschat kunnen worden, waardoor de propensity score minder naar waarheid is, blijft de verbetering beperkt. De schatting

is daardoor dus slechter dan in Figuur 4.8.

In Figuur 4.9 komt naar voren dat wanneer de caliper verkleind wordt de schatting beter wordt. Een gevaar kan zijn waneer de calipers te klein worden dat er te weinig matches kunnen worden gemaakt wat de schatting niet ten goede zal komen. In zowel Figuur 4.8 als Figuur 4.9 komt dan ook naar voren dat bij de kleine calipers de standaardfout vergroot. Merk op dat wanneer de calipers groter worden gemaakt, de verschillen tussen de schattingen in Figuur 4.8 kleiner zullen zijn.

Wat herhaaldelijk terugkomt, is dat er veel keuzes kunnen worden gemaakt in de matching strategie, waardoor de schatting kan worden verbeterd. Wel zitten er haken en ogen aan, waardoor voorzichtigheid en een goede argumentatie geboden is. Zo kan een verkeerd gekozen caliper of het matchen op de verkeerde variabelen de schatting mogelijk niet ten goede komen. In Artikel I wordt voor Nearest Neighbour matching gekozen met gebruik van een caliper, waarbij aanvullend variabele matching wordt gebruikt. Geen van deze keuzes wordt echter beargumenteerd. Voor de keuze van de drie covariaten waar extra op wordt gematcht is bijvoorbeeld geen motivatie gegeven. Wanneer de lezer van Artikel I naar Tabel 1 (Appendix A) kijkt, waarin de uitkomsten van het regressiemodel zijn gepresenteerd, lijken de gekozen covariaten significant te zijn volgens de uitgevoerde toets. Maar waarom precies deze covariaten zijn gekozen en niet een combinatie van andere 'significante' covariaten blijft onduidelijk.

4.3 Generaliseerbaarheid van de resultaten

4.3.1 'De eigenlijke toetsing'

> *"Het artikel van Rosenbaum en Rubin vermeldde in de titel reeds dat het ging over sampling methods. Bij Wermink et al. was echter geen sprake van sampling: data van de complete populaties van veroordeelden waren beschikbaar. Wanneer in zekere zin de complete populatie onderzocht is, is het hanteren van een statistische significantietoets vreemd en onnodig. In de statistiek spreken we met name over significantie om aan te geven in welke mate resultaten, gevonden op basis van steekproeven, van betekenis geacht mogen worden voor de hele populatie."*
>
> — Groenendijk & van Delft (2013a), p.62.

In het algemeen verkrijgen we niet de data van ieder individu uit een populatie waar we een uitspraak over willen doen. Als oplossing nemen we bijvoorbeeld een willekeurige trekking uit de populatie en proberen de gevonden resultaten van deze steekproef uit te breiden naar de gehele populatie. Wat Groenendijk en van Delft problematisch vinden is dat de gebruikte data uit Artikel I geen steekproef is uit een bestaande populatie. We moeten daardoor een imaginaire populatie, die als het ware de eeuwigheid representeert, accepteren. Groenendijk en van Delft zijn het hiermee oneens en vinden dat je de resultaten die je vindt enkel kunt betrekken op de data die er bestaat, de veroordeelden uit 1997, en niet over ieder die ooit in aanmerking kan komen voor een veroordeling.

We kunnen ons daarom het volgende afvragen: Is het reëel om aan te nemen dat de veroordeelden van nu zich net zo gedragen als de veroordeelden uit 1997, waarvan de data is gebruikt? Ofwel, mogen we de resultaten die we vinden uitbreiden naar een grotere populatie, naar alle "criminelen" die ooit in aanmerking komen voor een werk- of gevangenisstraf?

Het gaat er hier niet zozeer om of de toetsing die Wermink et al. toepassen wel juist is, maar eerder in algemene zin of statistiek in deze vorm gebruikt mag worden om uitspraken te doen. Het probleem is dat wanneer we niet aannemen dat er een imaginaire populatie bestaat, statistiek enkel nog toegepast kan worden wanneer bijvoorbeeld een willekeurige steekproef wordt genomen

uit een bestaande populatie. In veel studies is dit niet het geval, waardoor statistiek dan niet meer gebruikt kan worden.

