Centraal Bureau
voor de Statistiek

Rapport

Schoolverlatersonderzoek 2017: schattingen voor kleine deelpopulaties van MBO schoolverlaters

Harm Jan Boonstra

CBS Heerlen
CBS-weg 11
6412 EX Heerlen
Postbus 4481
6401 CZ Heerlen
T +31 45 570 60 00

projectnummer 303221
Sector Procesontwikkeling en Methodologie (BPM)
06-04-2018

Met het Schoolverlatersonderzoek (SVO) worden schattingen gemaakt van variabelen die betrekking hebben op de aansluiting van opleiding bij werk. Dit rapport beschrijft de manier waarop zeer gedetailleerde schattingen zijn gemaakt, naar onderwijsinstelling en opleidingsrichting, voor een populatie van MBO schoolverlaters. Evenals voor het SVO 2016 is deze populatie integraal benaderd, maar de respons op SVO 2017 ligt lager. Een binomiaal multilevel model wordt gebruikt om op basis van de respons voor een aantal doelkenmerken schattingen te maken voor de vele deelpopulaties. Dit methodologisch rapport is een update van Boonstra (2017b) over de gedetailleerde schattingen op basis van SVO 2016. Samenvattend zijn de belangrijkste veranderingen:

- personen van 50 jaar en ouder (een relatief kleine groep) behoren nu wel tot de doelpopulatie
- de belangrijkste deelpopulatie is gewijzigd van de werkzamen naar de niet-studerende werkzamen. Schattingen van de 'aansluitingsvariabelen' worden gepercenteerd op deze deelpopulatie.
- de schattingen worden niet meer uitgesplitst naar de meest gedetailleerde opleidingsclassificatie (CREBO) maar naar een iets ingedikt niveau (beroep)
- er worden extra tabellen geschat met uitsplitsingen naar 4 MBO niveaus
- het model is iets vereenvoudigd in verband met de lagere respons, de kleinere deelpopulatie van niet-studerende werkzamen en het iets lagere detailniveau van schattingen:
  - een klein aantal covariaten zijn ingedikt of uit het model weggelaten
  - drie van de vijf gemodelleerde (random) effect termen in het model hebben een iets lager detailniveau
- tegelijkertijd is het model iets uitgebreid qua afhankelijkheid van MBO niveau, omdat een aantal tabellen hiernaar uitgesplitst worden

Vooral vanwege het buiten beschouwing laten van de studerende werkzamen zijn de schattingen niet goed vergelijkbaar met die van SVO 2016.

1 Inleiding

Het Schoolverlatersonderzoek (SVO) heeft als doel inzicht te krijgen in de aansluiting van school naar werk of naar vervolgopleiding. CBS doet dit onderzoek in opdracht van het Ministerie van Onderwijs, Cultuur en Wetenschappen (OCW) en werkt daarbij samen met het Researchcentrum voor Onderwijs en Arbeidsmarkt (ROA).

De doelpopulatie van het Schoolverlatersonderzoek 2017 is een deelverzameling van personen die in het schooljaar 2015/2016 zijn geslaagd voor MBO, VMBO, HAVO of VWO, of het onderwijs hebben verlaten zonder in dat schooljaar een diploma te hebben behaald (van Berkel, 2016). Alleen personen die in Nederland wonen en deel uitmaken van een niet-institutioneel huishouden behoren tot de doelpopulatie. Er is geen afbakening op basis van leeftijd. Dit is een wijziging ten opzichte van SVO 2016 waar voor de analyse personen boven de 50 jaar werden uitgesloten.

Deze studie naar gedetailleerde schattingen wordt toegepast op de deelpopulatie van MBO schoolverlaters die geslaagd zijn. Deze populatie, binnen het SVO ook aangeduid met groep 1, is integraal benaderd. Uiteindelijk is van $n = 27317$ personen een bruikbare respons ontvangen, wat neerkomt op 19,9% van de doelpopulatie, die uit $N = 137441$ personen bestaat. Ondanks de beperking tot personen jonger dan 50 jaar lagen al deze cijfers voor SVO 2016 hoger: $n = 37252$, $N = 142339$ en daarmee een responsfractie van 26,2%. Overigens is de 50+ groep relatief klein met 2561 personen in de SVO 2017 populatie waarvan 1087 bruikbare responsen.

OCW gebruikt de informatie uit het SVO op een zeer gedetailleerd niveau, namelijk op het niveau van instelling gekruist met opleiding. Om op dat niveau schattingen van de belangrijkste doelvariabelen te kunnen maken is een weging niet toereikend omdat er voor de meeste combinaties te weinig en soms zelfs helemaal geen respondenten zijn. Daarom is een modelgebaseerde methode ontwikkeld waarmee de gewenste domeinschattingen worden gemaakt. Omdat de doelkenmerken binaire (0/1) variabelen zijn wordt een binomiaal multilevel model gebruikt. Multilevel modellen worden vaak gebruikt voor het modelleren van data op meerdere niveaus zoals in dit geval instelling en opleiding, en worden binnen de officiële statistiek gebruikt voor het maken van gedetailleerde schattingen, ook aangeduid met kleine-domeinschattingen. Zie Rao en Molina (2015) voor een overzicht van kleine-domeinschattingsmethoden en, bijvoorbeeld, Gelman en Hill (2007) voor meer informatie over multilevel modellen.

De deelpopulaties waarvoor schattingen gemaakt worden zijn op populatieniveau vaak al klein. Dit betekent dat er deelpopulaties zijn zonder respons maar ook deelpopulaties die volledig zijn waargenomen. Het multilevel model wordt op de gehele respons geschat en vervolgens wordt voor alle non-respondenten bijgeschat. Hierbij wordt rekening gehouden met verschillen tussen respondenten en non-respondenten naar achtergrondkenmerken die bekend zijn voor de hele populatie.

Om de nauwkeurigheid van de schattingen te kunnen beoordelen worden ook standaardfouten en 95% intervallen berekend. Deze laten zien dat de modelgebaseerde schattingen gemiddeld een stuk nauwkeuriger zijn dan directe schattingen op basis van de respondenten per cel. Toch zijn er nog veel domeinen waarvoor ook de

modelgebaseerde schattingen grote standaardfouten hebben. Op basis van de (relatieve) omvang van de standaardfouten en onthullingsrisico's kan worden besloten welke schattingen worden onderdrukt bij publicatie.

De rest van deze nota is als volgt opgebouwd. In paragraaf 2 worden de doelvariabelen en de deelpopulaties beschreven. In paragraaf 3 wordt de schattingsmethode uiteengezet. Daarna volgen enkele resulaten in paragraaf 4 en conclusies in paragraaf 5. In de bijlagen zijn meer details over het model en de gebruikte achtergrondvariabelen te vinden, en verschillende grafieken waarin directe en modelschattingen worden vergeleken.

2 Doelvariabelen, deelpopulaties en te schatten tabellen

De doelvariabelen waarvoor gedetailleerde schattingen worden gemaakt zijn

0. $werk$: werkzaam volgens de internationale definitie en niet studerend
1. $aansluiting$: voldoende of goede aansluiting tussen de gevolgde MBO-opleiding en de huidige functie
2. $niveau$: het vereiste niveau voor de huidige functie is gelijk aan of hoger dan dat van de gevolgde MBO-opleiding
3. $richting$: de gevolgde MBO-opleiding is qua richting verwant/gelijk aan de voor de huidige functie benodigde opleiding

Een wijziging ten opzichte van SVO 2016 is dat in de definitie van de werkzame deelpopulatie nu studerende werkzamen worden uitgesloten. Variabelen $aansluiting$, $niveau$ en $richting$ zijn alleen van toepassing op werkzame personen. De schattingen voor deze variabelen worden voor SVO 2017 gepercenteerd op de deelpopulatie van niet-studerende werkzamen. Alle vier doelvariabelen zijn binair met mogelijke waarden wel/niet, gecodeerd als 1/0. De doelvariabelen zijn alleen bekend voor de respondenten. Variabele $werk$ is bekend voor alle respondenten, terwijl voor de andere drie variabelen een klein aantal waarden bij (niet-studerende) werkzame respondenten ontbreken: 26 voor $aansluiting$, 79 voor $niveau$ en 213 voor $richting$. Voor de gehanteerde modelgebaseerde schattingsmethode is dit geen groot probleem, omdat per variabele alle ontbrekende waarden worden bijgeschat.

Van de $n = 27317$ respondenten zijn 23390 werkzaam volgens de internationale definitie, waarvan 17272 niet studeren volgens de daarvoor gehanteerde definitie. Door de beperktere afbakening en de lagere respons is het aantal waarnemingen waarop de schattingen van $aansluiting$, $niveau$ en $richting$ worden gebaseerd een stuk kleiner geworden vergeleken met SVO 2016.

Schattingen voor de doelvariabelen worden gemaakt voor uitspluitsingen naar de volgende kenmerken, die bekend zijn voor de gehele populatie van MBO-schoolverlaters:

1. $BRIN$: (codering voor) onderwijsinstelling; er zijn 68 instellingen voor MBO-onderwijs in de doelpopulatie
2. $beroep$: (code voor) opleiding; dit is een indeling in 413 opleidingscategorieën
3. $leerweg$: of de MBO-opleiding beroepsbegeleidend of beroepsopleidend was
4. $NIVEAUMBO$: niveau van de MBO opleiding in 4 klassen (MBO-1 tot MBO-4)

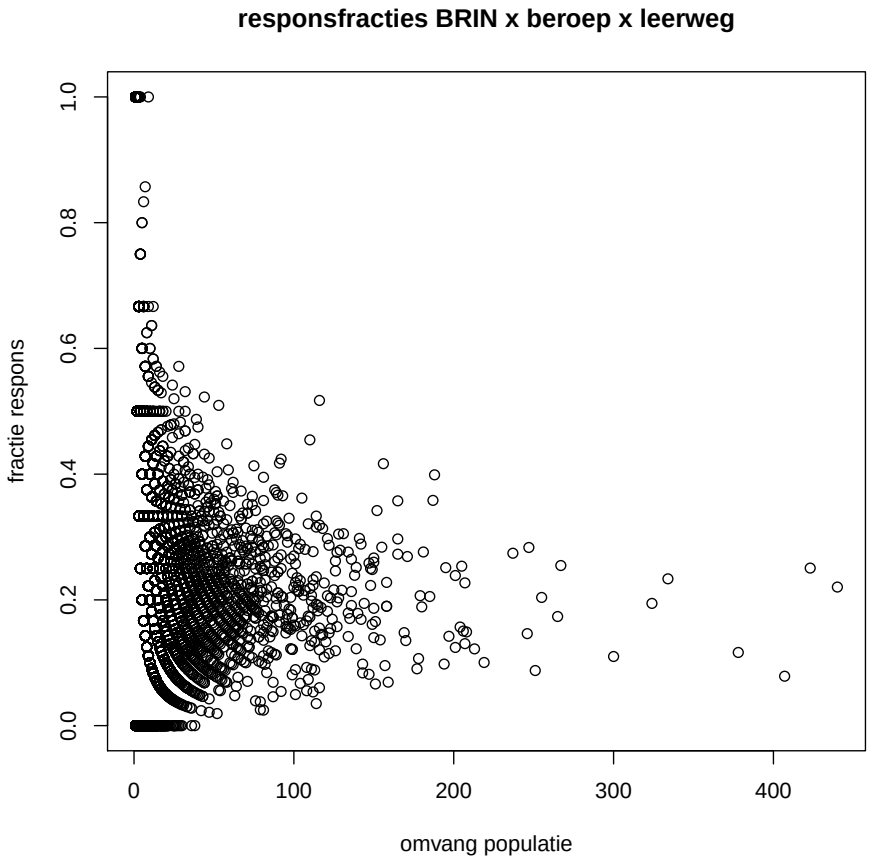
Voor SVO 2016 is in plaats van beroep de CREBO opleidingscode gebruikt als meest gedetailleerde uitsplitsingvariabele. Deze is iets gedetailleerder (644 niet-lege klassen in de SVO 2017 doelpopulatie), maar besloten is dat het detailniveau naar beroep voldoende is.