Naar mijn mening kunnen we statistisch onderzoek als hulpmiddel gebruiken voor gevolgtrekkingen. Wanneer dit onderzoek zou aanwijzen dat er geen reden is om aan te nemen dat werkstraf voor minder recidive zorgt, is onze interesse naar het onderzoek waarschijnlijk snel verdwenen en houden we ons bij de standaarden die er al waren. Echter, een land wil innoveren en verbeteren. Om dit te kunnen doen moeten we een stap durven zetten. Denk eens aan Einstein, die pas bij de aanname van een eindig heelal tot resultaten kon komen. Ondanks dat we misschien nog niet overtuigd zijn van de eindigheid van het heelal heeft hij een mogelijkheid gecreëerd tot ontwikkeling op dit gebied.

Het gevonden resultaat ten gunste van de werkstraf daagt ons uit om een stap vooruit te durven zetten. We kunnen als land niet alleen miljoenen besparen, we kunnen ons ook voorbereiden op de nieuwste ontwikkelingen, zoals het gebruik van een enkelband.

> *"Het huidige gebrek aan kennis over bedoelde en onbedoelde gevolgen van straffen bij juristen en criminologen is in onze ogen zorgelijk en dient te worden gelenigd. Met ons onderzoek hebben we een empirische bijdrage willen leveren aan de vermeerdering van kennis over de effecten van straf, en meer in het bijzonder aan die over het effect van werkstraf."*
>
> — Wermink et al. (2013), p.67.

Ten aanzien van bovenstaand citaat van Wermink et al. sluiten zij zich aan bij de mogelijkheid tot ontwikkeling. Naar mijn mening moet er, bij het trekken van de conclusie, vermeld worden dat de auteurs aannemen dat er een steekproef wordt genomen uit een imaginaire populatie. Het gevaar is namelijk dat media e.d. de conclusies klakkeloos overnemen, zonder zich bewust te zijn van mogelijke beperkingen in de generaliseerbaarheid.

4.3.2 'Correlatie versus causaliteit'

> *"De onderzoekers claimen een correlatie te hebben aangetoond tussen de soort straf na een eerste veroordeling en het aantal malen dat dezelfde persoon later opnieuw opgepakt en veroordeeld werd. Correlatie toont geen causaliteit aan."*
>
> – Groenendijk & van Delft (2013a), p.63.

Het discussiepunt dat hier wordt aangesneden is van belang voor de interpretatie van de verkregen resultaten, omdat men de uitkomsten wil gebruiken als interventie. Er zijn talloze voorbeelden te bedenken van correlerende factoren welke niet causaal geïnterpreteerd kunnen worden, omdat er een causale factor is voor beide. Groenendijk en van Delft geven een voorbeeld van 'sociale intelligentie', wat een causale factor voor zowel de straftoewijzing als het aantal recidive kan zijn.

> *"Veroordeelden die in staat zijn om de rechter de indruk te geven dat ze hun leven kunnen en willen verbeteren en daarom een taakstraf krijgen opgelegd, zullen om twee redenen minder veroordeeld worden voor nieuwe vergrijpen. Ofwel omdat de indruk klopte en men geen nieuwe vergrijpen pleegde, ofwel omdat de doortraptheid die hen hielp om de rechter om de tuin te leiden ook in hun voordeel werkt voor wat betreft de kans om aangehouden, in staat van beschuldiging gesteld en veroordeeld te worden voor nieuwe vergrijpen."*
>
> – Groenendijk & van Delft (2013a), p.63.

Correlatie versus causaliteit is een veelbesproken en ingewikkeld probleem, vooral uit filosofisch en psychologisch oogpunt. De vraag is en blijft hoe en of we de verkregen resultaten causaal mogen interpreteren.

> *"Can we ever estimate the causal effect? The answer is: sometimes. In particular, random assignment to treatment makes it possible to estimate the causal effect."*
> – Wasserman (2004), p.331.

Gezien de moeilijkheid van het vinden van de juiste verstorende factoren blijft de 'terugkeer' naar een gerandomiseerd experiment met behulp van de propensity score methode moeilijk, waardoor een gevonden effect niet causaal hoeft te zijn. Dit probleem is in Artikel I ook erkend.

> *"Onze resultaten zijn niet gebaseerd op een experiment met volledige random toewijzing van werk- en gevangenisstraffen en daarom blijft voorzichtigheid geboden bij het interpreteren van de gevonden verschillen in termen van causaliteit. Het risico blijft immers bestaan dat ondanks de matching op geobserveerde variabelen een of meer niet-gemeten variabelen verantwoordelijk zijn voor de gevonden verschillen in recidive."*
> – Wermink et al. (2009), p.224.

Kunnen we ooit de gevonden resultaten gebruiken als interventie?

> *"One single observational study is not, by itself, strong evidence. Remember that when you read the newspaper."*
> – Wasserman (2004), p.338.

Echter, uit Artikel I citeren we het volgende:

> *"Voor Nederland maar ook internationaal, is dit de* ***eerste*** *grootschalige matching studie, gebaseerd op observationele data over een langere periode, die de recidive na werkstraffen vergelijkt met die na gevangenisstraffen."*
> – Wermink et al. (2009), p.223.

De moeilijkheid van 'de generaliseerbaarheid van de resultaten' ligt misschien niet eens zozeer bij bijvoorbeeld de aannames die we moeten doen, maar bij het feit dat niet iedereen zich bewust is van de beschreven problemen. Ik vermoed dat dit precies is waar Groenendijk en van Delft moeite mee hebben; dat enkel de conclusie 'minder recidive' wordt overgenomen zonder dat benadrukt wordt dat dit nog niet het einde van de zoektocht is, maar slechts het begin.

Conclusie

In deze scriptie is de gebruikte statistische methodiek van het artikel 'Recidive na werkstraffen en na gevangenisstraffen' bestudeerd. Met behulp van de propensity score methode wordt geprobeerd antwoord te geven op de vraag: 'In hoeverre zijn werkstraffen een goed alternatief voor gevangenisstraffen in relatie tot de recidive van de gestraften na afloop van hun straf?' Er wordt daarbij tevens antwoord gegeven op de vraag van mevrouw Helder hoe we personen moeten vergelijken, terwijl zij eigenlijk niet vergelijkbaar zijn, wat geciteerd is in de introductie.

We hebben gezien dat ondanks de selecte toewijzing tot een experimentele groep, of controlegroep er een oplossing kan worden gevonden, waardoor toch het effect van een werkstraf t.a.v. recidive kan worden bepaald. Dat de methode intuïtief voldoet, komt naar voren in de conditionering op de covariaten, de variabelen die het experiment als het ware verstoren. Onder deze conditionering en het accepteren van twee aannames bestaat de mogelijkheid het gewenste effect te bepalen. Daarbij hebben we gezien dat het volstaat om op een functie van de covariaten te conditioneren, de propensity score.

Wat opvalt aan de beschreven methode is dat deze een wirwar is van verschillende mogelijkheden, waardoor er veel aan de onderzoekers zelf kan worden overgelaten. Dit loopt uiteen van de keuze voor het te bepalen effect tot welke (gespecificeerde) matching strategie moet worden gebruikt. Echter, uiteindelijk hebben de onderzoekers allemaal hetzelfde doel: het gewenste effect bepalen. De verschillende mogelijkheden binnen de propensity score methode zijn er dan ook niet voor niets, denk aan variabele matching waarmee de zuiverheid van de schatting mogelijk verbeterd kan worden. Naar mijn inzicht kunnen sommige mogelijkheden daardoor een verrijking zijn binnen de methode, mits een behoorlijke argumentatie gegeven is. Wanneer een verantwoorde keuze kan worden gemaakt, welke bijvoorbeeld het beste past bij de gebruikte data set, kan het de uiteindelijke schatting ten goede komen. Naar de beste keuzes is al veel onderzoek gedaan, bijvoorbeeld over de optimale caliper grootte in Austin (2010a), en dit zal zich in de toekomst hopelijk nog verder uitbreiden.