De deelpopulaties zijn de cellen in de kruistabel $BRIN \times beroep \times leerweg$. Dit is een erg grote tabel met potentieel $68 \times 413 \times 2 = 56168$ cellen. Maar omdat in de praktijk de meeste instellingen slechts een (klein) deel van de opleidingen aanbieden geldt dat

verreweg de meeste cellen leeg zijn in de populatie, zodat er geen schattingen voor gemaakt hoeven worden. Tabel 2.1 geeft een overzicht van het aantal niet-lege cellen in populatie en respons, alsook minimum, gemiddelde en maximum omvang van deze cellen. Uit de tabel blijkt ook dat er 1792 niet-lege populatiecellen zijn die geen respons bevatten. Figuur 2.1 toont de responsfracties in de 6966 (niet-lege) cellen van $BRIN \times beroep \times leerweg$. Naast cellen zonder respons zijn er ook cellen die integraal zijn waargenomen. Beide uitersten komen alleen voor bij kleine cellen.

	aantal (niet-lege) cellen	minimum omvang	gemiddelde omvang	maximum omvang
populatie	6966	1	19,7	440
respons	5174	1	5,3	106

Tabel 2.1 Het aantal niet-lege cellen in de kruistabel $BRIN \times beroep \times leerweg$, en minimum, maximum en gemiddelde omvang van de niet-lege cellen in populatie en respons.



Figuur 2.1 Histogram van de responsfracties in alle 6966 niet-lege cellen van de kruistabel $BRIN \times beroep \times leerweg$.

De gedetailleerde schattingen volgens tabel $BRIN \times beroep \times leerweg$ kunnen eenvoudig naar ingedikte niveaus geaggregeerd worden, met behulp van de populatie-aantallen. Om standaardfouten voor geaggregeerde schattingen te berekenen is er daarnaast informatie vanuit de geschatte modellen nodig over de covarianties tussen de schattingen. Als deze worden verwaarloosd dan worden de standaardfouten

te klein geschat.

Voor de volgende set van tabellen zijn schattingen en standaardfouten berekend:

1. $BRIN \times beroep \times leerweg$
2. $BRIN \times beroep$
3. $BRIN \times beroepsopleiding \times leerweg$
4. $BRIN \times beroepsopleiding$
5. $BRIN \times hoofdgroep \times leerweg$
6. $BRIN \times hoofdgroep$
7. $BRIN \times leerweg$
8. $BRIN$
9. $BRIN \times NIVEAUMBO$
10. $BRIN \times leerweg \times NIVEAUMBO$
11. $BRIN \times hoofdgroep \times NIVEAUMBO$

Opleiding wordt hierbij op verschillende ingedikte niveaus gebruikt:

1. $beroep$: een indikking van $CREBO$ in 413 klassen
2. $beroepsopleiding$: een indikking van $beroep$ in 165 klassen
3. $hoofdgroep$: een indikking van $beroepsopleiding$ in 17 klassen

Voor elke cel van deze tabellen worden de fracties werkzaam[1)], en de fracties met goede aansluiting, niveau en richting onder de werkzamen geschat. In formulevorm is de fractie werkzaam voor een cel h,

$$\theta_h^{(0)} = \frac{1}{N_h} \sum_{i \in U_h} y_i^{(0)}, \tag{1}$$

met U_h de verzameling populatie-eenheden in cel h, en N_h de omvang van U_h. Omdat de andere doelvariabelen alleen van toepassing zijn op werkzame personen worden de desbetreffende fracties gepercenteerd op de werkzame deelpopulatie:

$$\theta_h^{(v)} = \frac{\sum_{i \in U_h} y_i^{(0)} y_i^{(v)}}{\sum_{i \in U_h} y_i^{(0)}}, \tag{2}$$

waar bovenschrift $v = 1, 2, 3$ staat voor $aansluiting$, $niveau$, of $richting$. Strikt genomen is $y_i^{(v)}$ niet gedefinieerd voor niet-werkzame personen. Voor later gemak bij het bij-voorspellen voor niet-waargenomen eenheden laten we de som in de teller over de hele deelpopulatie lopen en worden niet-werkzame personen buiten beschouwing gelaten door te vermenigvuldigen met $y_i^{(0)}$.

Voor een klein aantal integraal waargenomen cellen zijn (1) en (2) direct te berekenen, maar voor de meeste cellen moet bijgeschat worden op basis van het multilevel model dat in de volgende paragraaf wordt beschreven.

[1)] Hier en in de rest van het rapport wordt de eerder gegeven definitie van de variabele $werk$ gehanteerd voor werkzaam. Het gaat dus om personen die werkzaam zijn volgens de internationale definitie en niet ook nog studeren.

3 Schattingsmethode

Op de gehele SVO respons wordt jaarlijks een weging uitgevoerd (Banning, 2017). De achtergrondvariabelen die in de weging zijn gebruikt zijn vooral bedoeld om te corrigeren voor selectiviteit van de respons. Om die reden nemen we de gebruikte weegvariabelen, voor zover ze van toepassing zijn op de MBO-populatie, ook mee in de modellen voor de domeinschatters.

Daarnaast gebruiken we extra achtergrondvariabelen die voorspellend zijn voor (één of meer) van de beschreven doelvariabelen. Verder houdt het model rekening met de gewenste uitsplitsingen. Uit praktische overwegingen is er voor gekozen om voor alle vier doelvariabelen dezelfde modelspecificatie te hanteren. Dit is redelijk omdat de meeste achtergrondvariabelen in het model voorspellend zijn voor meerdere en soms alle doelvariabelen, en omdat voor alle doelvariabelen dezelfde uitsplitsingen worden gehanteerd. Net zoals bij wegen gebruiken we dus één model voor alle doelvariabelen, maar anders dan bij wegen moet het (univariate) model voor elke doelvariabele apart geschat worden.

Het gehanteerde model is een logistisch multilevel model,

$$y_i \overset{\text{ind}}{\sim} \text{Be}(p_i),$$
$$p_i = \text{logit}^{-1}\left(\beta' x_i + \sum_{g=1}^{G} \gamma^{(g)'} z_i^{(g)}\right), \tag{3}$$

met $\text{logit}^{-1}(x) = 1/(1+e^{-x})$ de logistische of inverse logit functie. De doelvariabele y_i volgt een Bernoulli-verdeling, zodat y_i met kans p_i gelijk is aan 1 en met kans $1-p_i$ gelijk aan 0. De kans p_i wordt via de logistische functie gerelateerd aan bekende achtergrondvariabelen x_i en $z_i^{(g)}$ vermenigvuldigd met bijbehorende modelcoëfficiënten β en $\gamma^{(g)}$. De vector van achtergrondvariabelen x bevat demografische variabelen uit de Basisregistratie Personen, arbeids- of uitkeringsgerelateerde variabelen uit de Polisadministratie en opleidingsgerelateerde variabelen uit de onderwijsnummerbestanden zoals opgenomen in het populatiekader. De bijbehorende coëfficiënten β zijn ongemodelleerde of vaste effecten. De vectoren $z_i^{(g)}$ bestaan uit indicatorvariabelen voor verschillende gedetailleerde indelingen naar $BRIN$, $leerweg$, $NIVEAUMBO$ en (indikkingen van) $beroep$ en combinaties daarvan. Omdat het om zeer veel coëfficiënten gaat en veel cellen maar weinig waarnemingen bevatten worden deze coëfficiënten gemodelleerd in een volgend 'level' van het model. Deze gemodelleerde effecten worden ook random effecten genoemd. De geselecteerde variabelen x en $z^{(g)}$ worden hieronder in paragraaf 3.1 beschreven.

Het model wordt voor doelvariabele $werk$ geschat op alle responsdata. Voor de andere doelvariabelen wordt het model geschat op de werkzame respondenten minus een klein aantal met item-nonrespons op de betreffende doelvariabele.

3.1 Het geselecteerde model

De selectie van geschikte achtergrondvariabelen x en z is op basis van een aantal praktische criteria gedaan:

- x-variabelen zijn voornamelijk geselecteerd op basis van hun bijdrage aan de adjusted R-squared in lineaire regressie modellen voor de doelvariabelen. Deze maat is gemakkelijk en snel voor veel verschillende combinaties van achtergrondvariabelen uit te rekenen.
- bij de selectie van de $z_i^{(G)}$ is geprobeerd zo veel mogelijk de indelingen waarvoor schattingen gemaakt worden mee te nemen. Een selectie van de belangrijkste termen is gemaakt met behulp van de formele modelcriteria DIC (Spiegelhalter et al., 2002) en WAIC (Watanabe, 2010).
- geschatte modelparameters en domeinschattingen zijn beoordeeld op plausibiliteit.

Voor de vector x van covariaten is uiteindelijk de volgende keuze gemaakt:

```
Cluster +
ink_ontbreekt*LFTcat +
Herkomst3 +
LANDSDEEL2016 +
TypeHuishouden +
hoofdgroep*oplSBI +
TYPEMBODPL +
oplSBI * leerweg * vervolgopl2 +
leerweg * basis_ink +
leerweg * NIVEAUMBO +
(oplSBI + leerweg) * HB_SBI2008VPBL8 +
oplSBI*(NIVEAUMBO + RICHTINGMBO7) +
GBAGESLACHT * (AflGeneratie + basis_ink) +
StedGem +
SEC +
SEC3*ink_ontbreekt
```

Enkele opmerkingen hierbij:

- alle variabelen zijn categoriaal en hun indelingen zijn te vinden in Bijlage III
- een belangrijke achtergrondvariabele is `basis_ink` een indeling van inkomen uit de Polisadministratie in verschillende klassen. Voor $werk$ heeft vooral de categorie inkomen ontbreekt (die ook elders in het model voorkomt als `ink_ontbreekt`) voorspellende waarde. Deze categorie hoort bij personen die niet in de Polisadministratie voorkomen. Ook beoordelen mensen met een hoger inkomen de aansluiting van opleiding bij werk gemiddeld beter.
- sociaal-economische klasse (SEC) heeft ook een hoge voorspellende waarde voor met name $werk$. De interactie `SEC3 * ink_ontbreekt` is opgenomen omdat voor zelfstandigen het inkomen meestal ontbreekt.
- de variabele `oplSBI` is samengesteld uit opleiding en de SBI bedrijfsindeling uit de Polisadministratie. Hiervoor is uit de kruistabel van $beroepsopleiding$ met de SBI-variabele `HB_SBI2008VJJJJ` voor elke SBI bepaald welke de grootste opleidingen zijn die samen de kleinste meerderheid vormen in de populatie. Personen die niet in de Polisadministratie voorkomen krijgen de waarde 'nvt'. Van de overige personen krijgt iemand met een beroepsopleiding die tot de meerderheidsopleidingen van diens SBI behoort de waarde 'meerderheid' en anders de waarde 'minderheid'. Deze afgeleide variabele heeft voorspellende waarde voor alle vier doelvariabelen.