Naast de bespreking van de propensity score methode is het geschreven discussiestuk van Groenendijk en van Delft bestudeerd. In de discussie in Hoofdstuk 4 is er aan de hand van de kritiek van Groenendijk en van Delft geconcludeerd dat er een beperkte wijze van presentatie is. Het geven van een fit van het regressiemodel is een belangrijke ontbrekende schakel in de zoektocht naar het te bepalen effect. Wanneer het verkregen regressiemodel niet fit, kan dit zeer van invloed zijn op de verkregen schatting. Daarnaast is de moeilijkheid erkend van het vinden van de juiste covariaten. Echter, juist gezien deze moeilijkheid is de beperkte motivatie over de covariaten zorgelijk en onvolledig. Hoewel de auteurs van Artikel I stellen dat zij een sensitiviteitsanalyse doen naar mogelijk ongeobserveerde covariaten is dit niet wat er daadwerkelijk is gebeurd. Om nog een ander voorbeeld van de beperkte presentatie te geven, kunnen we denken aan een ontbrekende motivatie voor het gekozen te bepalen effect. Sterker nog, of τ of τ_e wordt bepaald, wordt überhaupt niet genoemd in het gehele artikel. In de conclusie ontbreekt dan ook dat het gaat om het effect van een werkstraf op de werkgestraften. Daarnaast had in de conclusie meer aandacht besteed moeten worden aan het feit dat bijna drie kwart van de werkgestraften niet kon worden gebruikt, wat de generaliseerbaarheid van de resultaten kan beperken. Ook is er geen motivatie gegeven voor de gekozen matching strategie en de keuzes die hierbij zijn gemaakt, zoals de caliper en variabele matching. De kritiek van Groenendijk en van Delft op het gebruik van een caliper is echter onterecht, want op grond van paragraaf 4.2.1 lijkt de keuze van een caliper zeker gerechtvaardigd. De terechte kritiek op Artikel I betreft dus de beperkte wijze van presenteren wat, wanneer er kennis is van de statistische methodiek, een onbevredigd gevoel geeft.

Er zijn ook meer filosofische discussiepunten besproken, waar niet zozeer een criterium 'goed' of 'fout' voor bestaat. Dit betreft bijvoorbeeld de bespreking van het uitgangspunt dat er een steekproef wordt genomen uit een grotere populatie, terwijl dit eigenlijk niet zo is. Daarnaast wordt bekritiseerd dat correlatie niet per se causaliteit aantoont. Wermink et al. hebben in de conclusie van Artikel I laten zien dat zij kennis hebben van de moeilijkheden rond causaliteit, dus naar mijn mening is de aanval van Groenendijk en van Delft over dit punt onterecht.

In het algemeen vind ik de uitingen van Groenendijk en van Delft op z'n minst ongepast. Bekijk daartoe de volgende citaten:

> *"We besloten serieus werk te gaan maken van deze uiting van beroerd wetenschappelijk niveau."*
> –Groenendijk & van Delft (2013b).

> *"De auteurs wijzen alle kritiek af: niet alleen door een zwakke argumentatie maar ook door eenvoudigweg op de belangrijkste punten van kritiek niet in te gaan. Het weerwoord toont daarom een tekort aan integriteit en competentie. Dit is funest voor het politieke debat over bestrijding van de criminaliteit en funest voor het aanzien van de wetenschap."*
> –Groenendijk & van Delft (2013b).

We hebben gezien dat Wermink et al. hebben geprobeerd een bijdrage te leveren aan de maatschappij. Hoewel de presentatie onvolledig is en verbeterd kan worden, is de gegeven kritiek van Groenendijk en van Delft op aspecten van de gebruikte methode een aantal maal onterecht en onjuist. Een voorbeeld is dat daders die niet konden worden gekoppeld wel worden meegenomen in het logistisch regressiemodel. Op basis daarvan vind ik het ongepast de auteurs van Artikel I op deze wijze af te branden en kan er beter een helpende hand worden aangeboden voor waar en hoe er volgens hen nog verbetering nodig is.

> *"Niet alleen betreft het hier maatschappelijk beslist relevant onderzoek: de claims worden op grote schaal klakkeloos overgenomen in media en politiek."*
> –Groenendijk & van Delft (2013b).

Ik heb begrip voor de zorgen van Groenendijk en van Delft wat betreft de overname van het resultaat. Dat de claims klakkeloos worden overgenomen is bijvoorbeeld bevestigd door de dialoog, welke is gegeven in de inleiding van deze scriptie. Vanuit dit oogpunt begrijp ik in ieder geval de discussiepunten over causaliteit en de toetsing, omdat men hier in het algemeen bij het opnemen van de resultaten geen of weinig kennis van krijgt. In paragraaf 4.3.2 hebben we geleerd dat één enkel onderzoek nog geen rotsvast bewijs levert, dus dat we niet direct alles moeten geloven wat in de kranten te lezen is.

In de strijd over dit gevoerde onderzoek naar recidive na een werk- of gevangenisstraf lijkt tot dusver niet echt een winnaar of verliezer te zijn. Het bestudeerde onderzoek toont wellicht aan dat er minder recidive na een werkstraf is, mits aan de beschreven tekortkomingen kan worden voldaan. Maar laten we benadrukken dat onder andere in termen van causaliteit dit enkele onderzoek nog geen bewijs is van minder recidive ten gunste van de werkstraf. We eindigen daarom met mijn favoriete citaat, welke voor alle partijen van toepassing is:

> *"Success is not final, failure is not fatal: it is the courage to continue that counts."*
> –Winston Churchill.

Referenties

Wermink, H., Blokland, A., Nieuwbeerta, P., Tollenaar, N. (2009), Recidive na werkstraffen en na gevangenisstraffen, *Tijdschrift voor Criminologie*, **51** (3), 211-227.

Rosenbaum, P., Rubin, D. (1983), The central role of the propensity score in observational studies for causal effects, *Biometrika*, **70**, 41-55.

Rosenbaum, P., Rubin, D. (1985), Constructing a control group using multivariate matched sampling methods that incorporate the propensity score, *The American Statistician*, **39**, 33-38.

Caliendo, M., Kopeinig, S. (2005), Some practical guidance for the implementation of propensity score matching, *IZA Discussion Paper* NO. 1588.

Bijma, F., Jonker, M., van der Vaart, A. (2013), *Inleiding in de statistiek*, Epsilon Uitgaven (Utrecht).

Austin, P.C. (2010a), Optimal caliper widths for propensity-score matching when estimating differences in mean and differences in proportions in observational studies, *Pharmaceutical Statistics*, **10**, 150-161.

Austin, P.C. (2010b), A data-generation process for data with specified risk differences or numbers needed to treat, *Communications in Statistics - Simulation and Computation*,Toronto.

Austin, P.C. (2011), An introduction to propensity score methods for reducing the effects of confounding in observational studies, *Multivariate Behavioral Research*, **46**, 399-424.

Rice, A. (2007), *Mathematical statistics and data analysis*, Brooks/Cole (Belmont).

Hosmer, D.W., Lemeshow, S. (2000), *Applied logistic regression*, Wiley&Sons.

Abadie, A., Imbens, G. (2012), *Matching on the estimated propensity score*, NBER Working Paper No. 15301.

Nieuwbeerta, P., Nagin, D., Blokland, A. (2007), Het meten van effecten van gevangenisstraf op crimineel gedrag in een niet-experimentele studie, *Mens & Maatschappij*, **82**, (3), 283.

Brookhart, M.A., Schneeweiss, S., Rothman, K.J., Glynn, R.J., Avorn, J., Stürmer, T. (2006), Variable selection for propensity score models, *American Journal of Epidemiology*, **163**, 1149-1156.

Wasserman, L. (2004), *All of statistics: a concise course in statistical inference*, New York: Springer, 327-343

Stürmer, T., Rothman, K., Avorn, J., Glynn, R. (2010), Treatment effects in the presence of unmeasured confounding: dealing with observations in the tails of the propensity score distribution - A simulation study, *American Journal of Epidemiology.*

Agresti, A. (2007), *An introduction to categorical data analysis*, Wiley&Sons, Inc., Publication, Florida, 66-67,75.