– van de uitsplitsingsvariabelen komen alleen $leerweg$ en $hoofdgroep$ voor als covariaten in x. De overige uitsplitsingvariabelen zijn te gedetailleerd om als ongemodelleerde effecten in het model op te nemen.

Ten opzichte van SVO 2016 is de selectie van covariaten iets gewijzigd:

– de variabele `Herkomst3` in 3 klassen vervangt de variabele `AflHerkomstCBS` met 7 klassen
– positie in het huishouden (`POSHHK`) ontbreekt. Het blijkt dat deze weinig meer toevoegt ten opzichte `TypeHuishouden`
– de interactie `oplSBI * leerweg * ink_ontbreekt` is weggelaten vanwege te kleine celvulling
– de term `leerweg * NIVEAUMBO` is toegevoegd omdat schattingen hiernaar worden uitgesplitst.

De covariaten x kunnen lang niet alle verschillen tussen de domeinen verklaren. Bij het maken van domeinschattingen is daarom belangrijk om ook expliciete domeineffecten in het model op te nemen. Vanwege de kleine aantallen waarnemingen per cel kunnen de domeinen niet als vaste effecten meegenomen worden, maar wel als gemodelleerde effecten. Voor de volgende classificaties worden gemodelleerde effecten $\gamma^{(g)}$ ($g = 1 \ldots 5$) in het model opgenomen:

1. $beroep$
2. $beroepsopleiding \times leerweg$
3. $leerweg \times BRIN$
4. $BRIN \times hoofdgroep \times NIVEAUMBO$
5. $BRIN \times beroep \times leerweg$.

Ten opzichte van SVO 2016 is er een iets aangepaste indeling gehanteerd: $beroep$ vervangt $CREBO$ in 1. en 5., in overeenstemming met het niveau waarop schattingen nu worden gemaakt, en $BRIN \times hoofdgroep \times NIVEAUMBO$ komt in de plaats van $BRIN \times beroep$.

Voor de gemodelleerde effecten wordt aangenomen dat ze, onafhankelijk van elkaar, een normaalverdeling volgen,

$$\gamma^{(g)} \sim \mathrm{N}(0, \sigma_g^2 I_{d_g}), \tag{4}$$

met variantieparameter σ_g^2, en I_{d_g} de eenheidsmatrix met dimensie gelijk aan de dimensie van $\gamma^{(g)}$, d.w.z. het aantal cellen van de kruistabel die in de populatie voorkomen. De derde term, $leerweg \times BRIN$, wordt omwille van een betere fit op een iets algemenere manier gemodelleerd, met per leerweg een aparte variantieparameter, en een correlatieparameter tussen beide leerweg klassen:

$$\gamma^{(3)} \sim \mathrm{N}(0, I_{68} \otimes \Sigma), \tag{5}$$

met Σ een algemene 2×2 covariantiematrix, en I_{68} de eenheidsmatrix van dimensie 68, het aantal instellingen.

3.2 Schatten van het model

We volgen een Bayesiaanse aanpak om het model te schatten. Dit betekent dat model (3) wordt uitgebreid met een priorverdeling voor de modelparameters, zie Bijlage I. Uit het model en de data volgt de posterior verdeling voor de modelparameters. Vanwege de complexiteit van het model kan de posterior verdeling niet direct berekend worden. Daarom gebruiken we een Markov Chain Monte Carlo (MCMC) methode. Dit is een simulatiemethode die, na een bepaalde opstartfase (burnin), trekkingen genereert uit de posterior verdeling. Met een voldoende aantal trekkingen kan de posterior verdeling goed benaderd worden. Zie Bijlage I voor meer informatie over de gebruikte MCMC methode.

De MCMC simulatie is uitgevoerd met 500 burnin iteraties. Dit bleek ruim voldoende om de reeksen naar het belangrijkste deel van de posterior verdeling te laten convergeren. Na de burnin is de simulatie voortgezet over 2000 iteraties. Elke iteratie geeft een trekking uit de posterior verdeling, en omdat het aantal modelparameters zeer groot is, is om geheugen te besparen alleen elke 4^e trekking bewaard. Er zijn steeds vier ketens parallel gesimuleerd, met onafhankelijk getrokken startwaarden. Al met al geeft dit $4 \times (2000/4) = 2000$ posterior trekkingen voor alle modelparameters. Dit bleek ruim voldoende, in de zin dat de gemaakte simulatiefout altijd veel kleiner was dan de posterior standaardfouten voor de parameters. Daarnaast is gecontroleerd op voldoende convergentie met behulp van de zogenaamde R-hat statistiek (Gelman en Rubin, 1992), de verhouding van de totale variantie over de parallelle ketens en de variantie binnen de ketens. Deze lag voor de meeste parameters zeer dicht bij 1 en voor alle parameters onder de 1,1, waarmee de simulaties voldoen aan deze vuistregel voor voldoende convergentie.

Het model is op deze manier voor elk van de vier doelvariabelen afzonderlijk geschat, waarbij, zoals eerder opgemerkt, voor $aansluiting$, $niveau$ en $richting$ alleen de respons met $werk = 1$ is gebruikt. De MCMC resultaten voor de modelparameters zijn vervolgens vertaald naar trekkingen uit de posterior verdeling voor de te schatten domeinparameters. Voor doelvariabele $werk$ geeft elke trekking r ($r = 1 \ldots 2000$) uit de posterior verdeling voor de bijbehorende modelparameters, voor een bepaald domein h, een waarde

$$\theta_h^{(0)(r)} = \frac{1}{N_h}\left(\sum_{i \in s_h} y_i^{(0)} + \sum_{i \in U_h \setminus s_h} y_i^{(0)(r)}\right), \tag{6}$$

met N_h de omvang van deelpopulatie U_h, s_h de respons en $U_h \setminus s_h$ de niet-waargenomen eenheden. De trekkingen $y_i^{(0)(r)}$ worden gegenereerd volgens

$$\begin{aligned} y_i^{(0)(r)} &\sim \mathrm{Be}\left(p_i^{(r)}\right), \\ p_i^{(r)} &= \mathrm{logit}^{-1}\left(x_i'\beta^{(r)} + \sum_g z_i^{(g)'}\gamma^{(g)(r)}\right), \end{aligned} \tag{7}$$

met $\beta^{(r)}$ en $\gamma^{(g)(r)}$ de trekkingen voor de modelcoëfficiënten uit de MCMC simulatie voor $werk$. Samen vormen $\left\{\theta_h^{(0)(r)}\right\}_{r=1\ldots2000}$ een benadering van de posterior verdeling voor de fractie werkzaam $\theta_h^{(0)}$, waarmee schattingen en standaardfouten of intervallen kunnen worden berekend.

Voor de overige doelvariabelen worden analoog trekkingen voor domeinparameters gegenereerd volgens

$$\theta_h^{(v)(r)} = \frac{\sum_{i \in s_h} y_i^{(0)} y_i^{(v)} + \sum_{i \in U_h \setminus s_h} y_i^{(0)(r)} y_i^{(v)(r)}}{\sum_{i \in s_h} y_i^{(0)} + \sum_{i \in U_h \setminus s_h} y_i^{(0)(r)}}, \tag{8}$$

waarbij de trekkingen $y_i^{(v)(r)}$ gebaseerd zijn op de MCMC simulatie voor de betreffende doelvariabele. We merken op dat deze doelvariabelen met een tweedelig model worden geschat: het model voor $werk$ en het model voor de betreffende doelvariabele, vergelijkbaar met tweedelige modellen die gebruikt worden bij het modelleren van data met extra nullen, zie bijvoorbeeld Pfeffermann et al. (2008) en Krieg et al. (2016).

4 Resultaten

Voor elke doelvariabele is het multilevel model geschat door middel van een MCMC simulatie. Vervolgens zijn simulatievectoren voor alle domeinparameters berekend zoals in de vorige paragraaf beschreven. Met deze Monte Carlo (MC) benaderingen van de posterior verdelingen voor de domeinparameters kunnen eenvoudig puntschattingen, standaardfouten of intervalschattingen gemaakt worden.

Als puntschatting gebruiken we (de MC benadering van) het posterior gemiddelde, als standaarfout (de MC benadering van) de posterior standaardfout, en als 95% interval het interval begrensd door (de MC benaderingen van) de 2,5 en 97,5 percentielpunten.

In de grafieken in Bijlage II worden de modelgebaseerde schattingen vergeleken met directe schattingen. De directe schattingen met standaardfouten zijn berekend als

$$\hat{\bar{Y}}_h^{(0)} = \frac{1}{n_h}\sum_{i\in s_h} y_i^{(0)} \quad en \quad \hat{\bar{Y}}_h^{(v)} = \frac{\sum_{i\in s_h} y_i^{(v)}}{\sum_{i\in s_h} y_i^{(0)}} \;\; voor \;\; v = 1,2,3$$
$$\text{se}\left(\hat{\bar{Y}}_h^{(v)}\right) = \sqrt{\frac{1-n_h/N_h}{n_h}\hat{\bar{Y}}_h^{(v)}\left(1-\hat{\bar{Y}}_h^{(v)}\right)} \;\; voor \;\; v = 0,1,2,3 \qquad (9)$$

met n_h de omvang van de respons s_h. Voor domeinen zonder respons zijn geen directe schattingen mogelijk, en deze domeinen komen dan ook niet voor in de grafieken. Voor $aansluiting$, $niveau$ en $richting$ wordt de (meestal kleine) extra bijdrage aan de standaardfouten als gevolg van het percenteren op $werk$ verwaarloosd.

Plots voor alle vier doelvariabelen worden getoond voor de meest gedetailleerde indeling naar $BRIN \times beroep \times leerweg$ (figuren II.1 t/m II.8) en voor de indeling naar enkel $BRIN$ (figuren II.9 t/m II.16). De figuren laten zien dat de modelschattingen een kleinere spreiding hebben, zeker voor de meest gedetailleerde tabel. Ook is te zien dat de verschillen tussen directe en modelschattingen het grootst zijn bij de kleinste domeinen (kleine cirkels). Voor grotere domeinen die in de regel veel waarnemingen bevatten zijn de verschillen kleiner (grote cirkels). Belangrijk is dat de standaardfouten van de modelschattingen, op enkele uitzonderingen na, beduidend kleiner zijn dan die van de directe schattingen.