Sekhon, J.S. (2013), Multivariate and propensity score matching with balance optimization. Beschikbaar op 28 oktober, 2013, van *http://cran.r-project.org/web/packages/Matching/Matching.pdf*, 16-25.

Wermink, H., Blokland, A., Nieuwbeerta, P., Tollenaar, N. (2013), De betere stuurlui roeien (ook) met de riemen die ze hebben, *Tijdschrift voor Criminologie*, **55**, (1), 66-74.

Groenendijk, F., van Delft, A. (2013a), Recidive, werkstraf en gevangenisstraf: een kritische bespreking, *Tijdschrift voor Criminologie*, **55**, (1), 59-65.

Groenendijk, F., van Delft, A. (2013b), Wermink, Blokland, Nieuwbeerta en Tollenaar duiken. Beschikbaar op 30 maart, 2013, van *http://www.keizersenkleren.nl/?p=386*.

A Tabellen

	B	S.E.	Sign.	Exp(B)
Persoonskenmerken				
vrouw	0,41	0,07	***	1,50
leeftijd(/10)	-1,29	0,22	***	0,28
leeftijd kwadraat(/10)	0,21	0,03	***	1,23
geboren in buitenland	-1,61	0,05	***	0,20
Uitgangszaak				
aantal feiten in uitgangszaak	0,08	0,03	***	1,08
ernst uitgangszaak	0,11	0,02	***	1,12
Type delict				
overige wetten	(ref.)	(ref.)		(ref.)
huis, lokaalvredebreuk	-1,41	0,86	*	0,24
openlijk geweld	-0,03	0,17		0,97
wederspannigheid	-0,25	0,50		0,78
overig openbare orde	-0,32	0,24		0,73
agressief ernstig letsel	-0,44	0,18	***	0,64
belediging	-2,09	1,30		0,12
bedreiging, mishandeling	-0,36	0,14	***	0,69
mishandeling zwaar	-0,60	0,24	***	0,55
brandstichting levensgevaar	-1,98	0,36	***	0,14
overige zaken, dieren	-0,81	0,22	***	0,45
zedengeweld	-1,00	0,32	***	0,37
ontucht jonger 16 jaar	-0,63	0,23	***	0,53
bevoordeling	-0,09	0,12		0,92
wegnemen	-1,26	0,13	***	0,28
gekwal. diefstal	-1,27	0,13	***	0,28
vermogensgeweld	-1,61	0,20	***	0,20
WvW	0,19	0,13		1,21
Opiumwet	-0,59	0,11	***	0,56
Vuurwapenwet	-0,17	0,18		0,84
Criminele geschiedenis				
aantal vermogensdelicten in afgelopen jaar	-0,36	0,07	***	0,70
aantal geweld in afgelopen jaar	-0,28	0,14	**	0,76
aantal overig in afgelopen jaar	0,03	0,08		1,03
aantal vermogen in laatste 10 jaar	-0,03	0,03		0,97
aantal geweld in laatste 10 jaar	0,03	0,06		0,97
aantal overig in laatste 10 jaar	-0,09	0,03	***	0,91
Constante	3,41	0,35	***	

* $p < 0,10$; ** $p < 0,05$; *** $p < 0,01$

Tabel 1: Resultaten logistisch regressiemodel van Artikel I (Tabel 1).

	Gemiddelde experimentele groep (N=2.123)	**Gemiddelde controlegroep (N=2.123)**	**Absoluut verschil**	**t-stat**	**D**
Vrouw	0,125	0,125	0,000	0,000	0,0
Leeftijd(/10)	2,723	2,723	0,000	0,000	0,0
Leeftijd kwadraat(/10)	9,003	9,011	-0,008	-0,052	-0,2
Geboren in buitenland	0,554	0,555	-0,001	-0,062	-0,2
Aantal feiten in uitgangszaak	1,581	1,574	0,007	0,253	0,8
Strafdreiging	4,315	4,181	0,134	1,547	4,7
Aantal feiten afgelopen jaar					
vermogen	0,057	0,049	0,008	0,907	2,8
geweld	0,021	0,018	0,002	0,507	1,6
overig	0,060	0,069	-0,009	-0,965	-3,0
Aantal feiten afgelopen tien jaar					
vermogen	0,314	0,329	-0,015	-0,545	-1,7
geweld	0,097	0,104	-0,007	-0,614	-1,9
overig	0,352	0,389	-0,036	-1,394	-4,3

* $p < 0,10$; ** $p < 0,05$; *** $p < 0,01$

Tabel 2: Resultaten na het matchen van Artikel I (Tabel 2).

	Gemiddelde experimentele groep (N=2.123)	**Gemiddelde controlegroep (N=2.123)**	**Absoluut verschil**	**t-stat**	**Sign.**	**Relatief verschil**
1 jaar						
totaal	0,273	0,683	-0,410	-3,229	***	-0,60
vermogen	0,132	0,398	-0,266	-2,404	**	-0,67
geweld	0,044	0,109	-0,065	-2,255	**	-0,60
overig	0,097	0,175	-0,079	-3,698	***	-0,45
3 jaar						
totaal	0,292	0,575	-0,283	-5,401	***	-0,49
vermogen	0,129	0,292	-0,162	-5,418	***	-0,56
geweld	0,052	0,100	-0,049	-2,755	***	-0,48
overig	0,111	0,183	-0,072	-2,048	**	-0,40

* $p < 0,10$; ** $p < 0,05$; *** $p < 0,01$

Tabel 3: Resultaten propensity score matching van Artikel I (Tabel 3).

B Appendix

Stelling 1. Voor gebeurtenissen A, B en C, waarbij $P(B|C) > 0$, geldt

$$P(A|B,C) = \frac{P(A,B|C)}{P(B|C)}.$$

Bewijs. De conditionele kans van A gegeven B is gedefinieerd als

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

Substitueer nu $(B \cap C)$ voor B, dan volgt

$$\begin{aligned} P(A|B \cap C) &= \frac{P(A \cap (B \cap C))}{P(B \cap C)} \\ &= \frac{P((A \cap B) \cap C)}{P(B \cap C)}. \end{aligned}$$

Twee keer toepassen van de definitie van de conditionele kans geeft direct

$$P(A|B,C) = \frac{P(A,B|C)P(C)}{P(B|C)P(C)}.$$

□

Stelling 2. (Tower Property of Conditional Expectation) Zij X, Y en Z stochasten, dan geldt

$$\mathbb{E}[X|Y] = \mathbb{E}[\mathbb{E}(X|Y,Z)|Y].$$

Bewijs. We zullen dit bewijzen voor het discrete geval. Het continue geval kan op gelijke wijze bewezen worden. Gegeven $Y = y$ is de kans op $\mathbb{E}[X|Y,Z] = \mathbb{E}[X|Y = y, Z = z]$ gelijk aan $P(Z = z|Y = y)$, omdat enkel z kan variëren over zijn mogelijke waardes. Daarom kunnen we schrijven:

$$\begin{aligned} \mathbb{E}[\mathbb{E}(X|Y,Z)|Y = y] &= \sum_z \mathbb{E}[X|Y = y, Z = z]P(Z = z|Y = y) \\ &= \sum_z \sum_x xP(X = x|Y = y, Z = z)P(Z = z|Y = y) \end{aligned}$$

Nu volgt:

$$\begin{aligned} \mathbb{E}[\mathbb{E}(X|Y,Z)|Y = y] &= \sum_{x,z} x \frac{P(X = x, Y = y, Z = z)}{P(Y = y, Z = z)} \cdot \frac{P(Z = z, Y = y)}{P(Y = y)} \\ &= \sum_{x,z} x \frac{P(X = x, Y = y, Z = z)}{P(Y = y)} \\ &= \sum_x x \frac{P(X = x, Y = y)}{P(Y = y)} \\ &= \sum_x xP(X = x|Y = y) \\ &= \mathbb{E}[X|Y = y]. \end{aligned}$$

□