De schattingen op de verschillende aggegatieniveaus zijn onderling consistent (op een kleine simulatiefout na voor $aansluiting$, $niveau$ en $richting$). Er is geen exacte consistentie afgedwongen met schattingen op een hoger geaggregeerd niveau die met behulp van de weging gemaakt kunnen worden.

In tabel 4.1 staan de geaggregeerde modelschattingen voor SVO 2016 en SVO 2017. Er zijn grote verschillen vooral door de andere definitie van de variabele $werk$ die voor SVO 2017 exclusief studerende werkenden is. Het is daarom niet verwonderlijk dat het percentage voor 2017 veel lager ligt. De percentages van de overige doelvariabelen liggen juist hoger, waarschijnlijk vooral omdat het werk van niet-studerende werkzamen beter aansluit bij de gevolgde studie.

	SVO 2016	SVO 2017
werk	84,0	65,4
aansluiting	72,2	76,5
niveau	67,1	76,1
richting	57,9	66,5

Tabel 4.1 Geaggregeerde modelschattingen gebaseerd op SVO 2016 en SVO 2017 voor de fracties *werk*, *aansluiting*, *niveau* **en** *richting* **(in %) voor de populatie van MBO schoolverlaters. Merk op dat de variabele** *werk* **anders gedefinieerd is voor SVO 2017.**

5 Samenvatting en discussie

Voor OCW zijn schattingen gemaakt van enkele belangrijke aspecten van aansluiting van opleiding bij werk voor MBO schoolverlaters naar instelling en opleiding. Omdat het om een erg gedetailleerd niveau gaat is het niet goed mogelijk om deze schattingen met de reguliere SVO gewichten te maken. In plaats daarvan zijn deze schattingen berekend op basis van een voor dit doel ontwikkeld multilevel model. Dit model houdt rekening met het binaire karakter van de doelvariabelen en de gewenste uitsplitsingen van de schattingen, en gebruikt een groot aantal uit registraties beschikbare hulpvariabelen om tot populatieschattingen te komen.

Voor het schatten van het model is een Bayesiaanse simulatie-aanpak gekozen. Uit de simulatie-output kunnen eenvoudig schattingen en bijbehorende standaardfouten of onzekerheidsintervallen berekend worden. Deze cijfers zijn voor een aantal vooraf gespecificeerde tabellen gemaakt. De standaardfouten van de modelgebaseerde schattingen zijn beduidend kleiner dan die van directe schattingen, en modelgebaseerde schattingen kunnen ook voor domeinen zonder waarnemingen worden gemaakt. Schattingen voor lineaire combinaties van cellen, zoals andere aggregaten of verschillen tussen opleidingen of instellingen, kunnen worden berekend als dezelfde lineaire combinaties van de schattingen per cel. Voor de berekening van bijbehorende standaardfouten of intervallen kan dan het best wel rekening gehouden worden met mogelijke correlaties. Dit gaat automatisch als deze schattingen direct uit de simulatie-output worden berekend.

De beschreven methode zou in principe elk jaar kunnen worden toegepast om schattingen naar instelling en opleiding te maken. Bij steekproefsgewijze in plaats van integrale benadering van de populatie moet het model mogelijk wel vereenvoudigd worden. Een mogelijkheid die we hier niet hebben onderzocht is om een model te specificeren over meerdere jaren waarbij (sommige) coëfficiënten tijdsafhankelijk worden gemodelleerd. De verwachting is dat met een dergelijk groter model nauwkeuriger schattingen gemaakt kunnen worden, met name van ontwikkelingen. Maar of dat de extra complexiteit en ontwikkelingskosten rechtvaardigt is de vraag.

Referenties

Banning, R. (2017). Weging SVO 2016. Memo 24-02-2017, CBS Heerlen.

Boonstra, H. J. (2017a). *mcmcsae: MCMC Small Area Estimation*. R package under development, version 0.8.1.

Boonstra, H. J. (2017b). Schoolverlatersonderzoek 2016: schattingen voor kleine deelpopulaties van MBO schoolverlaters. Methodologisch rapport, mei 2017, CBS Heerlen.

Gelfand, A. en Smith, A. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association* 85, 398--409.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis* 1 (3), 515--533.

Gelman, A. en Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Gelman, A. en Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science* 7 (4), 457--472.

Geman, S. en Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattn Anal. Mach. Intell.* 6, 721--741.

Krieg, S., Boonstra, H. J., en Smeets, M. (2016). Small-Area Estimation with Zero-Inflated Data--a Simulation Study. *Journal of Official Statistics* 32 (4), 963--986.

O'Malley, A. en Zaslavsky, A. (2008). Domain-Level Covariance Analysis for Multilevel Survey Data with Structured Nonresponse. *Journal of the American Statistical Association* 103 (484), 1405--1418.

Pfeffermann, D., Terryn, B., en Moura, F. A. (2008). Small area estimation under a two-part random effects model with application to estimation of literacy in developing countries. *Survey Methodology* 34 (2), 235--249.

Polson, N. G., Scott, J. G., en Windle, J. (2013). Bayesian inference for logistic models using Pólya--Gamma latent variables. *Journal of the American statistical Association* 108 (504), 1339--1349.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rao, J. N. en Molina, I. (2015). *Small area estimation*. John Wiley & Sons.

Spiegelhalter, D., Best, N., Carlin, B., en van der Linde, A. (2002). Bayesian Measures of Model Complexity and Fit. *Journal of the Royal Statistical Society B* 64 (4), 583--639.

van Berkel, K. (2016). Steekproef schoolverlatersonderzoek 2016. Memo 25-07-2016, CBS Heerlen.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research* 11 (Dec), 3571--3594.

Bijlage

I Binomiaal multilevel model

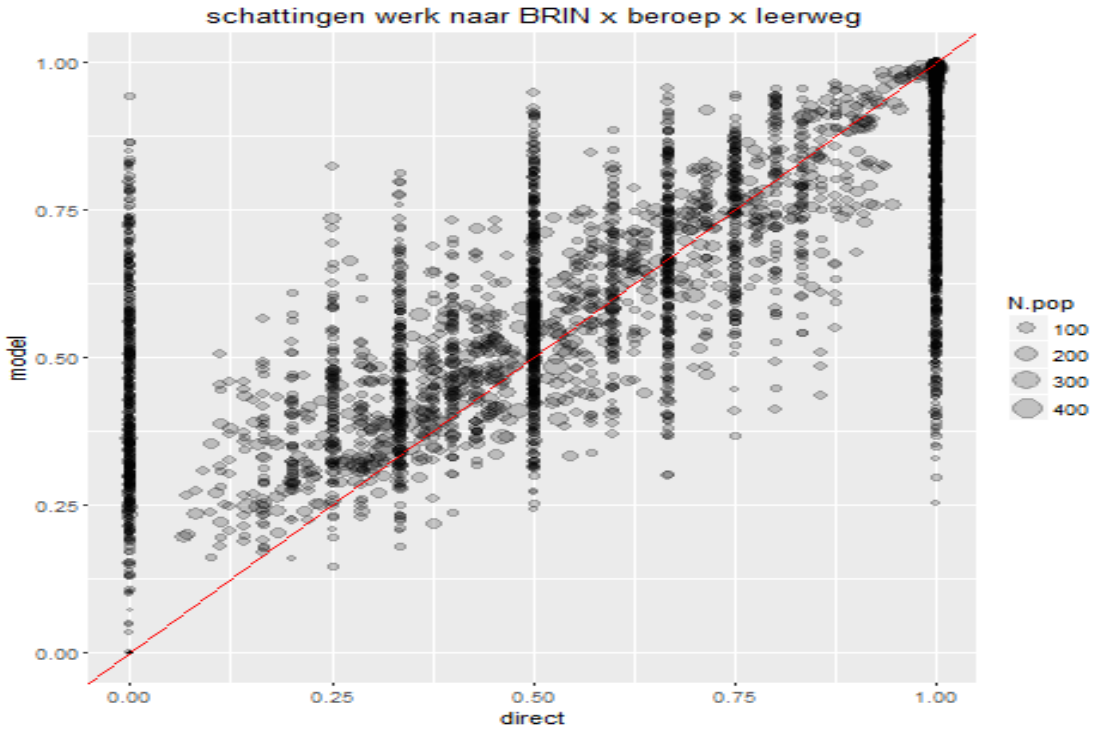
Het gebruikte model is

$$
\begin{aligned}
y_i &\overset{\text{ind}}{\sim} \text{Be}(p_i)\,, \\
p_i &= \text{logit}^{-1}\left(\beta' x_i + \sum_g \gamma^{(g)'} z_i^{(g)}\right), \\
\gamma^{(g)} &\sim \text{N}(0, \sigma_g^2 I_{d_g}) \quad voor\ \ g = 1,2,4,5\,, \\
\gamma^{(g)} &\sim \text{N}(0, I_{68} \otimes \Sigma) \quad voor\ \ g = 3\,.
\end{aligned}
\tag{10}
$$

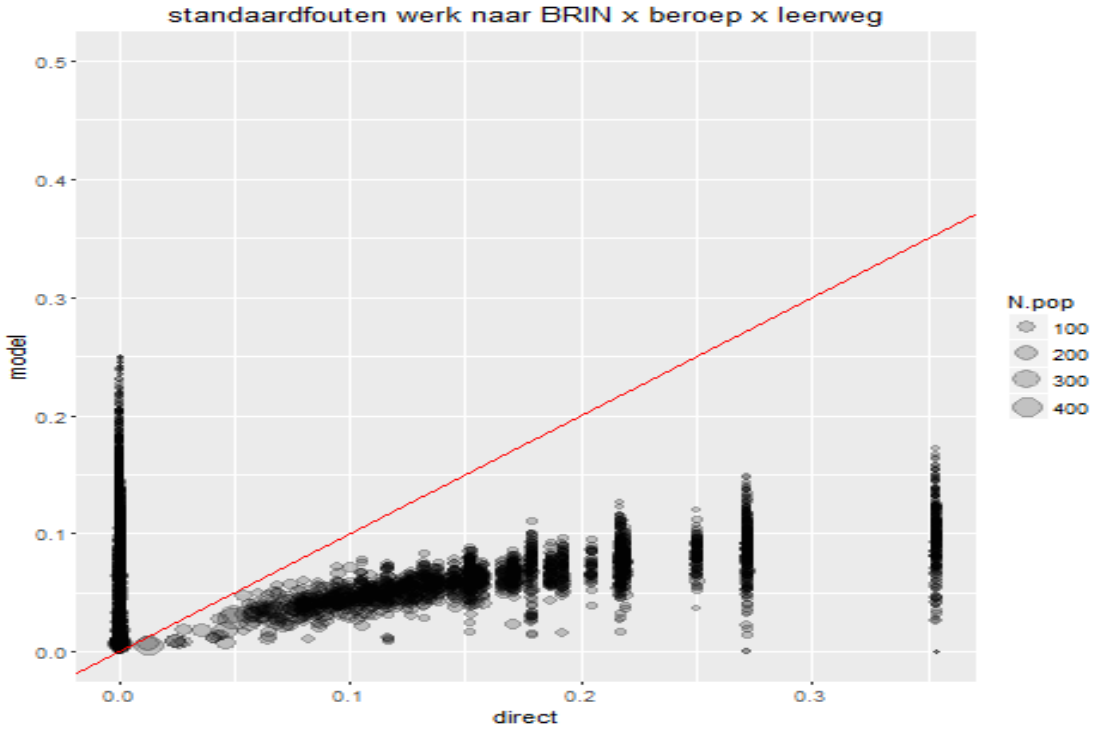
Omdat een Bayesiaanse aanpak wordt gevolgd moeten ook priorverdelingen voor de overige modelparameters worden gespecificeerd. Voor de componenten van β zijn onafhankelijke normale priorverdelingen gebruikt met gemiddelde 0 en variantie 4. Voor logistische regressie met categoriale covariaten is dit een vrij diffuse prior. Voor de standaarddeviaties σ_g zijn half-Cauchy priors gebruikt met schaalparameter 1. In Gelman (2006) wordt aangetoond dat dit geschikte standaardpriors zijn voor standaarddeviatieparameters in multilevel modellen . Ten slotte wordt voor de volledig geparametriseerde covariantiematrix Σ een zogenaamde scaled-inverse-Wishart verdeling gebruikt met 3 vrijheidsgraden en eenheidsschaalmatrix, zoals voorgesteld in O'Malley en Zaslavsky (2008) en aanbevolen in Gelman en Hill (2007). Deze prior kan ook gezien worden als generalisatie van de gehanteerde priors voor de (gekwadrateerde) standaardfouten.

Het model wordt geschat met een MCMC simulatiemethode, in het bijzonder de Gibbs sampler (Geman en Geman (1984); Gelfand en Smith (1990)). Hierbij wordt herhaaldelijk uit de conditionele posterior verdelingen van (blokken van) modelparameters getrokken. We maken daarbij gebruik van de representatie van de binomiale likelihood als mixture van een normale verdeling (Polson et al., 2013) waardoor alle conditionele verdelingen eenvoudig zijn. De berekeningen zijn uitgevoerd in R (R Core Team, 2015), met package mcmcsae (Boonstra, 2017a).

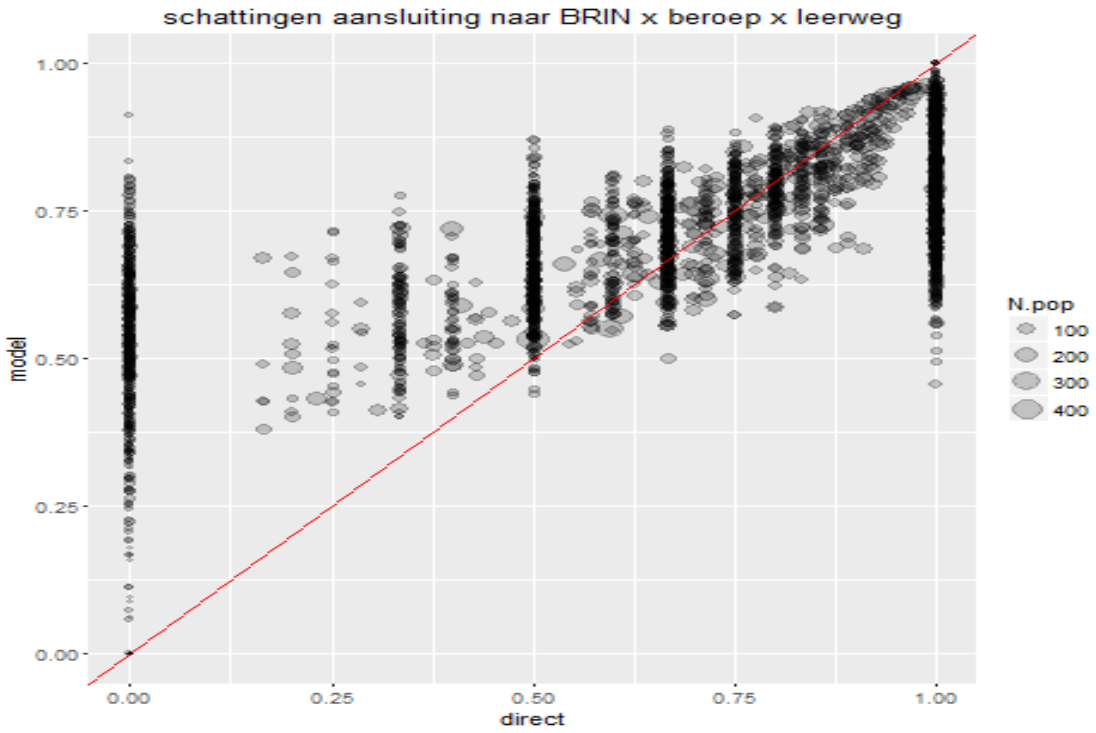
II Vergelijking met directe schattingen



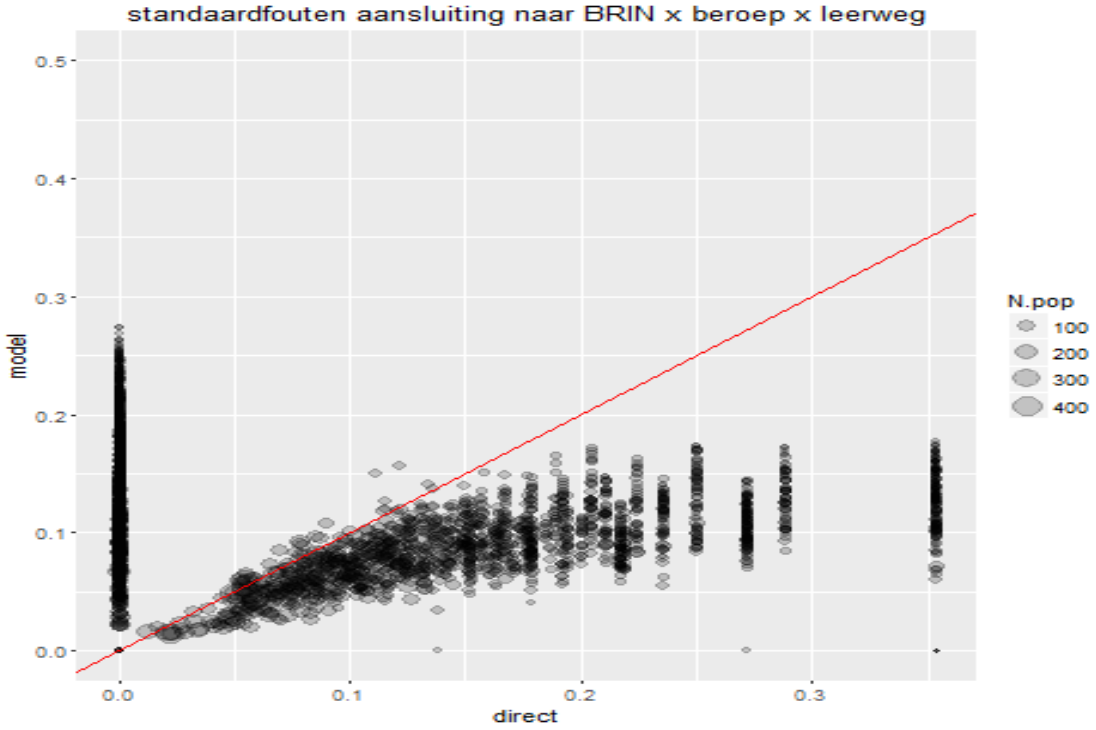
Figuur II.1 Modelgebaseerde en directe schattingen voor fracties *werk* **naar** $BRIN \times beroep \times leerweg$**.**



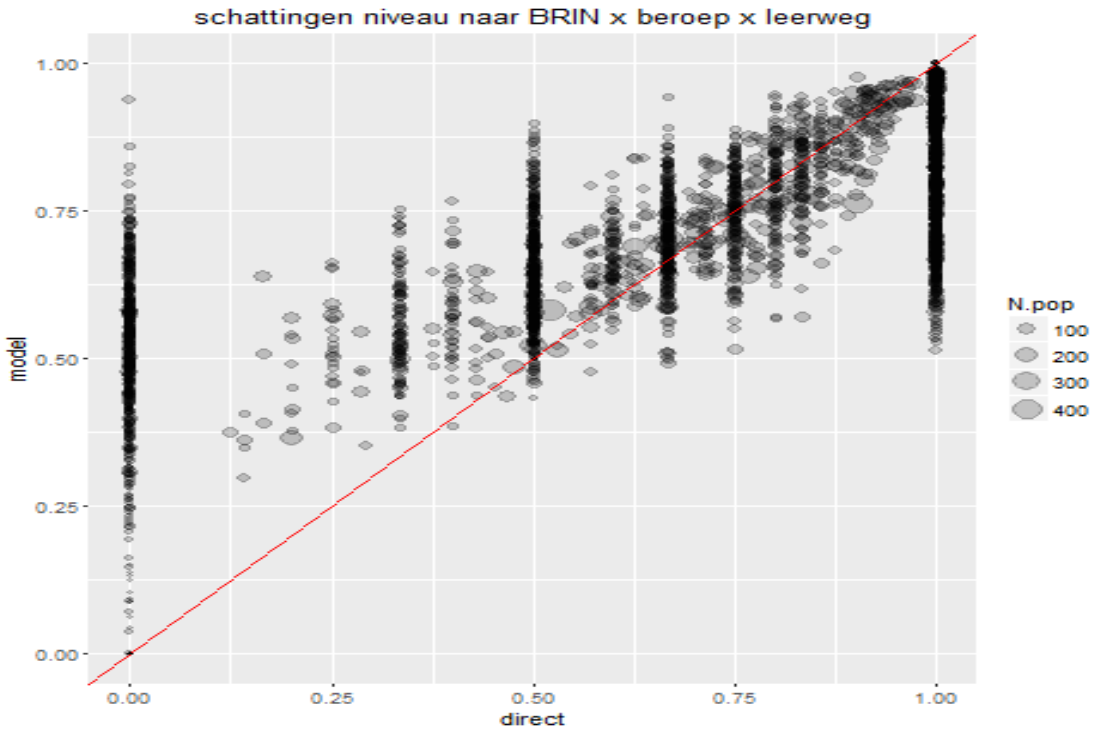
Figuur II.2 Standaardfouten voor modelgebaseerde en directe schattingen voor fracties *werk* **naar** $BRIN \times beroep \times leerweg$**.**



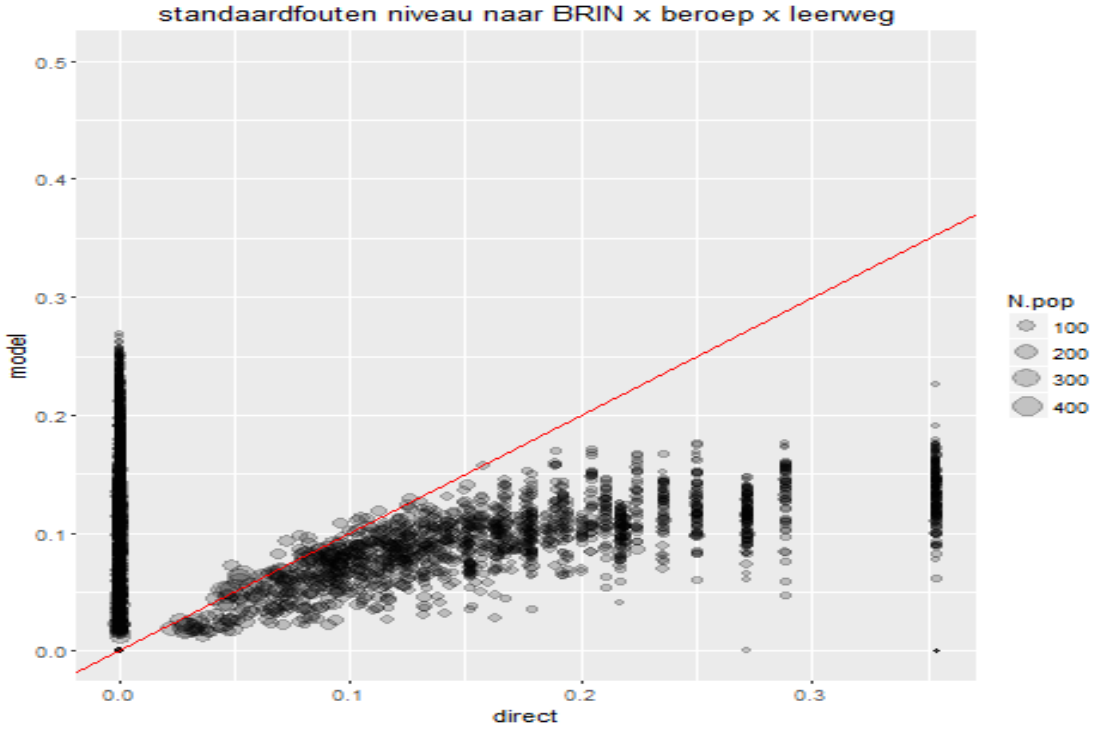
Figuur II.3 Modelgebaseerde en directe schattingen voor fracties *aansluiting* **naar** $BRIN \times beroep \times leerweg$**.**



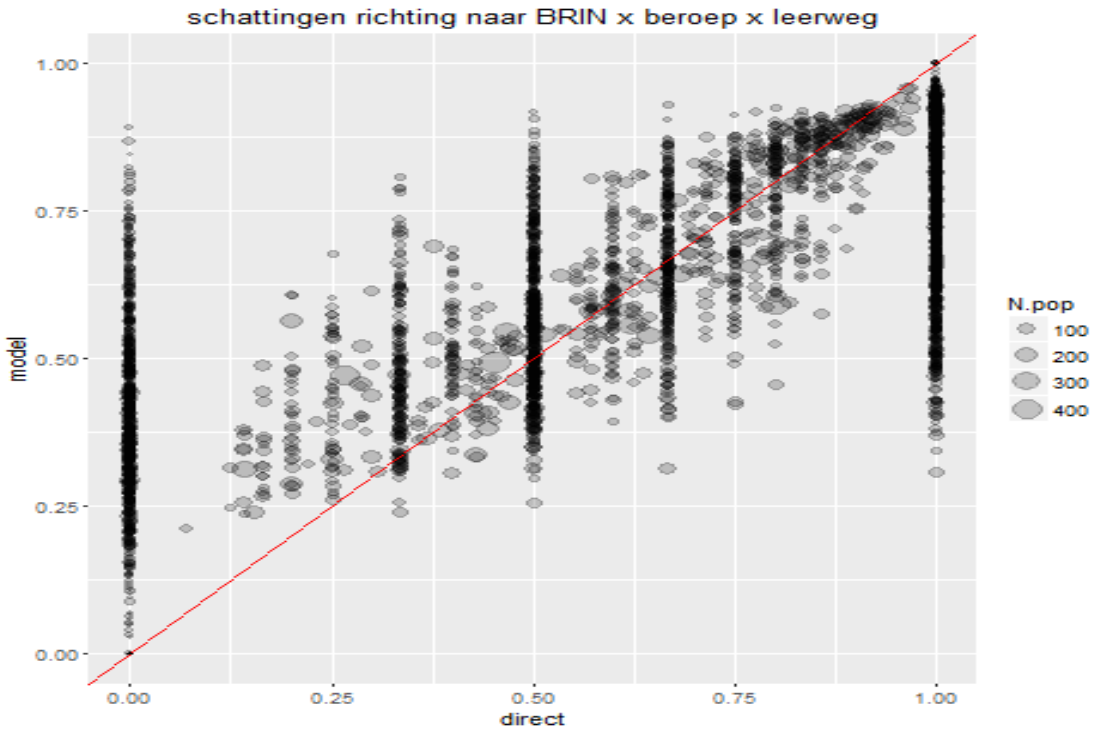
Figuur II.4 Standaardfouten voor modelgebaseerde en directe schattingen voor fracties *aansluiting* **naar** $BRIN \times beroep \times leerweg$**.**



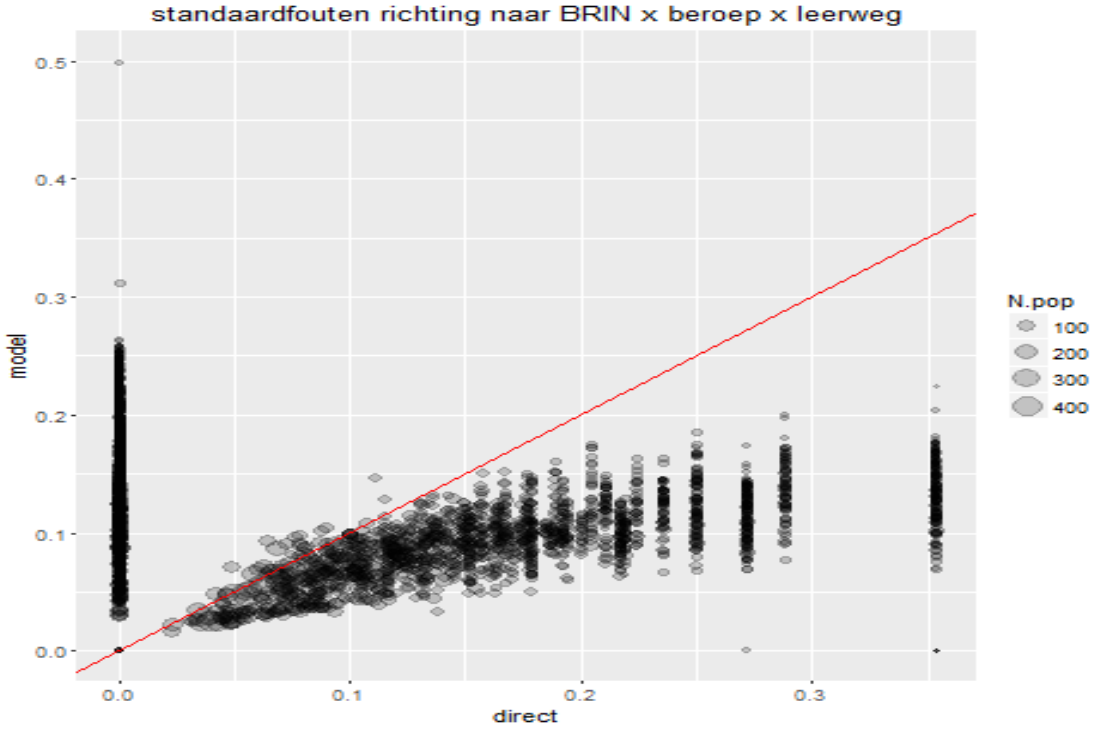
Figuur II.5 Modelgebaseerde en directe schattingen voor fracties *niveau* **naar** $BRIN \times beroep \times leerweg$**.**



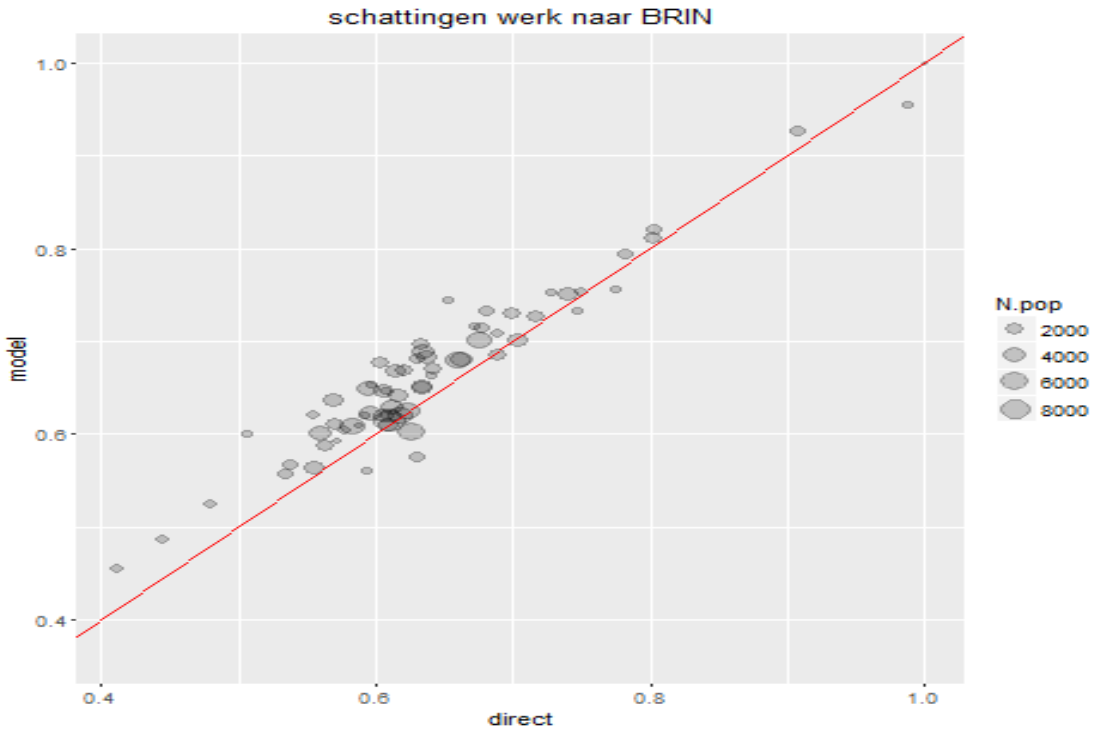
Figuur II.6 Standaardfouten voor modelgebaseerde en directe schattingen voor fracties *niveau* **naar** $BRIN \times beroep \times leerweg$**.**



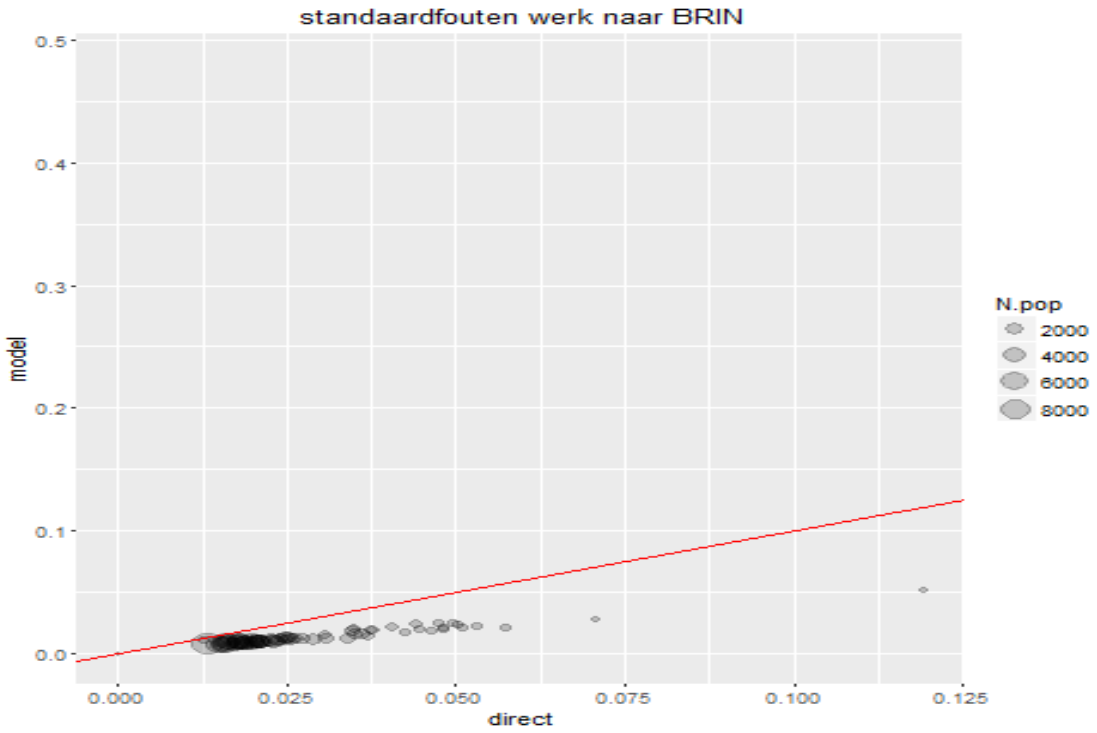
Figuur II.7 Modelgebaseerde en directe schattingen voor fracties *richting* **naar** $BRIN \times beroep \times leerweg$**.**



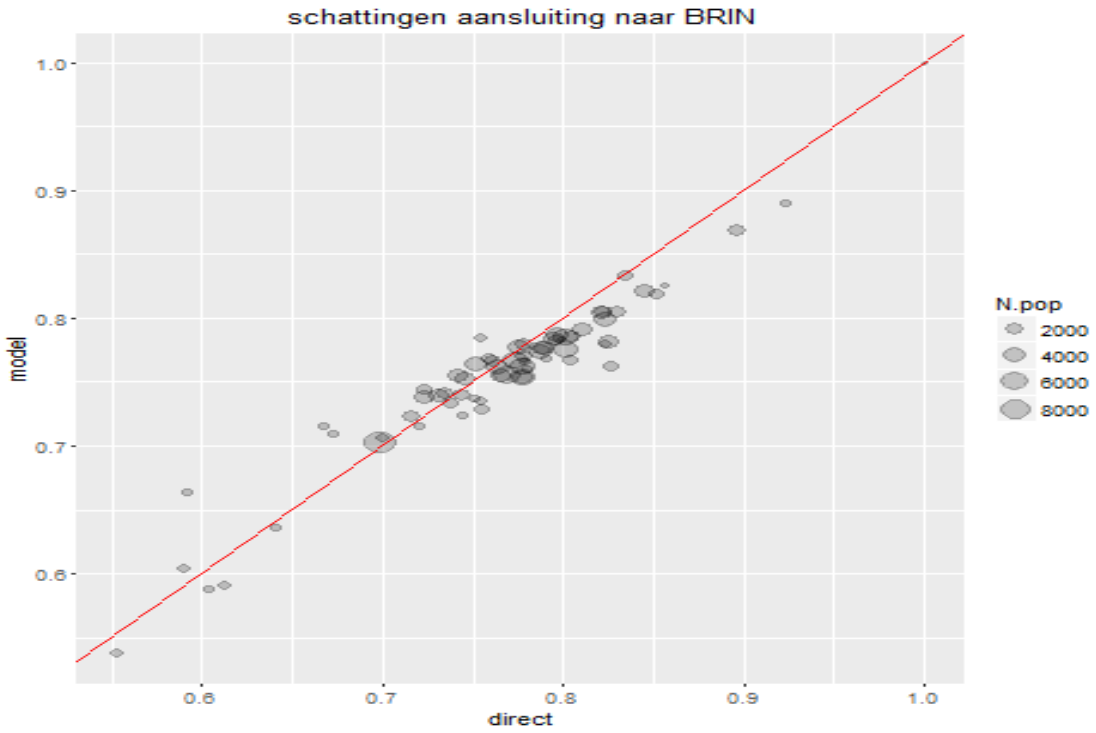
Figuur II.8 Standaardfouten voor modelgebaseerde en directe schattingen voor fracties *richting* **naar** $BRIN \times beroep \times leerweg$**.**



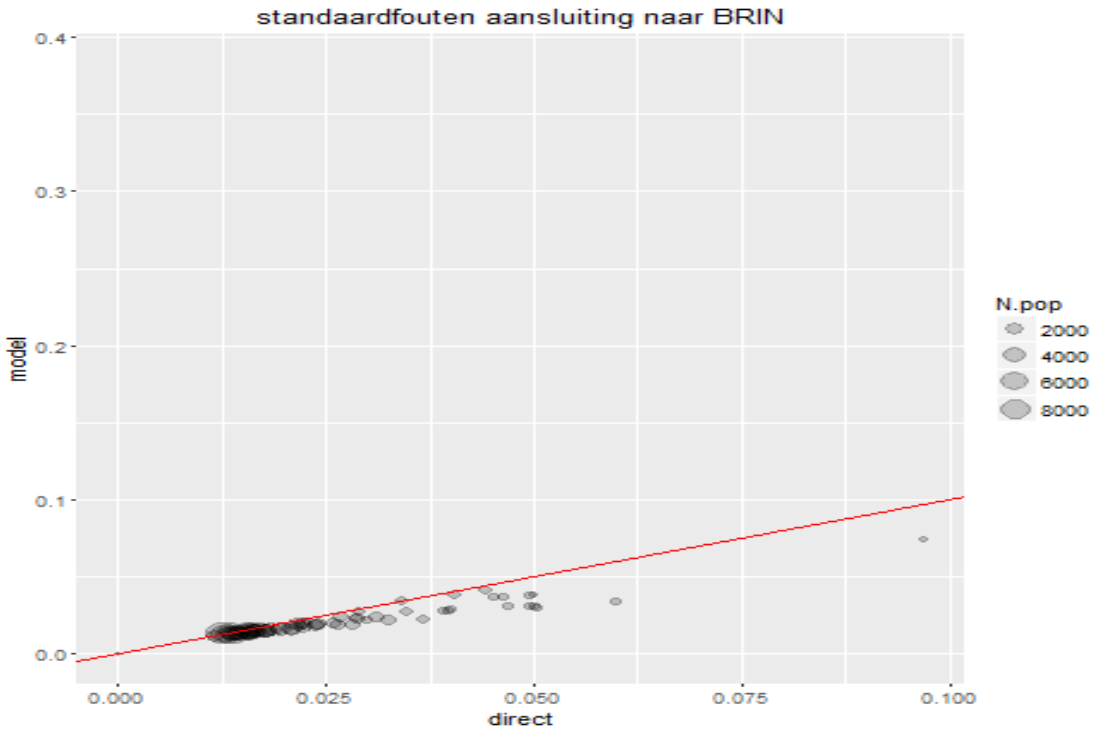
Figuur II.9 Modelgebaseerde en directe schattingen voor fracties *werk* naar *BRIN*.



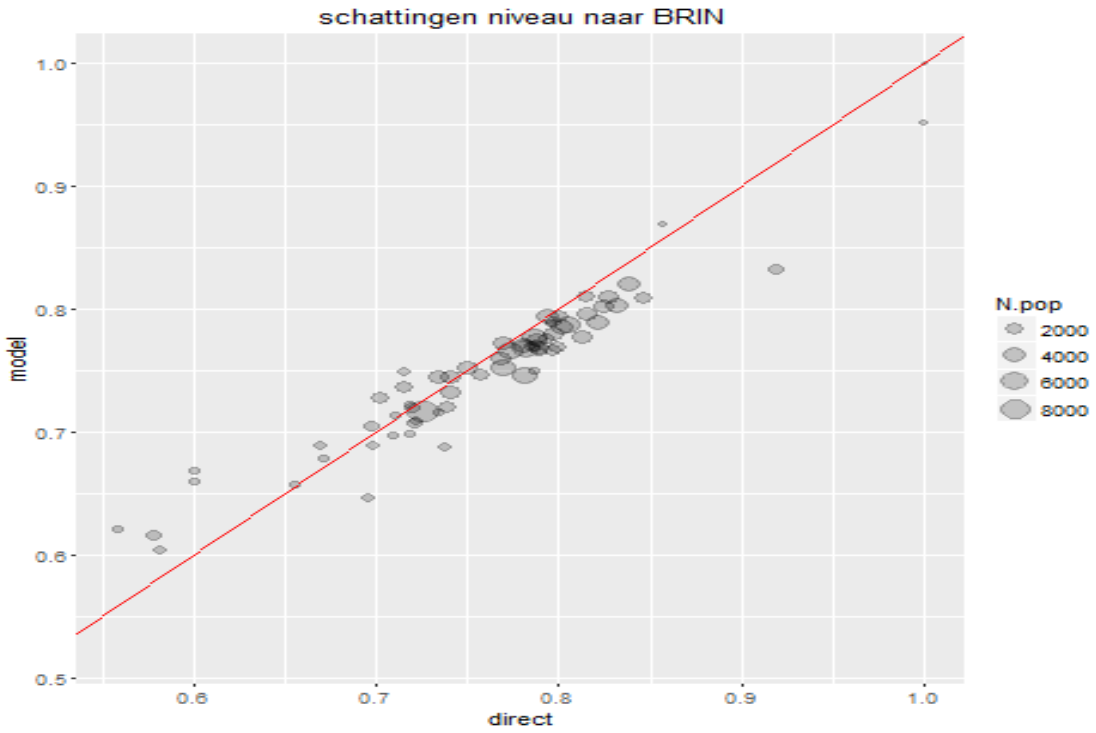
Figuur II.10 Standaardfouten voor modelgebaseerde en directe schattingen voor fracties *werk* naar *BRIN*.



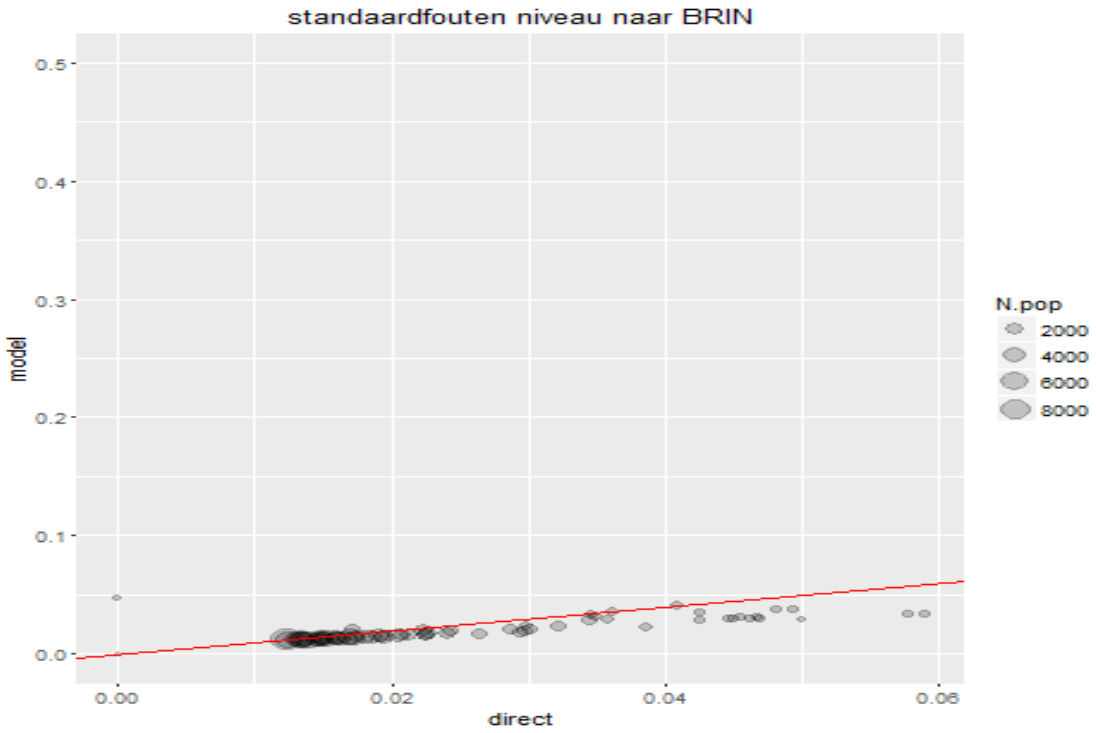
Figuur II.11 Modelgebaseerde en directe schattingen voor fracties *aansluiting* **naar** *BRIN*.



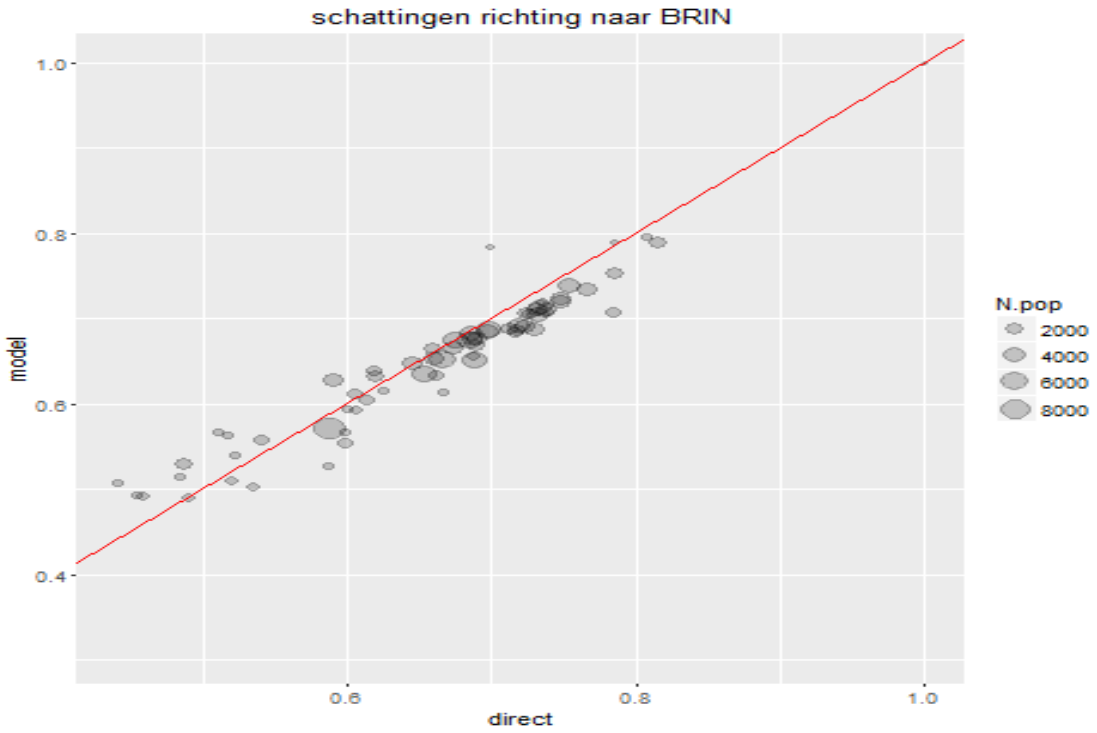
Figuur II.12 Standaardfouten voor modelgebaseerde en directe schattingen voor fracties *aansluiting* **naar** *BRIN*.



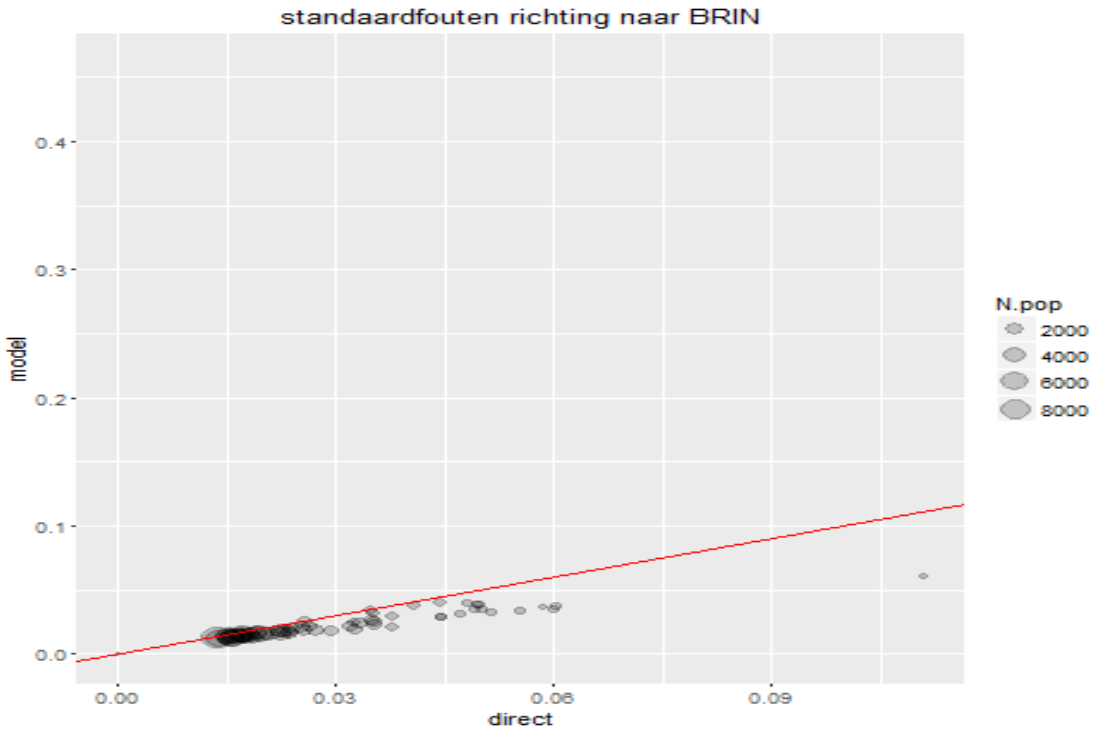
Figuur II.13 Modelgebaseerde en directe schattingen voor fracties $niveau$ **naar** $BRIN$**.**



Figuur II.14 Standaardfouten voor modelgebaseerde en directe schattingen voor fracties $niveau$ **naar** $BRIN$**.**



Figuur II.15 Modelgebaseerde en directe schattingen voor fracties *richting* **naar** *BRIN*.



Figuur II.16 Standaardfouten voor modelgebaseerde en directe schattingen voor fracties *richting* **naar** *BRIN*.

III Indeling in klassen van de achtergrondvariabelen x

Cluster
 - 1 MBO, uitgezonderd Cluster 5
 - 5 MBO-1, 12-22 jaar, geen opleiding in 2016/2017

GBAGESLACHT
  - man
  - vrouw

LFTcat, gebaseerd op leeftijd op 1 oktober 2016
  - jonger dan 19 jaar
  - 19 jaar
  - 20 jaar
  - 21 jaar
  - 22 jaar
  - 23 t/m 25 jaar
  - 26 t/m 30 jaar
  - 31 t/m 40 jaar
  - ouder dan 40 jaar

Herkomst3
  - Personen met een Nederlandse achtergrond
  - Personen met een westerse migratieachtergrond
  - Personen met een niet-westerse migratieachtergrond

AflHerkomstCBS
  - 0 Personen met een Nederlandse achtergrond
  - 1 Marokko
  - 2 Turkije
  - 3 Suriname
  - 4 Voormalige Nederlandse Antillen en Aruba
  - 5 Overige niet-westerse landen
  - 6 Overige westerse landen

AflGeneratie
  - 0 Personen met een Nederlandse achtergrond
  - 1 eerste generatie niet-Nederlandse achtergrond
  - 2 tweede generatie niet-Nederlandse achtergrond

LANDSDEEL2016, landsdeel in 4 klassen

StedGem, stedelijkheidsgraad in 5 klassen

POSHHK
  - Hoofdkostwinner

- Hoofdkostwinner, met partner
- Gehuwde partner
- Ongehuwde partner
- Minderjarig kind
- Meerderjarig kind
- Overig huishoudenslid

TypeHuishouden, type huishouden in 7 klassen

basis_ink, inkomensklassen gebaseerd op HB_SBASISLOON
- ontbreekt
- < 400
- 400 - 1000
- 1000 - 2000
- > 2000

ink_ontbreekt, indikking van basis_ink
- 0 ontbreekt niet
- 1 ontbreekt

SEC
- ambtenaar
- bijstandsontvanger
- ontvanger van ov soc voorz
- ontvanger werkloosheidsuitk
- overig niet actief
- student werknemer particulier bedrijf
- zelfstandig

SEC3, indikking van SEC
- zelfstandig
- werknemer
- overig

hoofdgroep, indikking van CREBO, zie hoofdtekst

oplSBI, afleiding uit beroepsopleiding en HB_SBI2008VJJJJ,
zie hoofdtekst

TYPEMBODPL, type MBO opleiding in 3 klassen
(leerweg is hier een indikking van)

vervolgopl2
- geen opleiding gevolgd in 2016/2017
- opleiding gevolgd in 2016/2017

HB_SBI2008VPBL8, publicatiegroep sbi2008 statline in 8 klassen

NIVEAUMBO, niveau MBO opleiding in 4 klassen (MBO-1 tot MBO-4)

RICHTINGMBO7, opleidingsrichting in 7 klassen, gebaseerd op code isced 97 field1